Complex statistical modelling of socio-economic variables in public health.

by

Robert W. Eyre

Thesis

Submitted to the University of Warwick

for the degree of

Doctor of Philosophy

Centre for Complexity Science

January 2018

Contents

List of Tables

List of Figures

Abbreviations

- Add Health - National Longitudinal Study of Adolescent to Adult Health
- AIC - Akaike Information Criterion
- AIDS - Acquired Immune Deficiency Syndrome
- AUROC - Area Under Receiver Operator Characteristic curve
- BIC - Bayesian Information Criterion
- BMI - Body Mass Index
- CAG - Community Advisory Group
- CDC - Center for Disease Control
- CDF - Cumulative Distribution Function
- CES-D - Centre for Epidemiological Studies Depression scale
- CPC - Candidate set of Parents and Children
- CPT - Conditional Probability Table
- DAG - Directed Acyclic Graph
- D-separation - Directional separation

- GLM - Generalised Linear Model(ling)
- HDSS - Health and Socio-demographic Surveillance System
- HIV - Human Immunodeficiency Virus
- HL - Homer-Lemeshow test
- KS - Kolmogorov-Smirnov test
- LINC - Learning, Information dissemination, and Networking with Community office
- LS - Lauritzen-Spiegelhalter algorithm
- MAP - Maximum A Posteriori estimate
- MLE - Maximum Likelihood Estimate/Estimation
- MMPC - Max-Min Parents and Children algorithm
- NDVI - Normalised Difference Vegetation Index
- ROC - Receiver Operator Characteristic curve
- SES - Socio-Economic Status
- WHO - World Health Organisation

Acknowledgments

First and foremost, thanks must go to my supervisors Frances Griffiths, Thomas House, and F. Xavier Gómez-Olivé. Without their guidance none of this would be possible, and they have helped train me into the researcher I am now. Many others have contributed along the way, but it is only without these three that this PhD could not have been this successful.

Thanks to Ed Hill and Andrea Apolloni for advice on the social contagion work.

Thanks to Jocelyn Watkins for the initial idea of analysing pregnancy in the Agincourt dataset. Thanks to Sulaimon Afolabi for extracting all the Agincourt data I needed and answering so many questions on it so quickly and so well. Thanks to the entire staff of the Agincourt research unit and the Witwatersrand Rural Facility for making my visit to the Agincourt study area so enjoyable and informative, especially Ryan Wagner and Fezile Mdluli.

Thanks to Martine Barons and Wayne Twine for advice on food security and Bayesian networks, as well as further thanks to Martine for comments on my thesis. Thanks to Simon Khoza, George Shabangu, Ralbon Godi, Polite Zitha, Happy Mnisi, Aubrey Nkuna, Sibangile Khosa, and Busi Mdluli for meeting with me and for acting as experts in my first elicitation, and to Betty Masinga, Thelma Nyathi, Thandiwe Whati, Aubrey Nkuna, Happy Mnisi, Larry Mzimba, Glory Khoza, and Monical Khosa for acting as experts in my second elicitation. Thanks especially to Rhian Twine for performing these elicitations for me.

Immense thanks must go to the staff and fellow students of the Centre for Complexity Science and Mathematics of Real World Systems Doctoral Training Centres for giving me the opportunity to do this research and establishing an atmosphere that has helped me learn so much. Thanks to Jim Skinner and Michael Pearce for advice on machine learning and data science, and for listening to and discussing my thoughts on my work. Thanks also to Marcus Ong and Dan Sprague for giving me a job to go on with after this that enables me to continue using the skills I learnt here (and I am not putting this here to suck up, I promise) and for advice on the writing up and viva process.

Thanks to my examiners Matthew Sperrin and Deirdre Hollingsworth for giving me a friendly and enlightening examination process.

This work was supported by the Engineering and Physical Sciences Research Council as part of the University of Warwick Complexity Science Doctoral Training Centre. This research uses data from Add Health, a programme project directed by Kathleen Mullan Harris, and designed by J. Richard Udry, Peter S. Bearman and Kathleen Mullan Harris at the University of North Carolina at Chapel Hill. Special acknowledgement is due to Ronald R. Rindfuss and Barbara Entwisle for assistance in the original design. It also uses data from the Agincourt Health and Socio-demographic Surveillance System, run by the Medical Research Council / University of the Witwatersrand Rural Public Health and Health Transitions (Agincourt) Research Unit. Finally, it also uses data from the MODIS/Terra sensor satellite image dataset. The MOD13Q1: MODIS/Terra Vegetation Indices 16-Day L3 Global 250 m SIN Grid V006 data was retrieved from the online Data Pool, courtesy of the NASA EOSDIS Land Processes Distributed Active Archive Center (LP DAAC), USGS/Earth Resources Observation and Science (EROS) Center, Sioux Falls, South Dakota, https://doi.org/10.5067/modis/mod13q1.006. Thanks must go to all these as this

research could not have been done otherwise.

Finally, some personal thanks. First to my mum Maureen Eyre, dad Geoffrey Eyre, and sister Vicky Eyre. Without all of you I would not have had the life that got me here. Thanks Dad for all the maths. It seems it was useful after all. Thanks Mum for all the love and care. I know you would be so proud of me.

My biggest personal thanks must go to my wife May Tsang-Eyre. Your strength is essential to everything I do, and safe to say supported me throughout this. From lifting me up and pushing me forward each time I felt too worn out to continue, to providing me the freedom to pursue this PhD in the first place and to go on to a great job afterwards. All I do, including this thesis, I do for you. It's only from how it helps us build our lives together that it gains its greatest meaning to me.

My absolute final thanks is to the little ball of fur who lazed around my flat as I wrote all this. Merlin, you were a dog so you would not have understood this even if you could have read it, but thanks for existing and being a warm presence throughout this. You were the best part of the last five years, and I miss you.

Declarations

This thesis is submitted to the University of Warwick in support of my application for the degree of Doctor of Philosophy. It has been composed by myself and has not been submitted in any previous application for any degree. Parts of the work presented in Chapter 3 and appendix A follow on from work that formed part of the author's MSc in Complexity Science at the University of Warwick in 2014.

The work presented was carried out by the author except in the cases outlined below:

- Add Health dataset used in Chapter 3 was provided by the Add Health study of the University of North Carolina at Chapel Hill,
- Agincourt HDSS dataset used in Chapters 4 and 5 was provided by the Medical Research Council / University of the Witwatersrand Rural Public Health and Health Transitions Research Unit,
- MODIS/Terra sensor satellite image dataset used in Chapter 5 was provided by the Land Process Distributed Active Archive Center,
- Figure 4.1 was reprinted from Kahn et al. (2012) with permission from Oxford University Press,
- The elicitations described in Chapter 5 were performed by Rhian Twine of the

Agincourt research unit following a process designed and instructions written by the author.

Parts of this thesis are published as:

- Eyre RW, House T, Hill EM, Griffiths FE. (2017) Spreading of components of mood in adolescent social networks. *R. Soc. open sci.* **4**: 170336. `http://dx.doi.org/10.1098/rsos.170336`
- Eyre RW, House T, Gómez-Olivé FX, Griffiths FE. (2018) Modelling fertility in rural South Africa with combined nonlinear parametric and semi-parametric methods. *Emerg. Themes Epidemiol.* **15**:5. `https://doi.org/10.1186/s12982-018-0073-y`

Optimisations in Chapter 3 and section 4.3 were performed using the GenSA package in the R software. Gaussian process regression was performed in Section 4.3 using an author-written code involving the GenSA package in the R software, which is available to view at `https://github.com/RWEyre/fertility_rates_example`. Gaussian process classification in Section 4.4 was performed using the GPy package in the Python software. Bayesian network structure learning and inference in Chapter 5 were performed using the gRain and bnlearn packages in the R software, and the d-separation analysis was performed using the pgmpy package in the Python software.

There are no wrong turnings.
Only paths we had not known we were meant to walk.

GUY GAVRIEL KAY

Abstract

The statistical inference of socio-economic variables in public health is key to the design of interventions to address the many health inequalities that exist across the world. However, such inferences are achieved commonly using a small standardised library of statistical methods. Meanwhile other fields such as computer science and systems biology have seen the development of many new methods allowing for more varied and useful analyses.

Here we present analyses in three important contextual areas of socio-economic variables in public health, bringing in modern and sophisticated methods in order to develop highly useful and flexible results and further expand the library of statistical methods in public health.

In the first, we further develop and apply a non-linear temporal model to analyse the spread of health aspects such as mood and weight over US adolescent friendship networks by a process known as *social contagion*. The use of this model improves our ability to more realistically reflect patterns we expect to see result in the data from contagion. This was achieved using analysis of the Add Health dataset.

In the second, we use the flexibility and complex features of Gaussian processes to analyse two different aspects of pregnancy in rural South Africa using the Agincourt HDSS dataset. First, the modelling of fertility-patterns over combinations of variables where some have established models and others do not, allowing us to incorporate such variables into our model without risking the enforcement of unjustified assumptions. Second, analysing social contagion of pregnancy risk behaviour where no social network data exists, demonstrating how the use of sophisticated methods can enable us to attempt complicated research questions.

Finally, in the third we build three possible Bayesian belief network models of household food security in the Agincourt study area. The structural features of these models make them potentially highly useful causal tools that enable us to model a wide range of interventions on our system.

Through these analyses we demonstrate the importance of expanding the library of statistical methods in public health to include the many modern and sophisticated methods being developed in other fields, whilst also producing findings and tools of great robustness, flexibility, and utility.

CHAPTER 1

Introduction

1.1 Socio-economic variables in public health

Public health is core to ensuring the health and welfare of a population, standing alongside other key areas of government policy such as education. According to the Centre for Disease Control (CDC), public health is defined as the "science of protecting and improving the health of families and communities through promotion of healthy lifestyles, research for disease and injury prevention and detection and control of infectious diseases" (CDC, 2017). This enables governments and international organisations like the World Health Organisation (WHO) to try and improve overall health at a population level. Through research and evidence based policies and interventions, public health has made great achievements in vaccinations, infectious disease control, motor vehicle safety, reduced child mortality, prevention and treatment of HIV/AIDS, and so much more (CDC, 1999; CDC, 2011a; CDC, 2011b).

Socio-economic variables, otherwise known as social determinants, are assuming an ever increasing importance in how public health tackles health problems throughout society. Such variables are usually thought of as "any non-medical factors influencing health, including health-related knowledge, attitudes, beliefs, or behaviour" (Braveman et al., 2011). Even as far back as the 1800s, many saw a correlation between the socio-economic status of individuals and their health. Biologists such as Rudolf Virchow are reported to have said that "medical science is intrinsically and essentially a social science" (Pellegrini Filho, 2011). In 1948, the constitution of the WHO recognised that health involves the social wellbeing of the individual, as well as the physical and mental (Pellegrini Filho, 2011).

Though knowledge of the social impact on health has existed for quite some time, it is only in the last decade or so that is has come in to focus within the aims of public policy and research. Social aspects such as the eradication of poverty, universal education, reduced inequality, and economic growth form core parts of both the United Nations Millennium Development Goals and their more recent Sustainable Development Goals (UN, 2015a; UN, 2015b). Most revealing of all is the establishment of the Commission on the Social Determinants of Health by the WHO, which delivered its final report in 2008 where it presented evidence of a social gradient in health that exists throughout all countries of the world (WHO, 2008). The addressing of this social gradient has become a central focus of public policy for many countries, such as both Sweden (Hogstedt, 2004) and the UK (Acheson et al., 1998).

Many different socio-economic aspects of our lives have been shown to impact on our health, including our class and race (Bleich et al., 2012); levels of social support and capital (Phongsavan et al., 2006); education (Cutler and Lleras-Muney, 2006); whether we have a job and the quality of that job (Barbeau et al., 2004; Bartley and Plewis, 2002); wealth (Pollack et al., 2007); and health knowledge (Coveney, 2005). In many health measures there are obvious disparities between different kinds of populations. For instance, the life expectancy of Japan was 81.9 years in 2005, whilst at the same time was only 34 years in Sierra Leone (Marmot, 2005). Many have argued that this therefore forms an issue of social justice (Marmot, 2005).

Considering the relative youth of active research in this area (despite how far back knowledge of its importance goes), we are still at a stage where evidence must be gathered to help inform the development of interventions and policies to address the inequalities in health across different socio-economic variables. Of great importance is the gathering of quantitative evidence to ensure that we can distinguish signals from noise in dealing with population level effects. Due to the nature of the variables involved, which include the existence of social support networks, employment, and wealth, much of this evidence must be inferred from observational data rather than experimentation which would prove both costly and possibly even unethical. However, inference from observational data comes with many difficulties, particularly in the inference of causality. As such, it is of utmost importance that we establish a library of statistical methods that can make these inferences with the greatest level of robustness, flexibility, and usefulness possible.

1.2 Statistical modelling in public health

The state of statistical methodology in the realms of public health, epidemiology, and medicine is that of a high level of standardisation. A small set of accepted methods are typically used for analysis of observational data. These methods form the core recommendations of medical statistics textbooks (Munro, 2005; Armitage et al., 2008; Kirkwood and Sterne, 2010; Petrie and Sabin, 2009; Zhou et al., 2009; Clayton and Hills, 2013; Peat and Barton, 2008), and see the most frequent use in journals of public health (Levy and Stolte, 2000; Nitta et al., 2010; Hayat et al., 2017) as well as in other medical disciplines (Brailsford et al., 2009; Horton and Switzer, 2005; Rigby et al., 2004).

Most of these methods are of three basic types: summarising data, hypothesis testing, and generalised linear modelling. Summarising data usually involves the calculation of summary statistics, or providing graphical plots of the data, and then drawing conclusions from what is directly observed in that particular data sample. Hypothesis testing is usually a dichotomous analysis of different possible realities. One in which the effect of interest (such as a medicine being effective) occurs, and one in which it does not. Data is gathered, and some suitable statistic of the data related to the effect of interest (such as the mean rate of symptom reduction in patients receiving the medicine) is measured. A p-value is then calculated from this statistic, in order to ascertain the probability of such a value for the statistic (or a more extreme one) occurring in one of the two possible realities. If the p-value crosses some chosen level, then it is deemed statistically significant, and is offered as evidence that the data sampled is unlikely to have occurred in the tested reality. Neither of these are strictly speaking models in the more traditional mathematical sense, though hypothesis testing is sometimes performed on models.

The most common form of actual modelling is done using generalised linear models (GLMs). Here the relationship of some unknown variable of interest (such as whether the medicine is effective or not) with certain known variables (e.g. height, weight, etc.) is modelled as being linear using a parameterised linear equation (which is described in more detail in Section 2.1). Often the values of the parameters of these models (which measure the association between the known variables and the unknown variable), or simply that they are non-zero, is what is of interest. As such, these are also subjected to hypothesis tests to try and ascertain evidence for them being non-zero. Use of GLMs are widespread, and they form the standard for statistical modelling throughout health statistics. They are even commonly used to take care of quite subtle and complex modelling issues such as causational confounding (Müllner et al., 2002).

The hypothesis testing side of this methodology has already seen great and widespread criticism. Particularly with the use of p-values, due to many reasons including that significant results can be discovered purely by chance as standard significance levels are often associated with high false positive rates (Cumming, 2014; Greenland et al., 2016; Goodman, 1999; Colquhoun, 2014). This, alongside issues related to practical research standards, has lead to the questioning of the validity of much medical research (Ioannidis, 2005; Leek and Peng, 2015). In addition, criticism has been levelled at the dichomotous nature of hypothesis tests, arguing that estimates of effect size rather than just the existence of an effect are what is really needed (Wilkinson, 2014). This stands as an effective argument for why modelling rather than testing can lead to higher utility. However, when effect size is being estimated it has been shown that great difficulties can arise if the real effect size is small. Attempting to estimate effects that are in reality close to zero can lead to estimation errors where the size of the effect can be grossly overestimated, or the sign of the effect can be wrongly estimated (Gelman and Weakliem, 2009).

Some of these issues already hint to the particular issue that we are concerned with in the research presented here. Namely the lack of use of more modern and sophisticated methods in modelling socio-economic variables in public health, that are better suited to capturing the nuances of the underlying system producing the data. Despite what link function is used in a GLM, and whether any of the independent variables are put under non-linear transforms, there is still an assumption of additive linearity built in. Due to the many subtle issues that can plague inference from observational data, relying on linear models can sometimes be far too restricting. Often a linear model is all that is needed, but when the question being asked comes with great levels of complexity from the system of which it is being asked, relying on a linear model can lead to issues that severely undermine the robustness of the results. On top of this, linear models come with only a certain level of flexibility, whilst other models may enable us to ask questions we could not with more standard methods.

Some more modern methods from areas such as machine learning have made their way to use in health science, but are still used sparsely and largely in only certain areas such as medical diagnosis (Crown, 2015; Kononenko, 2001; Altman, 2000). Effort has been made to design more robust versions of the standard methods (Erceg-Hurn and Mirosevich, 2008; Farcomeni and Ventura, 2012) but these do not answer the issue we attempt to establish here about the limitations of these methods.

1.3 Research aims

The core aim of this research is to utilise more modern methods better suited to more complex problems (that have seen use in areas such as machine learning, computer science, and systems biology) in selected areas of analysis of socio-economic variables in public health where such methods have seen much less use. In so doing, we aim to increase the flexibility of the models and the robustness and utility of the findings and research performed above the limitations that have been found in using linear models. We perform these analyses in different areas with different purposes, seeking to show various different ways where these more complex models can improve upon the traditionally used linear models. In addition to this, either finding results of greater robustness or building tools of greater flexibility and therefore providing greater usefulness to the public health literature on socio-economic variables.

In Chapter 2, we introduce two mathematical methods from outside the standard methodology that we use throughout many of the analyses performed here. These are the semi-parametric Bayesian methods of Gaussian process regression and classification, as well as Bayesian belief network modelling. We give basic explanations in order to aid the completeness and understanding of the work presented here. We also give a brief overview of the GLM methodology used commonly in most public health research.

In Chapter 3, we examine the spread of behaviours such as weight changes and mood over US adolescent friendship networks in a process known as *social contagion*. In doing so, we hope to first find results that pass the extreme criticisms those from past studies have had (which we describe in the chapter), and second show how the use of non-linear parametric models can be much more robust in causal inference than linear models due to the greater assumptions built in to them.

In Chapter 4, we look at two possible analyses, related by the context of pregnancy in rural South Africa and the use of Gaussian processes. First we employ a combination of non-linear parametric and semi-parametric methods in the estimation of fertility patterns over different socio-economic variables, arguing that the use of flexible non-linear methods like Gaussian processes can overcome the issue of having to make unjustified a-priori assumptions of linear relationships between these variables. Second we take advantage of certain features of Gaussian process classification to examine the possibility of social contagion of adolescent pregnancy risk behaviour in a data set that contains no social network data, therefore demonstrating how the flexibility of non-linear complex methods can allow for a much greater range of possible research into questions we would find difficult to answer by relying on linear methods.

In Chapter 5, we build a causal Bayesian belief network model of household food security in rural South Africa. Here we aim to employ community involvement in the model building process, and seek to build a tool that can be of great use in future research for this community and communities like it. By doing this, we also aim to show the greater utility of complex models that involve probabilistic indirect relationships and built in directional structure rather than simply providing single measures of direct effects.

Finally, in Chapter 6, we argue for the status of public health as what is known as a complex system, due largely to its emphasis on population level behaviour, and how this forms a core justification for this research. We summarise and discuss the advantages in the methods we have used, from the methodological advantages over the linear models that have come before to the significance of the findings, also considering the importance of trying to involve those who will be most affected by the research in the design and performance of the research. Then we examine the limitations of the analyses we performed, both in terms of more practical limitations as well as more subtle issues such as the rise of more powerful black box models and the issues involved in causal modelling. At the end, we demonstrate how we have met the aims described here, what potential there is for future work, and why in addition to all the main benefits this type of research also demonstrates the continuing need for collaboration between mathematical and social scientists.

Overall we hope to demonstrate the need for greater use of modern and sophisticated statistical methods in what is in most areas of the world an urgent area of health research. We do not aim to fully supplant the standard methodology, but rather to improve in areas where the methodology can be shown to be unsuitable.

CHAPTER 2

Mathematical preliminaries

Here we present the core mathematical methods that will see widespread mention in several chapters. The first is a very brief description of generalised linear models, as mentioned in Section 1.2. The second is Gaussian processes which we use in both Chapter 3 and Chapter 4. The last is Bayesian belief networks that we use in Chapter 5.

2.1 Generalised linear models

Generalised linear modelling (GLM) involves the modelling of some variable Y that is of great interest as a random variable with expectation $E\left(Y \mid \boldsymbol{X}\right)$ linearly dependent on a set of independent variables $\boldsymbol{X}$. The mathematical form of this model is then

$$E\left(Y \mid \boldsymbol{X}\right) = g^{-1}\left(\boldsymbol{X}^T\boldsymbol{\beta}\right) \tag{2.1}$$

where g is some link function related to the nature of Y, and $\boldsymbol{\beta}$ are a set of parameters governing by how much a change in each independent variable (when all others are held fixed) results in a change in the expectation of Y. These parameters are therefore related to the correlation between each independent variable and Y given all the other independent variables.

There are many different variations of GLM, each defined by a particular link function g. Standard linear regression models an infinite and continuously valued $Y \in \mathbb{R}$ as simply just a linear combination of the independent variables $\boldsymbol{X}$, and therefore has link function

$$g\left(y\right) = y \tag{2.2}$$

i.e. the identity function.

Another common model is logistic regression, which models a binary valued $Y \in \{0, 1\}$ on the independent variables $\boldsymbol{X}$ using the logit function as a link function

$$g\left(y\right) = \log\left(\frac{y}{1-y}\right) . \tag{2.3}$$

Another common form for Y is when it forms count data, where $Y \in \mathbb{N}_0$ (i.e. the set of non-negative integers). Here we use Poisson regression, which has link function

$$g\left(y\right) = \log\left(y\right) . \tag{2.4}$$

There are many other variations, all dependent on the the nature of Y. Even for the ones mentioned here, there are sometimes alternative link functions such as the inverse probit function $\Phi^{-1}\left(y\right)$ for logistic regression. However, at their core they all model the expectation of the variable of interest as some function of an additive linear combination of the independent variables. It is possible to further complicate these models by inserting simple non-linear functions of the independent variables, such as non-parametric functions of individual variables or products of two or more of them, as effectively new variables in the linear equation. For example,

$$E\left(Y \mid \boldsymbol{X}\right) = \beta_0 + \beta_1 \log\left(X_1\right) + \beta_2 X_2 X_3 \, . \tag{2.5}$$

This still requires no further parameterisation beyond the additive linear parameters, which is essentially what constrains the usefulness of GLMs. They are still of great use in many situations, but many systems being analysed are complicated enough such that additive linear parameterisation can not effectively aid the finding of useful inferences. More sophisticated methods, such as those described in the rest of this chapter, are necessary.

2.2 Gaussian processes

Gaussian process regression and classification are semi-parametric Bayesian methods that make only minimal assumptions about the relationships being modelled, such as smoothness and continuity, letting the data dictate the shape of said relationships.

An indepth explanation of the method is given in Rasmussen and Williams (2006). Here we give a brief simplified explanation to aid the understanding of the models used in later chapters.

2.2.1 Regression

Gaussian process regression is a method that aims to find a distribution over functions $f(\cdot)$ that relate a set of covariate observations $X = \{\mathbf{x}_i\}$ to a set of dependent variable observations $\mathbf{y} = \{y_i\}$ by $y_i = f(x_i) + \epsilon_i$ where ϵ_i is Gaussian noise (Figure 2.1). By incorporating data we can calculate a posterior distribution of possible functions, where predictions of new function values $\mathbf{f}^*$ for new observations with covariates X^* can be drawn from the posterior predictive distribution

$$\mathbf{f}^* | X^*, X, \mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Sigma}) \tag{2.6}$$

where

$$\boldsymbol{\mu} = \boldsymbol{K}(X^*, X)\left(\boldsymbol{K}(X, X) + \sigma_n^2 \boldsymbol{I}\right)^{-1} \mathbf{y} \tag{2.7}$$

$$\boldsymbol{\Sigma} = \boldsymbol{K}(X^*, X^*) - \boldsymbol{K}(X^*, X)\left(\boldsymbol{K}(X, X) + \sigma_n^2 \boldsymbol{I}\right)^{-1} \boldsymbol{K}(X, X^*) \tag{2.8}$$

where the noise variance σ_n, acting on the identity matrix $\boldsymbol{I}$, accounts for the noise in the data. The best prediction, and therefore the typically chosen function, comes from the mean of the distribution.

The method is called semi-parametric as we do not get a parametric relationship between the dependent and independent variables as a result, but instead parameters are used to define the covariance function $\boldsymbol{K}(\mathbf{x}_i, \mathbf{x}_j)$. In so doing we let the data dictate the shape of how the distribution of $f(x)$ varies over x, whilst being able to assert assumptions about such things as the level of smoothness and the existence of trends and periodicities. There are many different covariance function to choose from, but a standard choice is the squared exponential covariance function, which results in a smooth and continuous relationship between the dependent and

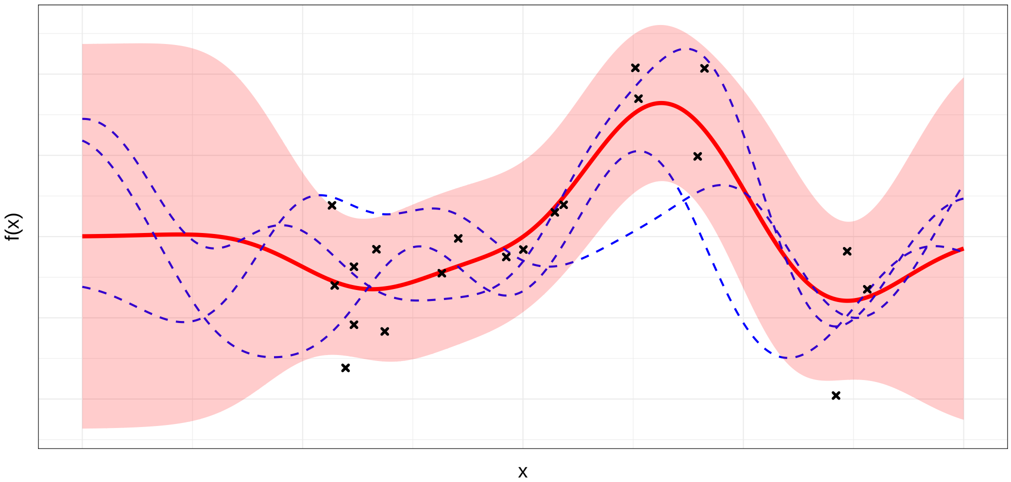


Figure 2.1: Example of a Gaussian process regression. The regression is performed on a set of noisy data (black crosses). The resulting distribution is Gaussian with a mean shown by the red line and 95% posterior intervals shown by the red shaded region. Three possible other functions for the regression drawn from this distribution are shown by the blue dashed lines.

independent variables, and is defined as

$$\boldsymbol{K}_{\mathrm{SE}}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sigma_f^2 \exp\left[-\frac{1}{2}\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \boldsymbol{M}\left(\mathbf{x}_i - \mathbf{x}_j\right)\right] \tag{2.9}$$

where $\mathbf{x}_i$ is the covariate vector for observation i, and the parameters of our covariance function are the signal variance σ_f^2, which governs the size of the covariance between pairs of observations, and $\boldsymbol{M} = \operatorname{diag}\left(\mathbf{l}\right)^{-2}$ where $\mathbf{l}$ is the vector of length parameters (one length parameter for each covariate). The length parameter for a particular covariate essentially governs how much our function varies over that covariate. For a small length parameter, $f\left(x\right)$ would vary greatly over x, and for a large length parameter the relationship would be almost flat. This essentially implements automatic relevance detection, where fitting the covariance function parameters will result in large length parameters for variables that have little relevance to $f\left(x\right)$ therefore forcing $f\left(x\right)$ to be (close to) constant over those variables (Figure 2.2) (Rasmussen and Williams, 2006).

Other potential covariance functions that result in similar smoothly varying functions to the squared exponential, but with greater flexibility, include the Matérn

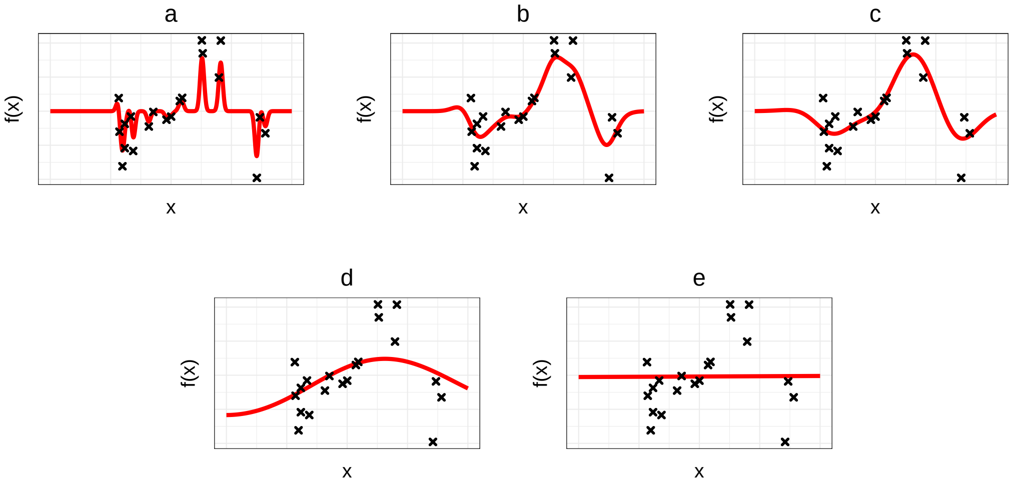


Figure 2.2: Gaussian process regression mean functions generated from the same dataset (shown as black crosses) with different lengthscale values of 0.1 (a), 0.5 (b), 1 (c), 5 (d), 100 (e). Note how the functions get smoother and flatter as the lengthscale increases. At the extreme value of $l = 100$ the function is completely flat, whilst at the other extreme of $l = 0.1$ it is clearly overfitting to the data.

and rational quadratic covariance functions. The Matérn has the form

$$\begin{aligned} \boldsymbol{K}_{\mathrm{M}}\left(\mathbf{x}_i, \mathbf{x}_j\right) =& \sigma_f^2 \frac{2^{1-\nu}}{\Gamma\left(\nu\right)} \left(2\nu \left(\mathbf{x}_i - \mathbf{x}_j\right)^T \boldsymbol{M} \left(\mathbf{x}_i - \mathbf{x}_j\right)\right)^{\nu/2} \times \\ & I_\nu \left(\sqrt{2\nu \left(\mathbf{x}_i - \mathbf{x}_j\right)^T \boldsymbol{M} \left(\mathbf{x}_i - \mathbf{x}_j\right)}\right) \end{aligned} \tag{2.10}$$

where $\Gamma\left(\cdot\right)$ is the Gamma function, $I_\nu\left(\cdot\right)$ is the modified Bessel function, and the positive parameter ν dictates the noisiness of the functions drawn from the resulting distribution. In the limit $\nu \to \infty$ the Matérn function becomes the squared exponential, though the cases of $\nu = 3/2$ and $\nu = 5/2$ are usually considered the most interesting and highly used. The rational quadratic has the form

$$\boldsymbol{K}_{\mathrm{RQ}}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sigma_f^2 \left(1 + \frac{\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \boldsymbol{M} \left(\mathbf{x}_i - \mathbf{x}_j\right)}{2\alpha}\right)^{-\alpha} \tag{2.11}$$

where positive parameter α gives extra control over the smoothness of the functions drawn from the resulting distribution. The rational quadratic is an infinite sum of squared exponential functions at different lengthscales. In the case $\alpha \to \infty$, similar to the Matérn function, it tends to a squared exponential with lengthscale matrix $\boldsymbol{M}$. Other covariance functions, such as the periodic, linear, and Brownian covariance functions, can also introduce features like periodicities and trends therefore making further assumptions beyond just the smoothness and continuity assumed by the above covariance functions. The values of the parameters of the covariance function can be found using maximisation of the marginal log likelihood, more details of which can be found in Rasmussen and Williams (2006).

2.2.2 Classification

Gaussian process regression assumes that the noise in the observations $\mathbf{y}$ is Gaussian distributed, which naturally requires the range of values for each y to be infinite and continuous. However, even if this is not the case, like for binary classification where each $y \in \{0, 1\}$, it is still possible to use Gaussian processes to model the dependent variables in terms of the covariates. In this case we model the outcome, for instance the classification probability $\pi(x) = \Pr(y = 1 \mid x)$, indirectly by performing a Gaussian process regression on a latent function f related to the outcome via a link function σ, i.e.

$$\pi(x) = \sigma(f(x)) \tag{2.12}$$

where in the case of binary classification σ is usually taken to be a logistic or probit function. The posterior predictive distribution for the class variable is then

$$\begin{aligned} \hat{\pi}(x^*) &= \Pr(y^* = 1 \mid X, \mathbf{y}, X^*) \\ &= \int \sigma(f^*)\, p(f^* \mid X, \mathbf{y}, X^*)\ df^* \end{aligned} \tag{2.13}$$

where the predictive posterior distribution of the latent function f is given by

$$\begin{aligned} p(f^* \mid X, \mathbf{y}, X^*) &= \int p(f^* \mid X, X^*, \mathbf{f})\, p(\mathbf{f} \mid X, \mathbf{y})\ d\mathbf{f} \\ &= \int \frac{p(f^* \mid X, X^*, \mathbf{f})\, p(\mathbf{y} \mid \mathbf{f})\, p(\mathbf{f} \mid X)}{p(\mathbf{y} \mid X)}\ d\mathbf{f}. \end{aligned} \tag{2.14}$$

As the likelihood $p(\mathbf{y} \mid \mathbf{f}) = \sigma(\mathbf{f})$ is non-Gaussian, this integral on the bottom line of (2.14) is analytically intractable. However, various methods exist to approximate the integral in order to find a solution, such as the Laplace approximation and expectation propagation (Rasmussen and Williams, 2006). The Laplace approximation is a method in which the posterior $p(\mathbf{f} \mid X, \mathbf{y})$ is approximated by a Gaussian distribution calculated from the second order Taylor expansion of the log posterior about the maximum of the posterior

$$q(\mathbf{f} \mid X, \mathbf{y}) \propto \exp\left(-\frac{1}{2}\left(\mathbf{f}-\hat{\mathbf{f}}\right)^T H \left(\mathbf{f}-\hat{\mathbf{f}}\right)\right) \tag{2.15}$$

where the posterior maximum $\hat{\mathbf{f}}$ and the negative log posterior Hessian

$$H = -\nabla^2 \log p(\mathbf{f} \mid X, \mathbf{y}) \mid_{\mathbf{f}=\mathbf{f}^*} \tag{2.16}$$

can be calculated analytically, enabling us to approximate the integral in (2.14) and make predictions (see Rasmussen and Williams (2006) for more detail).

2.3 Bayesian belief networks

2.3.1 Introduction

As with Gaussian processes, an in-depth description of Bayesian belief networks can be found in many other sources (Murphy, 2012; Korb and Nicholson, 2003; Koller and Friedman, 2009). Here we give a short introduction to them, followed by a brief description of the concepts of d-separation and inference using Bayesian networks which will be of importance during the research in Chapter 5.

In purely mathematical terms, a Bayesian network is a probabilistic graphical model in the form of a directed acyclic graph (DAG) representing a particular factorisation of the joint probability of the system (Figure 2.3). Essentially each set of in-going links in the network going in to a particular node represents a factor in the joint distribution that is the marginal conditional distribution of the variable that node represents. The distribution is conditioned upon the nodes that those links are outgoing from. These nodes are referred to as the 'parents' of the node of interest, and the node of interest is a 'child' of those nodes. The joint distribution of the system is then given by

$$\Pr(\boldsymbol{X}) = \prod_i \Pr(X_i \mid \mathbf{pa}(X_i)) \tag{2.17}$$

where $\boldsymbol{X}$ is the set of variables defining our system, X_i is one of these variables, and $\mathbf{pa}(X_i)$ are the set of parents of that variable.

Within public health, Bayesian networks are often used to give a causal representation of a system (or at the very least a representation of a set of causal beliefs about the system often obtained from experts on the particular system). We define

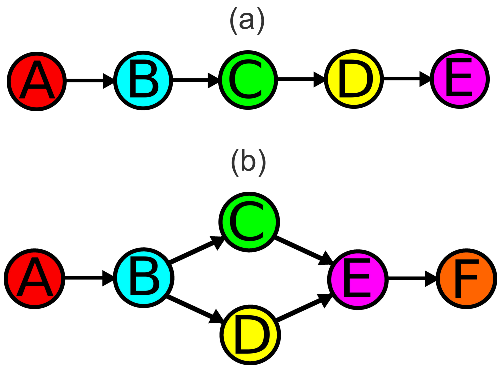


Figure 2.3: Graphical representations of specific factorisations of the joint distribution over the variables A, B, C, D, E, and F. (a) $\Pr(A, B, C, D, E) = \Pr(E \mid D) \Pr(D \mid C) \Pr(C \mid B) \Pr(B \mid A) \Pr(A)$. (b) $\Pr(A, B, C, D, E, F) = \Pr(F \mid E) \Pr(E \mid C, D) \Pr(C \mid B) \Pr(D \mid B) \Pr(B \mid A) \Pr(A)$.

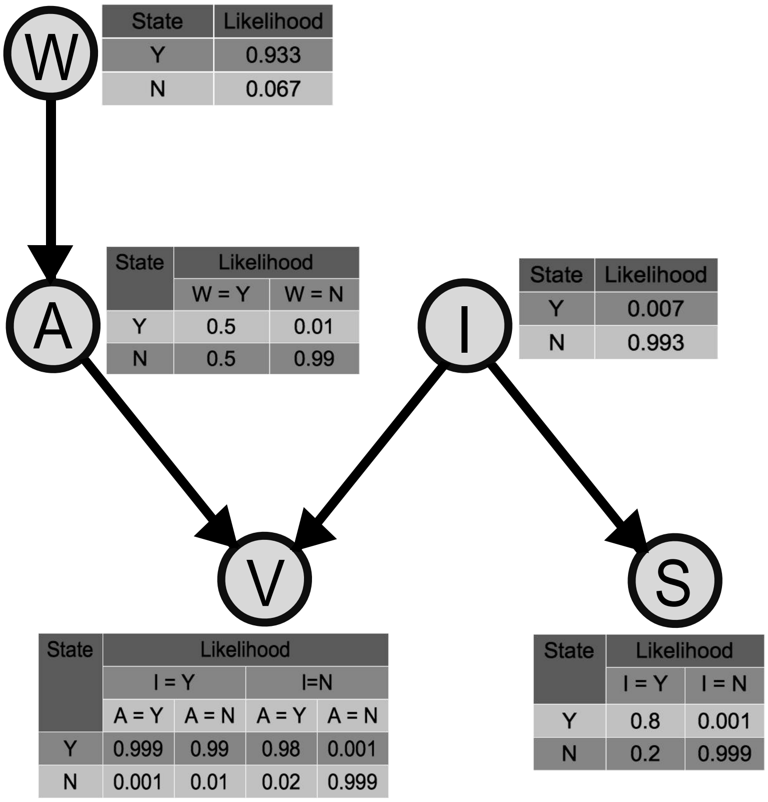
our system by a set of variables, and then a link going from variable A to variable B implies that A causes B (as well as that B is directly probabilistically dependent on A). For example, consider we have a small dog called Merlin. As Merlin is a greedy dog, sometimes when he is taken out on a walk he will eat something off the floor that is bad for him. This can potentially cause him to vomit later. There is always some chance that he may not be walked on a particular day, and even if he is walked he may not find anything to eat. If he is not walked, he may find something accidentally left on the kitchen floor to eat that could also be bad for him. Potentially he may be ill, which could also lead to him vomiting despite not eating anything bad for him. If he is ill then he may also show other symptoms. These relationships are encoded in a Bayesian network shown in Figure 2.4.

As we have described, each of the possible states of each variable are inherently probabilistic in their possibility. We encode these probabilities as conditional probability tables (CPTs) for each variable. Each entry in a table states the probability of a given state of that variable given the particular states its parents are inhabiting. Which variables are linked to which other variables forms the structure of our model, and the conditional probability tables form the parameters of our model, i.e.

$$\theta_i^{jk} = \Pr\left(X_i = j \mid \mathbf{pa}\left(X_i\right) = \boldsymbol{k}\right) . \tag{2.18}$$

Note that for Bayesian network model parameters we signify nodes by lower indices and states (or node and parent state combinations) by upper indices.

As we can see in our example of Merlin the dog, we are now able to ask questions



State	Likelihood
Y	0.933
N	0.067

State	Likelihood	
	W = Y	W = N
Y	0.5	0.01
N	0.5	0.99

State	Likelihood
Y	0.007
N	0.993

State	Likelihood			
	I = Y		I=N	
	A = Y	A = N	A = Y	A = N
Y	0.999	0.99	0.98	0.001
N	0.001	0.01	0.02	0.999

State	Likelihood	
	I = Y	I = N
Y	0.8	0.001
N	0.2	0.999

Figure 2.4: Bayesian belief network for the “Merlin the dog” example. Each of the variables has two states, yes (Y) or no (N). W - Merlin was walked. A - Merlin ate something bad for him. I - Merlin is ill. V - Merlin has vomited. S - Merlin has exhibited other illness symptoms (e.g. temperature, diarrhoea etc). Each node is accompanied by their conditional probability table, showing the likelihood of each state dependent on the parents of that node (if any).

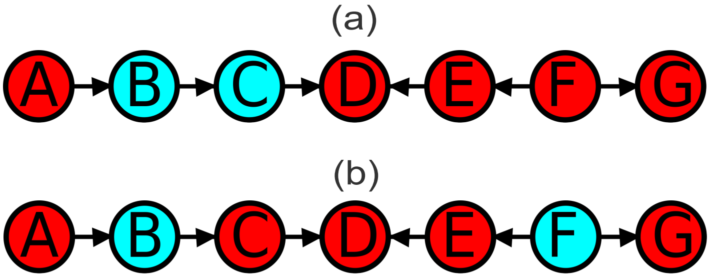


Figure 2.5: Examples of d-separation relationships. (a) $A \perp G \mid\mid \{B, C\}$. (b) $A \perp G \mid\mid \{B, F\}$.

such as "if we can prevent him from eating something he should not, how much can we reduce the likelihood of him making a mess of our floors?" and "if Merlin is ill and vomits, what is the likelihood that he has also eaten something he should not?". These types of queries rely on concepts of conditional independence which we describe next.

2.3.2 D-separation

As mentioned before, one factor that makes a Bayesian belief network suitable for modelling causal systems is how they are able to model conditional independencies between variables where conditioning on one or more variables renders two or more other variables probabilistically independent of each other, which we will write as

$$A \perp B \mid C \tag{2.19}$$

when A is independent of B given C.

There are three basic structures in Bayesian networks involving triplets of linked nodes where conditioning on one node in particular renders the other two either independent or dependent of each other depending on the structure type. Each of these are exhibited in our example of Merlin the dog (Figure 2.4).

The simplest is a chain (or causal chain) shown by the triplet $W \rightarrow A \rightarrow V$. Here V, whether Merlin vomits or not, is indirectly dependent on W, whether he has been walked or not. However, if we condition on A then $V \perp W \mid A$ as knowing whether he has been walked or not supplies no extra information to whether he will vomit or not if we already know he ate something he should not have done.

The second is a fork (or common cause) shown by the triplet $V \leftarrow I \rightarrow S$. Here if we do not know whether Merlin is ill, I, then knowing whether he exhibits further symptoms, S, affects our belief of whether he will vomit or not, V, as it affects our belief of whether he is ill in the first place. If we condition on I then $V \perp S \mid I$ as if we know he is ill then we do not need to infer the possibility of him vomiting from the other symptoms he exhibits.

The final structure is a collider (or common effect) shown by the triplet $A \rightarrow V \leftarrow I$. Here there is no effect from I on A until we condition on V, so we find $A \not\perp I \mid V$. If we know Merlin has definitely vomited, then finding out he is ill impacts on the probability of him having eaten something he should not have eaten (for example we may find it reduces the likelihood he ate something as the illness explains the vomiting, and so 'explains away' the possibility of him having eaten something bad).

What is more useful is being able to find conditional independencies between nodes far away from each other in the network. This can be done using the concept of d-separation (directional separation), which we will use later as a way to validate our Bayesian belief network model structure in Chapter 5. Two nodes are said to be d-connected if at least one path between them does not have a collider on it, such that there are no independencies on the path effectively 'blocking' the 'flow of information' via the probabilistic relationships between the nodes. If all paths are 'blocked' then the nodes are d-separated. A node A is d-connected to a node B given some other set of nodes C ($A \not\perp B \mid\mid C$) if any paths with a collider on them pass through C, i.e. by conditioning on C we 'unblock' the paths between A and B, and are d-separated by C otherwise ($A \perp B \mid\mid C$) as conditioning on C 'blocks' (or fails to unblock) the paths between them. Two examples of d-separation are shown in Figure 2.5.

Though d-connectedness does not necessarily imply dependence, d-separation always implies independence. Therefore by finding all the d-separation relationships in the network we can also find all the independencies (conditional and otherwise) in the network. This can be useful for both validating against expectations of how the system should act and for finding out previously unknown independencies in the system. Efficient algorithms for finding d-separation relationships can be found elsewhere (Koller and Friedman, 2009; Smith, 2010).

2.3.3 Inference

In order to answer queries like those detailed in Section 2.3.1, we need to perform inferences on our network. This gives the greatest mathematical advantage of

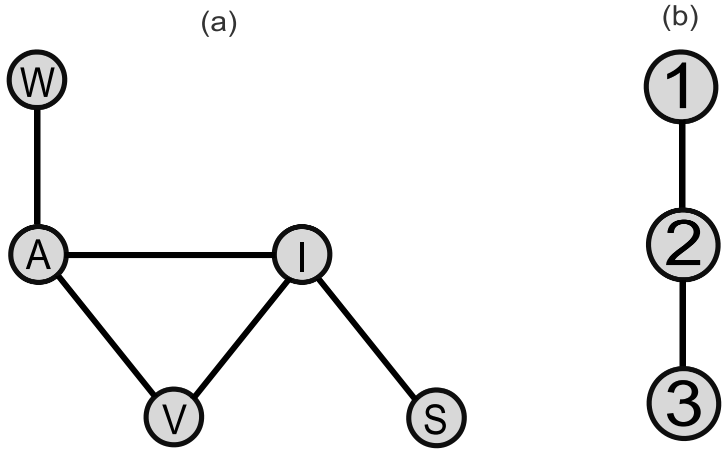


Figure 2.6: Stages of inference from the "Merlin the dog" example (Figure 2.4). (a) The moralized graph (see Section 2.3.3), providing an equivalent undirected graph to the Bayesian network. (b) The junction tree graph, where each node represents a clique in (a). Clique 1 contains W and A, 2 contains A, I, and V, and 3 contains I and S.

graphical representations of joint distributions such as a Bayesian network. Typically to answer a query such as "what is the likelihood Merlin will vomit given he ate something he should not have eaten" (from the network in Figure 2.4) requires calculations such as

$$\Pr\left(V \mid W = Y\right) = \frac{\sum_{a \in \mathbb{S}_A} \sum_{i \in \mathbb{S}_I} \sum_{s \in \mathbb{S}_S} \Pr\left(V, W = Y, A = a, I = i, S = s\right)}{\Pr\left(W = Y\right)} \tag{2.20}$$

where $\mathbb{S}_X$ is the set of states of variable X. In this calculation, enumeration of the joint distributions can require an exponentially large number of calculations depending on the number of nodes and states involved.

The advantage of a Bayesian network is that we can use the structure of the network to simplify these calculations. There are many different small variations on the algorithms used to achieve this, typically coming under the heading of belief propagation and variable elimination (Murphy, 2012; Koller and Friedman, 2009). In our inferences in Chapter 5 we use the Lauritzen-Spiegelhalter (LS) algorithm (Lauritzen and Spiegelhalter, 1988), as implemented by the gRain package in the software R (Højsgaard, 2012).

The LS algorithm is simply a specialised variation of belief propagation. When performing inferences on a Bayesian network, we take advantage of the specific factorisation of the joint distribution of the system that the network structure implies in order to simplify the calculations involved in the marginalisation necessary. This is achieved by summing out the different variables in a specific order that reduces down the number of calculations needed as much as possible, which is a process known as variable elimination.

Belief propagation is a specific form of variable elimination performed on tree-like networks. The tree structure defines a specific ordering to the marginalisation that results in the smallest number of calculations possible. We start at the leaf nodes and enact a so-called 'message passing' process where the factors in the joint distribution associated with each node are summed over and passed on to the next node until we reach the root node. To introduce evidence, or condition on a variable being in a particular state, we simply disclude from the summations all contributions where the observed variables are not in their observed states (and then normalise the result at the end of the calculations).

In the LS algorithm, in order to perform the belief propagation we first need a tree structure equivalent to our network. This is done by finding the clique tree network (Figure 2.6). We first moralise the graph, which finds the equivalent undirected network to our directed network by forming links between unlinked parent nodes with common child nodes and then removing the directions. Then the network is triangulated by placing links within loops such that the largest loops with no links crossing through them only contain three nodes. Finally, we find the set of smallest possible cliques (every smallest set of nodes fully linked to each other) and perform belief propagation on the graph formed by these cliques. The LS algorithm uses a variant of belief propagation that finds a factorisation of the joint distribution of our system in terms of the marginal distributions of the cliques, further simplifying and speeding up the calculations needed to perform inferences.

CHAPTER 3

Non-linear parametric modelling of social contagion

3.1 Social contagion of mood and weight changes

Depression and other associated mood disorders, as well as obesity and being overweight, form an increasing burden upon the health of modern society. The World Health Organisation estimates that 350 million people are affected by depression throughout the world, leading to morbidity through a reduced ability to work and socialise, as well as mortality due to suicide (WHO, 2014). Meanwhile, being overweight has been associated with greater incidence of comorbidity such as cardiovascular disease, type 2 diabetes, and osteoarthritis (Must et al., 1999). Between 1976 and 2010, prevalence of obesity amongst children and adolescents in the United States almost tripled (Fryar et al., 2012).

Evidence suggests both mood change and weight changing behaviour may spread from person to person via a process known as *social contagion*. Past research has suggested various plausible mechanisms over which social contagion could occur. Previous studies have found social support and befriending to be beneficial to mood disorders in adolescents (Dean and Ensel, 1982; Ueno, 2005; Rueger et al., 2010; Mead et al., 2010), whilst recent experiments suggest that an individuals emotional state can be affected by exposure to the emotional expressions of social contacts (Kramer et al., 2014). Associations have also been found between social norms and weight perception, suggesting that the weight of our peers can skew our idea of a healthy weight (Etilé, 2007). Studies have shown that average population weight has increased inline with changing perception of what weight is considered overweight in both the UK and US (Johnson et al., 2008; Burke and Heiland, 2007).

In recent years, evidence has been found to suggest that some behaviour-based health aspects, including happiness and obesity as well as others such as smoking cessation, can spread from person to person via social contagion (Hill et al., 2010a; Christakis and Fowler, 2013; Ali et al., 2012; Christakis and Fowler, 2007, 2008; Hill et al., 2010b; Fowler and Christakis, 2008; Balbo and Barban, 2014; Centola, 2010). However, such work has come under criticism for being unable to distinguish contagion from other possible phenomena that could confound any positive findings of contagion (Lyons, 2011; Cohen-Cole and Fletcher, 2008a; Thomas, 2013; Aral et al., 2009; Noel and Nyhan, 2011). The two simplest confounding phenomena are homophily, where individuals become friends due to sharing the same behaviour, and shared context, where individuals tend towards the same behaviour whether they are friends or not due to some outside influence (Lyons, 2011).

Hill et al. (2015) recently developed a model that distinguishes contagion from homophily and shared context. In this approach, they assess statistically whether the probability of an individual changing between binary states over time forms a better fit to the data when risk is stratified by the number of same or opposing state friends the individual has, or when risk is independent of the state of the individuals friends. This showed that while healthy mood spreads, depression does not, although treating a complex set of mood states as either 'ill' or 'not ill' can be an oversimplification. Doing this in the case of depression ignores all individuals with sub-threshold levels of depressive symptoms, despite their public health importance (Das-Munshi et al., 2008).

In previous work we generalised the model of Hill et al. (2015) to multiple mood states (Eyre, 2014). In doing so it was shown that mood is generally socially contagious, i.e. having more friends with better mood increases the likelihood of an individual improving in mood and decreases their likelihood of worsening, and vice versa for worse mood friends. This stands in seeming contradiction to the results of Hill et al. (2015) and shows the importance of considering non-binary states. Doing so reveals greater detail of the phenomena occurring and therefore provides greater information to effect policy and the design of health interventions. However, this is not to say that these results render those of Hill et al. (2015) invalid, and it is important to explore the reason behind this contradiction further.

In this chapter we first introduce our generalised multi-state social contagion model, describing how it overcomes the criticisms of previous work and shows the importance of using non-linear multi-state models to analyse complex health systems and overcome the criticisms of using generalised linear models (GLMs). We apply this model to social contagion of weight change as a further validation and to provide insight into the spread of obesity amongst adolescents, an issue we have seen to be of great concern due to the health implications to said adolescents as described above. Following this we employ a Gaussian process regression model to reconcile our previous results for mood with those of Hill et al. (2015). Finally, we explore some possible ways to further improve this model in order to more accurately reflect what is occurring within the system.

3.2 Multi-state model for social contagion

Typically social contagion models in the previous studies have been performed using GLMs (Christakis and Fowler, 2013), with a form like

$$g\left(E\left[Y_{t+1}^{i}\right]\right) = \beta_0 + \beta_1 y_t^i + \beta_2 y_{t+1}^j + \beta_3 y_t^j + \sum_{k=4}^{M} \beta_k x_k \tag{3.1}$$

where $g\left(\cdot\right)$ is the link function of the particular generalised regression, y_t^i is the state of individual i at time t, x_k are any other covariates considered in the regression, β_k are the regression parameters, and the model is considering effects between pairs of friends. The claim is then that if the coefficient β_2 is significantly positive this is evidence of a causal influence on the state of individual i by the state of individual j.

The model protects against homophily due to the inclusion of the y_t^j term. Modelling y_{t+1}^i as dependent on y_t^i leads to the model acting as a Markov chain, resulting in y_{t+1}^i being independent of both the state of individual i and that of individual j at the time they became friends. It protects against shared context by a found asymmetry in the impact of individual j on i from that of individual i on j.

As mentioned, this method has come under a lot of extreme (in some cases perhaps too extreme) criticism. One such criticism is that considering β_2 to be the effect of interest is a strange choice as causality should occur over time, so rather β_3 should contain the contagion effect (Lyons, 2011). Another argued that the model can be used to provide evidence for contagion of nonsensical things such as height (Cohen-

Cole and Fletcher, 2008a). Yet another argued that whilst the lagged friend term may control for homophily causing the initial creation of the friendship, it does not control for homophily causing the retention of the friendship (Noel and Nyhan, 2011). There were also criticisms as to the size and directions of the effects found, such as that commonly β_2 and β_3 were found to be of opposing sign (which would produce a negative homophily effect insinuating that any possible homophily leads to dissimilarity rather than similarity) and that the differences in the impact of individual j on i from that of individual i on j were commonly found to be insignificant (Lyons, 2011). Despite all these criticisms it has been shown that the model can successfully support the null hypothesis of no contagion (VanderWeele et al., 2012).

However, the sheer amount of criticism, and the lack of support for the models ability to show a significant result of actual social contagion, shows the difficulty in using such linear models when attempting the difficult task of inferring evidence for causal effects that are highly prone to confounding. It is therefore here that we present our first application of a more sophisticated statistical model to what is arguably a complex underlying system involving socio-economic variables by way of the social connections. This is done in the form of a non-linear parametric model, specifically that initially developed by Hill et al. (2015), where we also show how the model must be generalised to improve its applicability.

If we let a component of mood for an individual at time t with k^+ friends with better mood and k^- friends with worse mood be represented by an integer random variable $Y(t)$, we can imagine a very general probabilistic model for mood in which

$$\Pr(Y(t+1) = y' \,|\, y(t) = y) = f(y', y, k^+, k^-) \ . \tag{3.2}$$

In practice, finding an appropriate function f for such a general model becomes too difficult and so we will normally need to consider special cases of this general model. In the work of Hill et al. (2015) they considered only binary states $Y(t) = D$ for an individual with depressive symptoms at time t and $Y(t) = N$ for a non-depressed (healthy) individual, and sought to distinguish between sigmoidal dependence on the number of friends in a given state and no such dependence.

Such an approach is robust to confounding from homophily and shared context, as shown in Hill et al. (2015). In simplified terms, this comes from the fact that this model considers transition probabilities between states, which are distinguishable for contagion and the other basic phenomena that could confound it, rather than stationary distributions, which are not. If it can be shown that the transition probability for an individual of going from being depressed to not depressed is stratified by the number of friends they have who are depressed at the initial time point, i.e. that the probability $\Pr(Y(t+1) = D \,|\, Y(t) = N)$ is higher, and the probability $\Pr(Y(t+1) = N \,|\, Y(t) = D)$ is lower, for individuals with more friends in state D than the baseline values (with zero friends in state D), then we can infer evidence of social contagion that is not confounded by the typical confounding phenomena. If homophily or shared context were occurring within the data, rather than contagion, then we would simply expect to see more clusters of same state individuals in the dataset and we would expect individuals within these clusters to transition together. We would not expect either of these phenomena to result in the stratification of the transition probability by the number of contagious state friends.

It is possible that some more complicated phenomena could be confounding the method, but that does not negate the improvement this method presents over previous methods. There is also the argument that the simplest possible explanation is the most likely. If this is true here, then ruling out shared context and homophily leaves contagion as the simplest explanation.

Despite the robustness this method does not account for the possibility of non-binary states, such as the different continuous numerical scores for the weights individuals have. To relax this assumption we now let $Y(t)$ be an integer (for discrete states) or a continuous number (for continuous states), and consider a trinomial model specified by three probabilities: the probability of increasing state, the probability of decreasing state, and the probability of remaining in the same state

$$\begin{aligned} &\Pr\left(Y_i\left(t+1\right) > Y_i\left(t\right)\right) = p \ , \\ &\Pr\left(Y_i\left(t+1\right) < Y_i\left(t\right)\right) = q \ , \\ &\Pr\left(Y_i\left(t+1\right) = Y_i\left(t\right)\right) = 1 - p - q \ . \end{aligned} \tag{3.3}$$

We can then examine whether these probabilities were dependent on the states of an individuals friends relative to their own at the first time point by comparing two different functional forms for p and q. The first is conditioned on the number of friends an individual had who had a higher/lower score at the first time point, k. This takes the form of a discrete S-shaped (sigmoidal) function, appropriate for behavioural contagion being a type of complex contagion (Centola and Macy, 2007;

Centola, 2010; Valente, 1996), with the following mathematical formulation:

$$\begin{aligned} p_k &= \alpha + \beta \sum_{l=0}^{k} \binom{10}{l} \gamma^l \left(1-\gamma\right)^{1-l} \ , \\ q_k &= \delta + \epsilon \sum_{l=0}^{k} \binom{10}{l} \zeta^l \left(1-\zeta\right)^{1-l} \ . \end{aligned} \tag{3.4}$$

Here the parameters α and δ correspond to the baseline transition probabilities. The contribution to the transition probabilities by each additional friend is given by the binomial terms dependent on the parameters γ and ζ with an overall amplitude given by β and ϵ, which leads to the sigmoidal complex contagion form required. For each individual with k higher or lower scoring friends, k binomial terms (from a binomial distribution with up to 10 possible successes, reflecting the maximum number of friends an individual is allowed to list in the data set) are added to the baseline transition probability to give the individual's transition probability given their friends' states. The second functional form for p and q is independent of the states of the friends:

$$p_k = \alpha \, , \qquad q_k = \delta \, . \tag{3.5}$$

Each one is dependent on only the baseline transition probability. Using each possible combination of these two functional forms gives us four models to compare. Model 1, where p_k and q_k are given by (3.4), has both increasing and decreasing state being dependent on friend states. Model 2, where p_k and q_k are given by (3.5), has neither increasing nor decreasing state being dependent on friend states. Model 3, where p_k is given by (3.4) and q_k by (3.5), has increasing state alone being dependent on friend states. Model 4, where p_k is given by (3.5) and q_k by (3.4), has decreasing state alone being dependent on friend states.

These models can each be fitted to data using maximum-likelihood estimation (MLE), with separate model variants conditioned either on higher scoring friends or lower scoring friends. The likelihood takes the form

$$L\left(\mathbf{n}, \mathbf{m} | \mathbf{p}, \mathbf{q}, \mathbf{N}\right) = \prod_k \binom{N_k}{n_k, m_k, N_k - n_k - m_k} p_k^{n_k} q_k^{m_k} \left(1 - p_k - q_k\right)^{N_k - n_k - m_k} \tag{3.6}$$

where n_k was the number of individuals with k higher / lower scoring friends who worsened, m_k was the number of individuals with k higher / lower scoring friends who improved, N_k was the total number of individuals with k higher / lower scoring friends, and the multinomial coefficient is

$$\binom{N_k}{n_k, m_k, N_k - n_k - m_k} = \frac{N_k!}{n_k! m_k! \left(N_k - n_k - m_k\right)!} . \tag{3.7}$$

Competing models are then compared using their Akaike Information Criterion (AIC) values in order to find the preferred model in each case

$$\mathrm{AIC} = 2\nu - 2 \log L \left(\mathbf{n}, \mathbf{m} | \hat{\mathbf{p}}, \hat{\mathbf{q}}, \mathbf{N}\right) \tag{3.8}$$

where ν is the number of parameters in the model, and $\hat{\mathbf{p}}$ and $\hat{\mathbf{q}}$ are the values of $\mathbf{p}$ and $\mathbf{q}$ dependent on the fitted parameters (Akaike, 1974).

We initially used this method to model social contagion for discrete mood states (Eyre, 2014). We now apply it to continuous weight states in order to examine the possibility of contagion of mood changes in an adolescent population.

3.3 Weight change contagion

3.3.1 Data

We applied the model to data from the National Longitudinal Study of Adolescent to Adult Health (Add Health), which includes health and socio-demographic data of a nationally representative group of US adolescents taken at different time points over several waves starting in 1994 and repeated in 1995-96, 2001-02, and 2008 (Harris et al., 2008). We used data from the first two waves of the in-home interview survey of Add Health, which were performed 6-12 months apart. These included records of adolescents' in-school friends.

To be included in our study sample analysing weight state, at both time points the adolescent student had to be from a saturated school (in which all students were given the in-home interview, eliminating selection bias and ensuring as complete a social network as possible), have provided complete height, weight, age, and gender information at both time points, and have been the least restricted in the number of school friends they were allowed to list (each student was either asked to list up to five male and five female friends, or were limited to only listing one male and one female friend). This gave us a sample size of 2161 individuals.

We defined the adolescents' weight states by their Body Mass Index (BMI) z-score, which is calculated by a Box-Cox transformation (which transforms non-normal variables to approximate normality) of their BMI score (Must and Anderson, 2006; Cole and Green, 1992)

$$\text{BMI z-score} = \begin{cases} \frac{(\text{BMI}/M)^{L}-1}{LS} & \text{if } L \neq 0 \\ \frac{\log(\text{BMI}/M)}{S} & \text{if } L = 0 \end{cases} \tag{3.9}$$

with parameters M, L, and S dependent on the individual's age and gender, the values of which were taken from the Centre for Disease Control and Prevention (CDC) 2000 growth charts (CDC, 2009). The BMI score is a measure of the ratio of weight to body size, often taken as a clinical measure of body fat, given by

$$\text{BMI} = \frac{\text{weight [lbs]}}{(\text{height [in]})^2} \times 703 \tag{3.10}$$

and the BMI z-score is then simply a measure of the position of the individual within the overall standardised distribution of BMI scores for some greater population. By using BMI z-scores instead of BMI scores this allows us to compare the states of adolescents with different age and gender, as BMI itself is confounded by these factors (CDC, 2014). It is worth noting that there is significant evidence that BMI can often be a poor measure of obesity, even misclassifying individuals as unhealthy when their cardiometabolic health is good (Tomiyama et al., 2016). It is arguable that BMI still works as a population level measure in studies such as this, but it is worthwhile in future studies considering other possible measures related to obesity and being overweight.

The weight state then is continuous with infinite range, and centered at zero. Due to this continuous nature, some changes in weight state were found to be incredibly small, which could obviously be due to measurement noise or expected variations in weight. The analysis described above was therefore performed twice in this case. Once where all changes in weight state were considered, no matter what their size. The second time, where only clinically significant changes in weight state were considered, taken as being any weight state change of magnitude greater than or equal to 0.2, a value taken from clinical trials (Sacher et al., 2010).

3.3.2 Results

When considering all weight changes (even non-clinically significant ones), the preferred model according to AIC value is Model 1 (Figure 3.1 and table 3.1). This would imply that both having more friends of a higher weight and more friends of

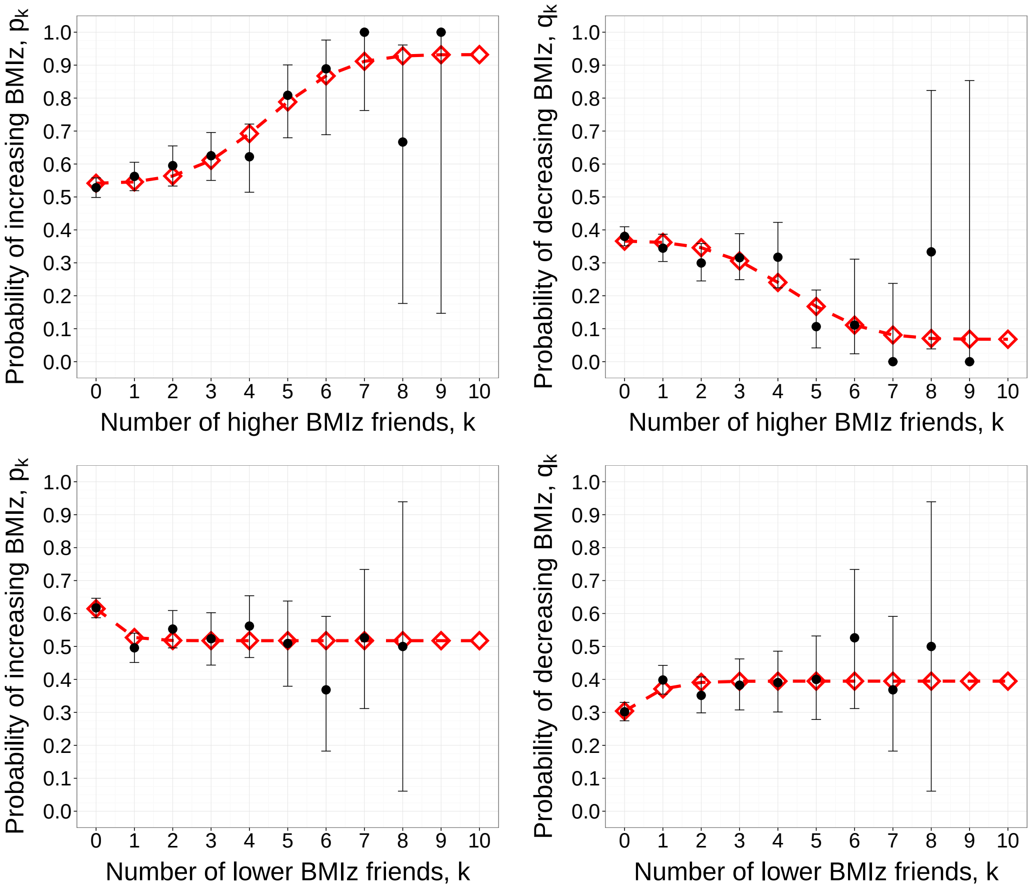


Figure 3.1: Probability of changing weight state as a function of either the number of lower weight state friends or the number of higher weight state friends. Observed data (black circles) is shown with 95% confidence intervals alongside the results of fitting (red diamonds) of the state change model to the Add Health data. The preferred model in this case for both lower weight and higher weight friends had both increasing and decreasing state being dependent (Model 1). Parameter values are shown in Table 3.1. For higher weight friends the AIC values were 91.89 for Model 1, 119.68 for Model 2, 113.54 for Model 3, and 123.02 for Model 4. For lower weight friends they were 96.45 for Model 1, 109.31 for Model 2, 108.84 for Model 3, and 111.77 for Model 4.

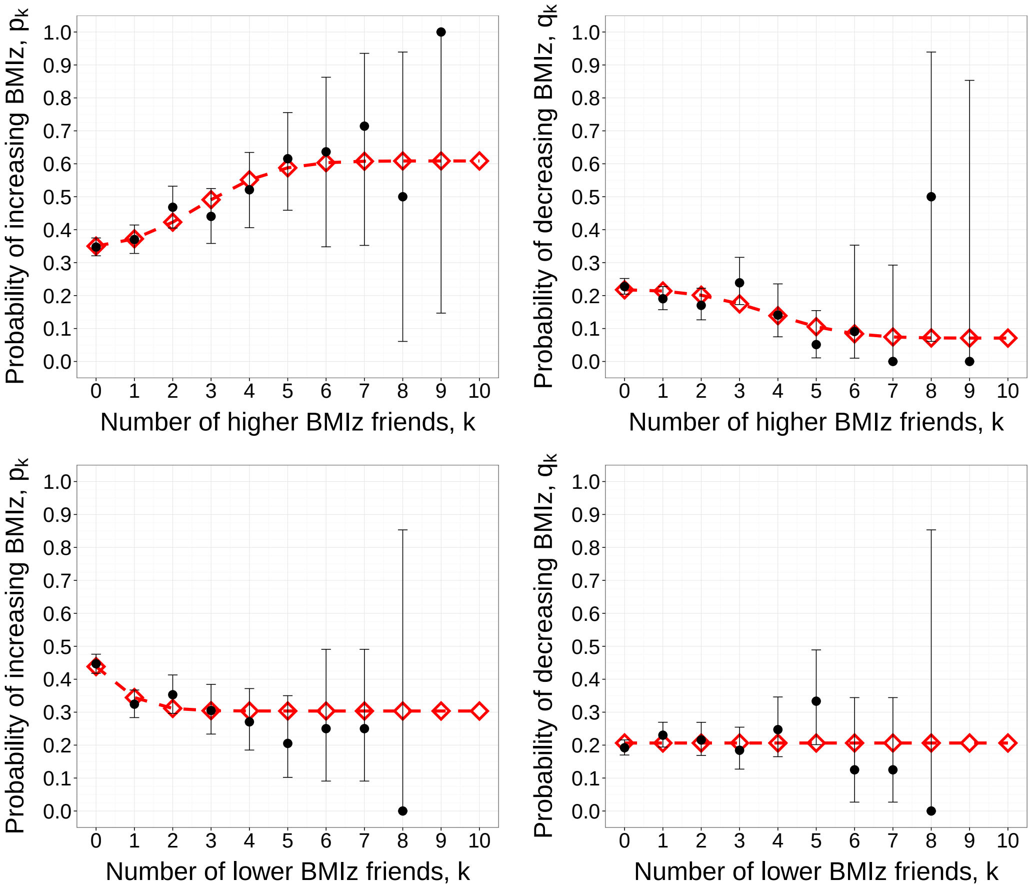


Figure 3.2: Probability of changing weight state by a clinically significant amount as a function of either the number of clinically significantly lower weight state friends or the number of clinically significantly higher weight state friends. Clinically significant refers to a difference in weight state, defined as BMI z-score, of 0.2 or more in size. Observed data (black circles) is shown with 95% confidence intervals alongside the results of fitting (red diamonds) of the state change model to the Add Health data. The preferred model in this case for higher weight friends had both increasing and decreasing state being dependent (Model 1). For lower weight friends, it had increasing state alone being dependent (Model 3). Parameter values are shown in Table 3.2. For higher weight friends the AIC values were 99.71 for Model 1, 125.56 for Model 2, 104.89 for Model 3, and 128.21 for Model 4. For lower weight friends they were 100.69 for Model 1, 132.73 for Model 2, 99.50 for Model 3, and 133.83 for Model 4.

Parameter	Higher weight friends model			Lower weight friends model		
	Value	Lower CI	Upper CI	Value	Lower CI	Upper CI
α	0.5411	0.5168	0.5655	1.000	-1.6194	3.6194
β	0.3908	0.2641	0.5175	-0.4824	-3.0771	2.1123
γ	0.4960	0.4055	0.5866	0.0222	-0.1160	0.1604
δ	0.3664	0.3425	0.3903	0.2007	-0.1492	0.5506
ϵ	-0.2983	-0.4221	-0.1745	0.1940	-0.1333	0.5214
ζ	0.4820	0.3653	0.5987	0.0612	-0.0951	0.2173

Table 3.1: Fitted parameter values for the preferred model (Model 1) of weight change dependent on higher or lower weight friends, with upper and lower values for their 95% confidence intervals calculated using the asymptotic normality of maximum-likelihood estimates.

Parameter	Higher weight friends model			Lower weight friends model		
	Value	Lower CI	Upper CI	Value	Lower CI	Upper CI
α	0.3456	0.3164	0.3745	0.5590	0.3154	0.8026
β	0.2630	0.1418	0.3841	-0.2555	-0.4766	-0.0343
γ	0.3357	0.2133	0.4582	0.0723	-0.0316	0.1761
δ	0.2180	0.1970	0.2390	0.2064	0.1893	0.2234
ϵ	-0.1470	-0.2572	-0.0369	-	-	-
ζ	0.4368	0.2360	0.6377	-	-	-

Table 3.2: Fitted parameter values for the preferred model (Model 1 for higher weight friends, Model 3 for lower weight friends) of clinically significant weight change dependent on clinically significant higher or lower weight friends, with upper and lower values for their 95% confidence intervals calculated using the asymptotic normality of maximum-likelihood estimates. Clinically significant refers to a difference in weight state, defined as BMI z-score, of 0.2 or more in size.

a lower weight impacts on an individuals likelihood of changing weight.

Indeed it can be seen that having an increased number of higher weight friends has the expected positive impact on the probability of weight increase and negative impact on that of weight decrease we would expect to see from contagion. The impact also appears quite large in magnitude, with the probability of increasing weight being driven from a baseline value of between 0.5 and 0.6 to up to above 0.9 for the maximum possible number of higher weight friends, and that of decreasing weight going from a baseline of just below 0.4 to below 0.1.

The same cannot be said for the impact of lower weight friends. There is a slight change between not having higher or lower weight friends and having at least one, but the confidence intervals in our data, as well as those of the model parameters themselves (Table 3.1), are large enough to undermine this possibility.

If we consider only clinically significant weight changes (Figure 3.2 and table 3.2) these results become more apparent. Model 1 is still preferred for the impact of higher weight friends, still supporting the conclusion of contagion from these (though with a much smaller impact on the probabilities of changing weight). However, for the impact of lower scoring friends Model 3 is preferred, so we see no evidence of an impact on the probability of an individual decreasing in weight. Therefore we find no evidence of a contagion effect from lower weight friends.

Although the results show a particular shape to the weight change probabilities over the number of higher and lower weight friends, due to the large confidence intervals about the data for high numbers of friends (caused by the lack of data in these regions) most conclusions that could be inferred from these shapes would not be particularly robust. Yet they do appear to highlight only a very small threshold on the number of friends with higher or lower weight needed to result in a contagion effect.

Goodness of fit tests were performed by comparing observed residuals of state changes to the empirical distributions of residuals found using parametric bootstrapping on the fitted model, the results of which support these conclusions (see Appendix A).

The fact that only higher weight friends are found to have an impact supports the possibility of contagion of obesity over that of weight change in general. However, it does not undermine the point of considering non-binary states, as it shows it is still important to consider higher weight friends in general, not just obese friends, when considering the risk of an individual becoming obese. Certainly such a contagion effect could imply a trend towards weight increases in US adolescents over time, consistent with results shown in the literature (Fryar et al., 2012). It is also important to consider this more general version of the method for a similar reason argued in our previous work on mood (Eyre, 2014), in that being sub-threshold to a clinical diagnosis of obesity still presents significant health risks and therefore it is important to not ignore such individuals and the risk to any individuals of getting closer to the obesity threshold in weight (Must et al., 1999).

Previous studies using methods such as the GLM discussed above have lead to conclusions similar to ours (Hill et al., 2010b; Ali et al., 2012; Christakis and Fowler, 2007). However, as mentioned before, these studies have come with some controversy, with other studies arguing that significant findings of obesity contagion are rendered insignificant when controlling for appropriate environmental factors (Cohen-Cole and Fletcher, 2008b). This is one of the core issues of trying to infer causal effects from observational data using GLMs. The legitimacy of results relies on successfully controlling for any other possible basic explanation for the outcome, which is notoriously difficult to do. Non-linear parametric models, such as that used in our method here, are in a way almost 'physically' modelling the phenomena we are looking for. By tailoring the assumptions of the model to the phenomena we make it more robust against confounding, such that if the model is found to be preferred in the dataset from an alternative devoid of such assumptions then we gain much more robust inferences of a causal effect.

As such, the results we have found here using this more robust contagion model provides much stronger evidence in support of obesity contagion that can give much more informed aid to the design of public policy and interventions against adolescent obesity.

3.4 Reconciling the multi- and bi-state results for mood contagion

Though the weight contagion results we have presented here largely agree with previous work, there was a noted contradiction between our previous work on mood contagion and other studies performed in that area (Eyre, 2014). In particular, the results of our mood contagion work seemed to contradict those of Hill et al. (2015). Here we introduce a complementary Gaussian process model that helps to demonstrate why our previous results are consistent with those of Hill et al. (2015).

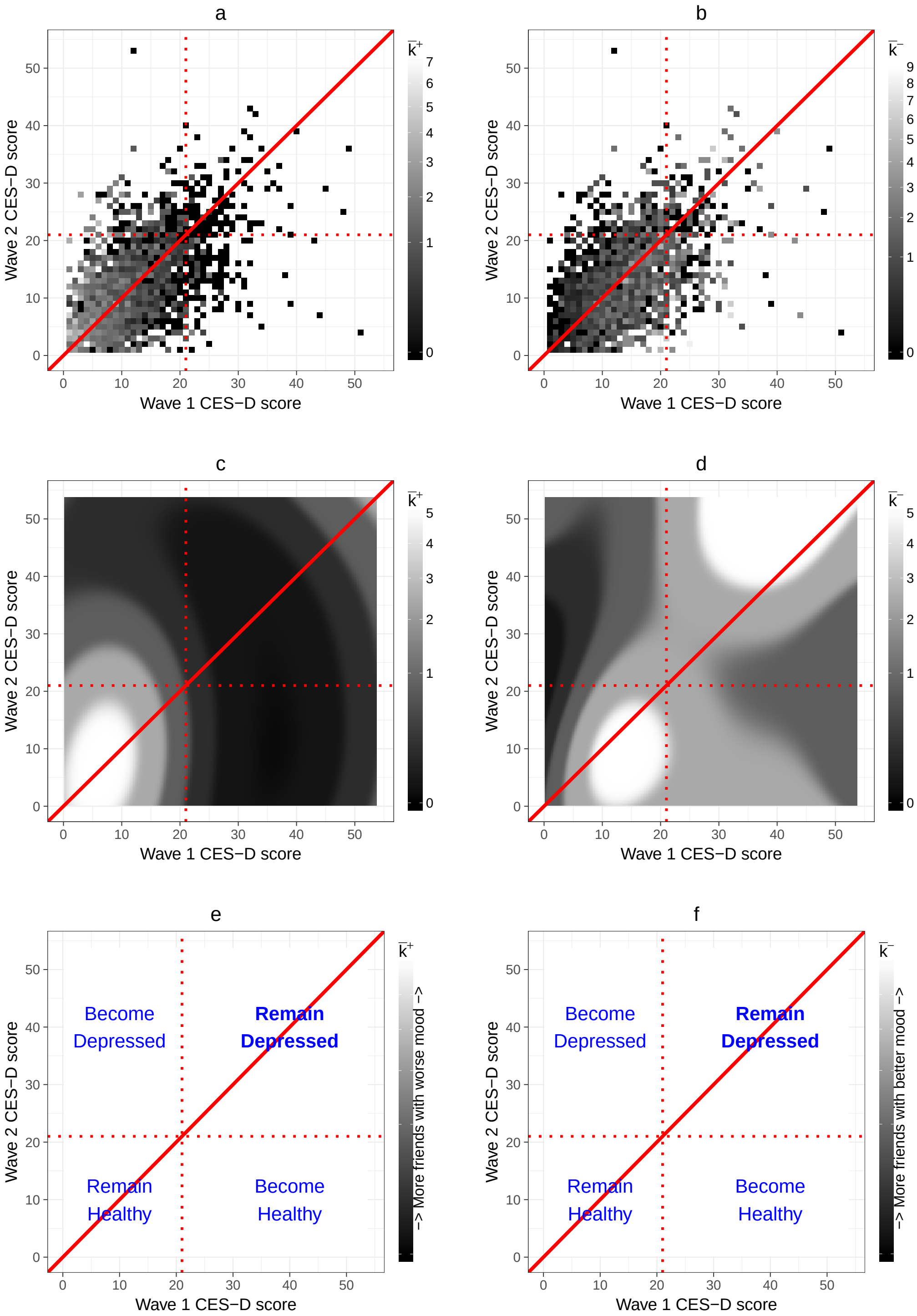


Figure 3.3: Gaussian process model results for the average numbers of better and worse mood friends dependent on the mood states at each time point.

Our mood contagion work was performed using the same method and same dataset as that used here for weight contagion. Instead of considering changes in weight states, defined as continuous BMIz values, we considered discrete mood states defined for each individual as their Centre for Epidemiological Studies Depression scale (CES-D) score calculated from the set of 18 CES-D questions asked within the Add Health survey (Radloff, 1977). This gave a discrete integer mood state for each individual ranging from 0 to 54, where a higher state indicated a worse mood. In the work of Hill et al. (2015), binary states of depressed and not depressed were defined relative to a threshold value in CES-D score. Individuals above the threshold were marked as depressed, and below were marked as not depressed.

The model described in Section 3.2 deals with the probability of a change of state $Y(t) \rightarrow Y(t+1)$ given a number k of better or worse scoring friends. We might instead assume that the initial state $Y(t)$ and the state at the next time point $Y(t+1)$ are known and treat the number of friends k (either better or worse) as the random variable to be modelled. As can be seen in the top row of Figure 3.3, the data in this form is very noisy and so we smooth the function $k(Y(t), Y(t+1))$ using Gaussian process regression (see Section 2.2 and Rasmussen and Williams (2006)).

In order to take into account the fact that k has both an upper and lower bound, we used warped Gaussian process regression (Snelson et al., 2004). In this variation of the method we pass the output k through a warping function in order to transform it to an infinite and continuous variable that the regression performs better in learning. The warping function we used was the inverse probit function.

The smoothing was performed using a squared exponential covariance function with lengthscales constrained to have a minimum value of 20 each in order to ensure high levels of smoothing. The hyperparameters of the covariance function were fitted to the data using maximisation of the marginal likelihood.

This Gaussian process model shows that most of the individuals with a greater number of higher scoring friends who were initially below the threshold for depression remained that way at the second time point, whilst the individuals with a greater number of lower scoring friends are more spread out in their score combinations such that many that started off above the threshold for depression passed below the threshold at the second time point (Figure 3.3).

To explain Figure 3.3, it shows the Wave 1 and Wave 2 CES-D scores for (a,b) our empirical sample and (c,d) the Gaussian process model. The left column (a,c,e) is coloured by the mean number of friends with worse mood $\bar{k}^+$ and the right column (b,d,f) is coloured by the mean number of friends with better mood $\bar{k}^-$. The set of states for those who have not changed in state is shown by the diagonal solid red line. The gender-averaged threshold boundary between the states of depressed and not depressed for each wave are shown by the dashed red lines, and the plots (e,f) show how to interpret the delineated quadrants. Individuals with greater numbers of worse or better mood friends (who we expect to experience a stronger contagion effect) are concentrated in the lighter regions of the plot. We see that individuals with more friends with worse mood (corresponding to higher scoring friends) are contained in the bottom left quadrant, meaning they remain below the depression threshold at both time points with any negative shift in mood caused by contagion seldom enough for the individual to transition to being classified as having depressive symptoms. Individuals with more better mood friends (corresponding to lower scoring friends) are spread out over the bottom two quadrants, meaning that they relatively often improve in mood to such an extent that they cross from being classed as depressed to being healthy in wave 2.

This then suggests that both better and worse mood are contagious, but whilst better mood is contagious enough to push individuals over the boundary from depressed to not depressed worse mood is not contagious enough to push individuals into becoming depressed. Consequently, we would not expect to find contagion-like characteristics for depression using a binary model.

We therefore observe a difference between depression, which we found not to spread, and relatively low mood below the threshold for depression, which we found did spread. This supports the view that there is more to clinical depression than simply low mood (although the latter may be indicative of the former). It is also in keeping with a tendency for a reduction in the normal social interactions that lead to spreading of mood during an episode of depression (Cruwys et al., 2014).

Of existing studies by other authors, the work of Hill et al. (2010a) is closest to ours, and using a different dataset these authors concluded that 'neutral' moods did not spread but both 'content' (threshold CES-D score 12 on the positively worded questions only) and 'discontent' (threshold CES-D score 16) moods did. This work tested models of the form $p_k = \alpha + \beta k$ using an ordinary least-squares fitting approach, selecting a spreading model if the p-value for a slope-free null hypothesis is under 0.05. While we argue that our methodology using a complex contagion of the form (3.4), maximum-likelihood estimation, and information-theoretic model selection is preferable to such an approach, we believe that the most important difference with the results presented here is our use of a CES-D threshold score of 20 (or 21) for presumptive depression - and in particular that the spreading of 'discontent' at CES-D scores in the 16-20 range is consistent with our results about the spreading of sub-threshold levels of depressive symptoms.

These results help emphasise the importance of generalising the model to non-binary states to be performed alongside use of the binary state version of the model. It would be of interest in future work to perform an analysis of binary obesity states using this method to compare with the results presented in this chapter.

3.5 Towards a combined friend model for social contagion

So far our social contagion analysis has modelled the effect of higher state and lower state friends separately. One possible way to further develop the model is to try and create a version dependent on the combined effect of both higher and lower state friends. In doing so this would create a more complete and realistic model of social contagion that would allow us to gain further insight into the possible contagion phenomena of different health aspects occurring within different populations. Here we present and compare a few possible ways of achieving this.

Our approach is to consider the effect of each individual friend to be like an impulse towards the individual themselves changing state, i.e. a higher state friend would contribute a positive impulse, a lower state friend would contribute a negative impulse, and a same state friend would contribute no impulse towards state change. We consider different models for how these impulses impact on the probability of an individual changing state.

Similar to the model we presented in Section 3.2, we consider the trinomial probabilities of increasing state p, decreasing state q, and remaining in the same state $1-p-q$, this time dependent on both the number of higher state friends k^+ and the number of lower state friends k^-. As before, for each different possible combined friend model, we compare different versions corresponding to different possibilities of p and q depending on k^+ and k^- or not using their AIC values, therefore attempting to capture whether a contagion or a no contagion model is preferred by the data. The fits are achieved through maximum likelihood estimation, where the

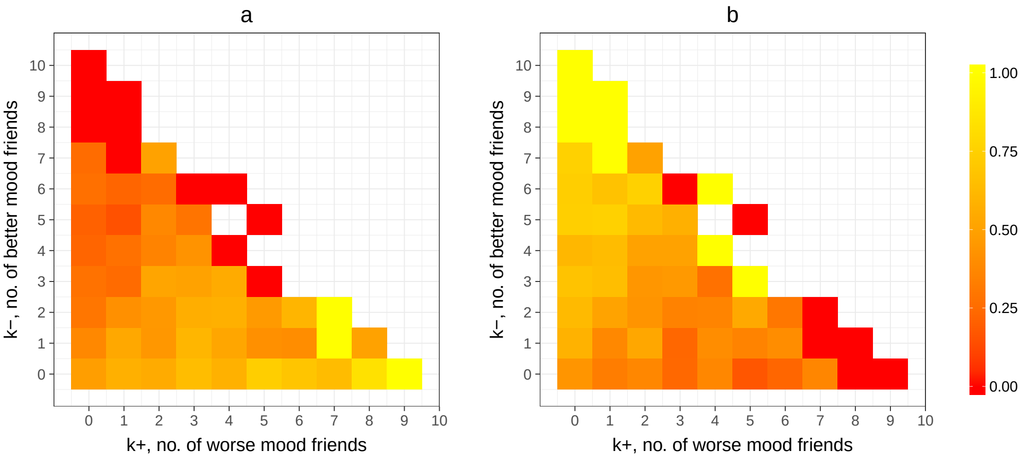


Figure 3.4: Empirical frequencies of individuals (from our sample from the Add Health dataset) worsening mood (a) and improving mood (b) for all combinations of number of worse mood (k^+) and better mood (k^-) friends the individuals have.

likelihood is this time given by

$$L(\mathbf{n}, \mathbf{m}|\mathbf{p}, \mathbf{q}, \mathbf{N}) = \prod_{k^+,k^-} \binom{N_{k^+,k^-}}{n_{k^+,k^-}, m_{k^+,k^-}, N_{k^+,k^-} - n_{k^+,k^-} - m_{k^+,k^-}} \times p_{k^+,k^-}^{n_{k^+,k^-}} \\ \times q_{k^+,k^-}^{m_{k^+,k^-}} \times \left(1 - p_{k^+,k^-} - q_{k^+,k^-}\right)^{N_{k^+,k^-} - n_{k^+,k^-} - m_{k^+,k^-}} \quad (3.11)$$

where N_{k^+,k^-} is the number of individuals with k^+ higher state friends and k^- lower state friends, n_{k^+,k^-} of whom have increased in state between the two time points, and m_{k^+,k^-} of whom have decreased in state between the two time points.

In order to compare the different models we present here, we use them to model social contagion of mood, like in our previous work (Eyre, 2014). We fit the various models to the same dataset as discussed in that work and Section 3.4. In this case as a higher state corresponds to a higher CES-D value, p is the probability of worsening mood, q is that of improving mood, k^+ is the number of worse mood friends, and k^- is that of better mood friends. Aside from comparing the AIC values of each proposed model, we will also compare how well they appear to capture the empirical frequencies of improving and worsening mood that occur within the dataset (Figure 3.4).

3.5.1 Impulse probability product model

The first possibility we propose is that, rather than p and q being explicitly dependent on the impulses provided by friends towards changing state, we define them as functions of the probabilities of receiving these impulses, i.e.

$$\begin{aligned} p_{k^+,k^-} &= g\left(k^+\right)\left(1 - h\left(k^-\right)\right) \ , \\ q_{k^+,k^-} &= \left(1 - g\left(k^+\right)\right) h\left(k^-\right) \end{aligned} \tag{3.12}$$

where g is the probability of getting a positive state change impulse from higher state friends and h is that of getting a negative state change impulse from lower state friends. These probabilities then take on two possible forms, a sigmoidal friend-dependent form

$$\begin{aligned} g\left(k^+\right) &= \alpha + \beta \sum_{l=0}^{k^+} \binom{10}{l} \gamma^l \left(1-\gamma\right)^{1-l} \ , \\ h\left(k^-\right) &= \delta + \epsilon \sum_{l=0}^{k^-} \binom{10}{l} \zeta^l \left(1-\zeta\right)^{1-l} \end{aligned} \tag{3.13}$$

and a constant friend-independent form

$$g\left(k^+\right) = \alpha \, , \qquad h\left(k^-\right) = \delta \tag{3.14}$$

like those seen in the separate friends models given by (3.4) and (3.5). Like before we then have four possible forms of this model to compare. Form 1 where both g and h are given by (3.13). Form 2 where g is given by (3.13) and h is given by (3.14). Form 3 where they are vice versa. Form 4 where both are given by (3.14).

The resulting preferred form by AIC value was found to be Form 1 (Figure 3.5). The model does successfully capture the changes in state change probability we see in the dataset (Figure 3.4) but not the gradient of these changes. For middle values of k^+ and k^- the probabilities remain reasonably constant. The extreme values of the probabilities also fail to stretch as far in value as the empirical frequencies do, remaining quite low throughout the range of k^+ and k^-. Overall this implies that the model performs poorly in capturing what is occurring within the dataset.

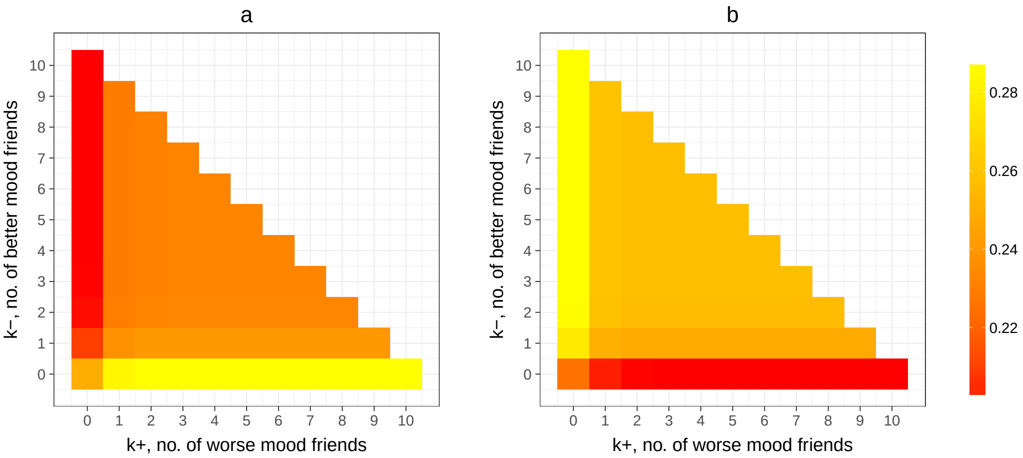


Figure 3.5: Fitted probabilities of individuals (from our sample from the Add Health dataset) worsening mood (a) and improving mood (b) for all combinations of number of worse mood (k^{+}) and better mood (k^{-}) friends the individuals have according to the impulse probability product model. General contagion was preferred with AIC value 2235.7.

3.5.2 Equal and opposite impulse models

An alternative approach is to define each friend as applying an individual impulse, where an impulse from a higher state friend is equal and opposite to that from a lower state friend. The overall impulse is then defined as

$$K = k^{+} - k^{-} \, . \tag{3.15}$$

We can then define p and q as being dependent directly on this overall impulse

$$\begin{aligned} p &= \alpha + \beta f\left(K\right) \, , \\ q &= \delta + \epsilon f\left(K\right) \end{aligned} \tag{3.16}$$

or independent in the case of no contagion

$$p = \alpha \, , \qquad q = \delta \, . \tag{3.17}$$

As before we then have four possible forms of our model to fit and compare, similar to those seen in Section 3.2 and Section 3.5.1. We do this for five possible functional forms of (3.16), acting as five separate proposed combined friend social contagion models.

The simplest of these models, Model A, has linear contagion forms

$$\begin{aligned} p &= \alpha + \beta K \,, \\ q &= \delta + \epsilon K \,. \end{aligned} \tag{3.18}$$

This ignores the status of social contagion as a form of complex contagion, where a threshold number of friends exhibiting the contagion behaviour are required to exist before the individual takes on that behaviour (Centola and Macy, 2007; Centola, 2010; Valente, 1996). We can bring in this model behaviour by incorporating an infinite discrete sigmoidal function of K. Here we choose the Poisson cumulative distribution function (CDF)

$$\text{PoissonCDF}\left(x, \lambda\right) = e^{-\lambda} \sum_{i=0}^{\lfloor x \rfloor} \frac{\lambda^i}{i!} \tag{3.19}$$

where $\lfloor \cdot \rfloor$ is the floor function.

There are four possible sigmoidal contagion forms for p and q we can consider, of varying complexity and flexibility. Model B has

$$\begin{aligned} p &= \alpha + \beta \operatorname{sign}(K) \ \text{PoissonCDF}(\left|K\right|, \lambda) \ , \\ q &= \delta + \epsilon \operatorname{sign}(K) \ \text{PoissonCDF}(\left|K\right|, \mu) \end{aligned} \tag{3.20}$$

where

$$\operatorname{sign}(K) = \begin{cases} 1 & \text{if } K > 0 \\ -1 & \text{if } K < 0 \\ 0 & \text{if } K = 0 \end{cases} . \tag{3.21}$$

In this case positive and negative overall impulses are assumed to have impacts on state change probabilities of equal magnitude where the threshold that must be passed to have an effect will also be equal. Model C has

$$\begin{aligned} p &= \alpha + (\beta \, \mathbb{1}(K > 0) + \gamma \, \mathbb{1}(K < 0)) \ \text{PoissonCDF}(\left|K\right|, \lambda) \ , \\ q &= \delta + (\epsilon \, \mathbb{1}(K > 0) + \zeta \, \mathbb{1}(K < 0)) \ \text{PoissonCDF}(\left|K\right|, \mu) \end{aligned} \tag{3.22}$$

where

$$\mathbb{1}(x) = \begin{cases} 1 & \text{if } x \text{ is true} \\ 0 & \text{if } x \text{ is false} \end{cases} . \tag{3.23}$$

This case assumes that the impact of positive and negative impulses on state change

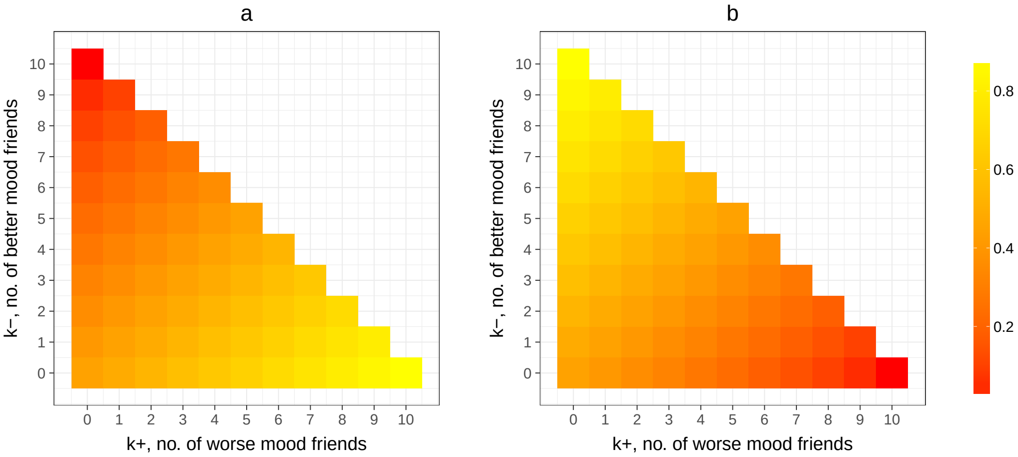


Figure 3.6: Fitted probabilities of individuals (from our sample from the Add Health dataset) worsening mood (a) and improving mood (b) for all combinations of number of worse mood (k^+) and better mood (k^-) friends the individuals have according to the equal and opposite impulse Model A. General contagion was preferred with AIC value 376.5.

probabilities have different magnitudes but equal thresholds. Model D has

$$\begin{aligned} p &= \alpha + \beta \, (\mathrm{PoissonCDF}(|K|\,, \lambda) \; \mathbb{1}(K > 0) - \mathrm{PoissonCDF}(|K|\,, \eta) \; \mathbb{1}(K < 0)) \; , \\ q &= \delta - \epsilon \, (\mathrm{PoissonCDF}(|K|\,, \mu) \; \mathbb{1}(K > 0) - \mathrm{PoissonCDF}(|K|\,, \nu) \; \mathbb{1}(K < 0)) \; . \end{aligned} \tag{3.24}$$

Here the impact of positive and negative impulses on state change probabilities have equal magnitudes but different thresholds. Finally, model E has

$$\begin{aligned} p &= \alpha + \beta \, \mathrm{PoissonCDF}(|K|\,, \lambda) \; \mathbb{1}(K > 0) + \gamma \, \mathrm{PoissonCDF}(|K|\,, \eta) \; \mathbb{1}(K < 0) \; , \\ q &= \delta + \epsilon \, \mathrm{PoissonCDF}(|K|\,, \mu) \; \mathbb{1}(K > 0) + \zeta \, \mathrm{PoissonCDF}(|K|\,, \nu) \; \mathbb{1}(K < 0) \; . \end{aligned} \tag{3.25}$$

In this last case the impact of positive and negative impulses on state change probabilities have both different magnitudes and different thresholds.

For every one of these models, the form where both p and q are dependent on K are preferred (Figures 3.6 to 3.10). In terms of then comparing AIC values between these different possible models, Models C and D come out as preferred (with only a narrow difference between them). All of the models manage to capture the variations in p and q that we see in the data (Figure 3.4) much better than the impulse probability product model (Figure 3.5). There is an expected difference between the linear model and the sigmoidal models, where the gradient of variation in p and q is much

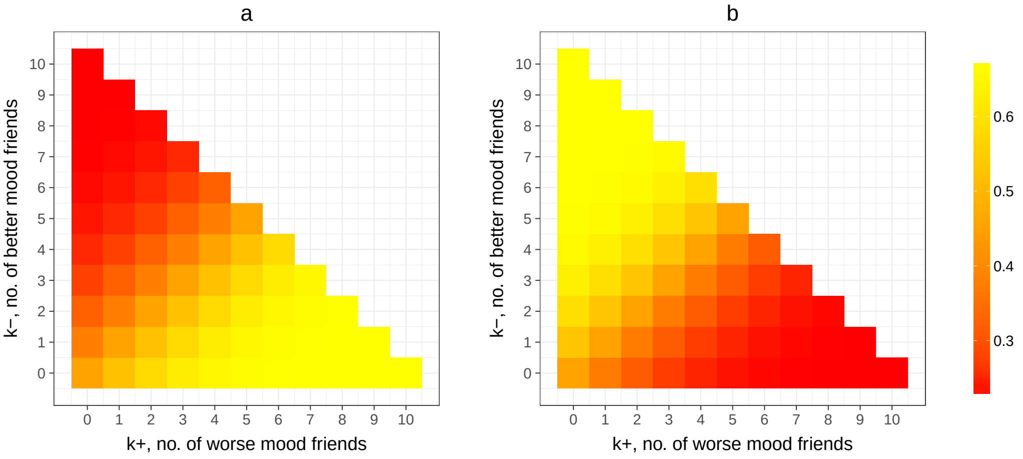


Figure 3.7: Fitted probabilities of individuals (from our sample from the Add Health dataset) worsening mood (a) and improving mood (b) for all combinations of number of worse mood (k^+) and better mood (k^-) friends the individuals have according to the equal and opposite impulse Model B. General contagion was preferred with AIC value 376.9.

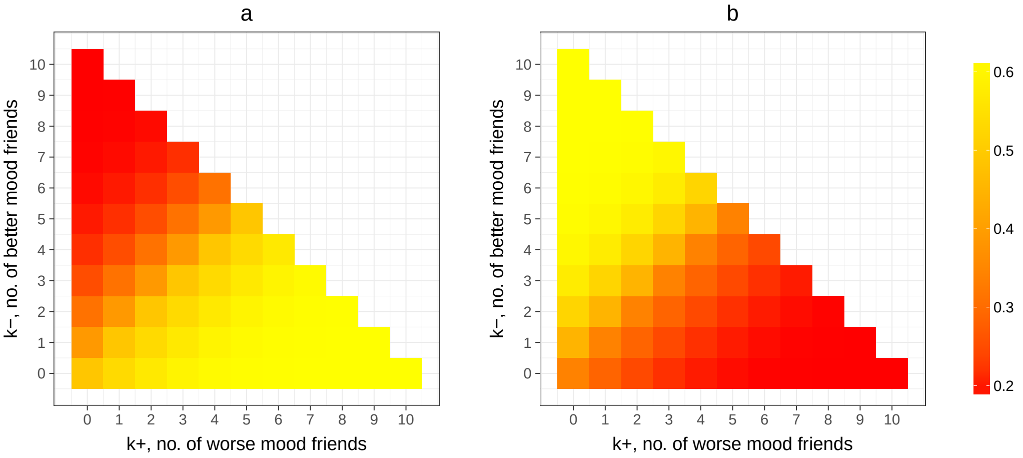


Figure 3.8: Fitted probabilities of individuals (from our sample from the Add Health dataset) worsening mood (a) and improving mood (b) for all combinations of number of worse mood (k^+) and better mood (k^-) friends the individuals have according to the equal and opposite impulse Model C. General contagion was preferred with AIC value 373.6.

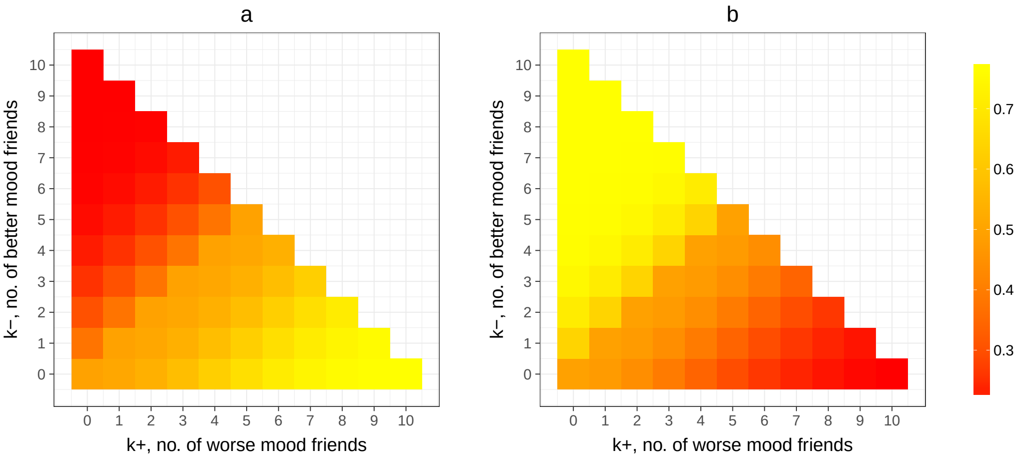


Figure 3.9: Fitted probabilities of individuals (from our sample from the Add Health dataset) worsening mood (a) and improving mood (b) for all combinations of number of worse mood (k^+) and better mood (k^-) friends the individuals have according to the equal and opposite impulse Model D. General contagion was preferred with AIC value 373.4.

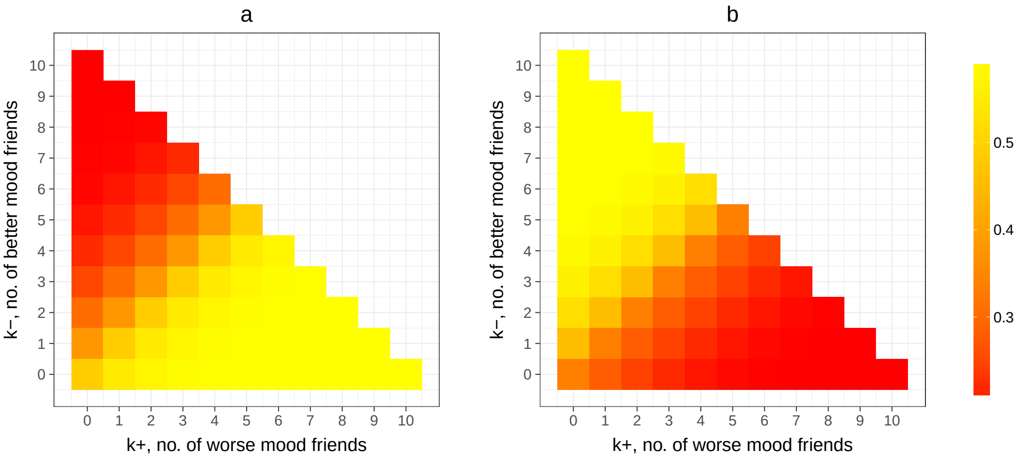


Figure 3.10: Fitted probabilities of individuals (from our sample from the Add Health dataset) worsening mood (a) and improving mood (b) for all combinations of number of worse mood (k^+) and better mood (k^-) friends the individuals have according to the equal and opposite impulse Model E. General contagion was preferred with AIC value 377.8.

smoother for the linear model. The sigmoidal models manage to capture the much coarser variation seen in the data that we would also expect of complex contagion, giving further reason to prefer Models C and D.

3.5.3 Sigmoidal impulse models

Instead of having equal and opposite impulses from higher and lower state friends, we can consider the opposing impulses to each be individual sigmoidal functions of k^+ and k^- respectively, therefore imposing the threshold effect of complex contagion on the impulses rather than the probabilities.

Using similar functional forms to the equal and opposite impulse models, we have four different possibilities which again allow for varying levels of complexity and flexibility. Model F where p and q are given by

$$\begin{aligned} p &= \alpha + \beta\left(\mathrm{PoissonCDF}\big(k^+,\lambda\big) - \mathrm{PoissonCDF}\big(k^-,\lambda\big)\right)\ , \\ q &= \delta - \epsilon\left(\mathrm{PoissonCDF}\big(k^+,\mu\big) - \mathrm{PoissonCDF}\big(k^-,\mu\big)\right)\ . \end{aligned} \tag{3.26}$$

In this case the impulses from k^+ and k^- are defined to have the same magnitude and thresholds. Model G where

$$\begin{aligned} p &= \alpha + \beta\left(\mathrm{PoissonCDF}\big(k^+,\lambda\big) - \mathrm{PoissonCDF}\big(k^-,\eta\big)\right)\ , \\ q &= \delta - \epsilon\left(\mathrm{PoissonCDF}\big(k^+,\mu\big) - \mathrm{PoissonCDF}\big(k^-,\nu\big)\right)\ . \end{aligned} \tag{3.27}$$

Here the impulses have the same magnitude but different thresholds. Model H where

$$\begin{aligned} p &= \alpha + \beta\,\mathrm{PoissonCDF}\big(k^+,\lambda\big) + \gamma\,\mathrm{PoissonCDF}\big(k^-,\lambda\big)\ , \\ q &= \delta + \epsilon\,\mathrm{PoissonCDF}\big(k^+,\mu\big) + \zeta\,\mathrm{PoissonCDF}\big(k^-,\mu\big)\ . \end{aligned} \tag{3.28}$$

For this case the impulses have different magnitudes but same thresholds. Finally, Model I where

$$\begin{aligned} p &= \alpha + \beta\,\mathrm{PoissonCDF}\big(k^+,\lambda\big) + \gamma\,\mathrm{PoissonCDF}\big(k^-,\eta\big)\ , \\ q &= \delta + \epsilon\,\mathrm{PoissonCDF}\big(k^+,\mu\big) + \zeta\,\mathrm{PoissonCDF}\big(k^-,\nu\big)\ . \end{aligned} \tag{3.29}$$

This final model has both different magnitudes and thresholds on the impulses.

As with the previous cases, for Models F, G, and H the forms for p and q dependent on k^+ and k^- (the so called contagion forms) are found to be preferred by the data over constant forms (i.e. no contagion forms) (Figures 3.11 to 3.13). The differences

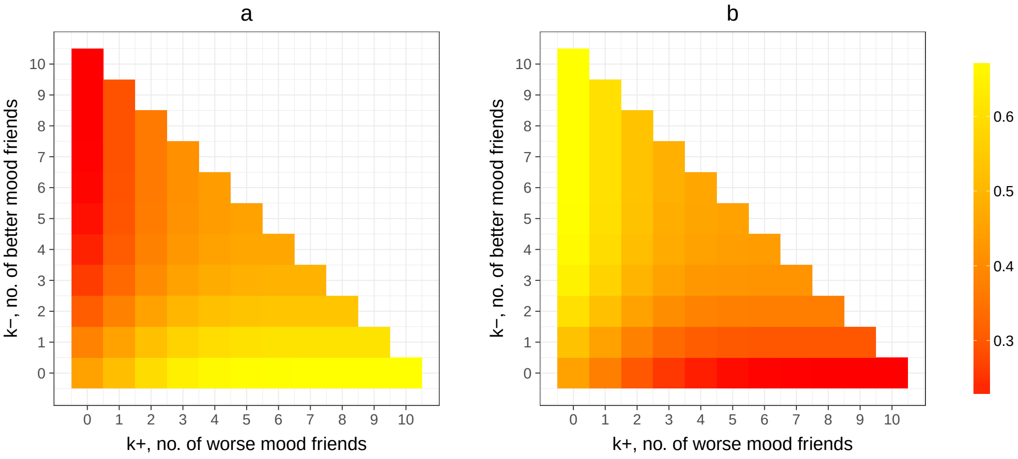


Figure 3.11: Fitted probabilities of individuals (from our sample from the Add Health dataset) worsening mood (a) and improving mood (b) for all combinations of number of worse mood (k^+) and better mood (k^-) friends the individuals have according to the sigmoidal impulse Model F. General contagion was preferred with AIC value 375.2.

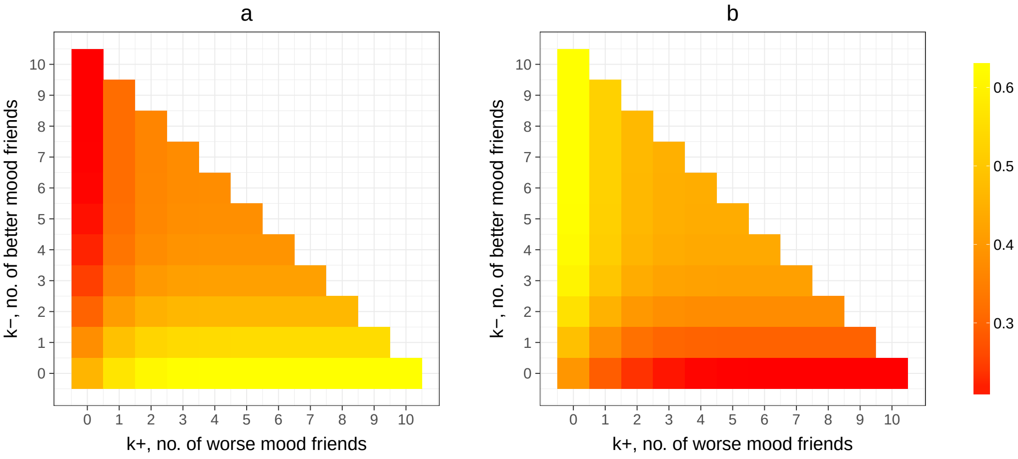


Figure 3.12: Fitted probabilities of individuals (from our sample from the Add Health dataset) worsening mood (a) and improving mood (b) for all combinations of number of worse mood (k^+) and better mood (k^-) friends the individuals have according to the sigmoidal impulse Model G. General contagion was preferred with AIC value 377.1.

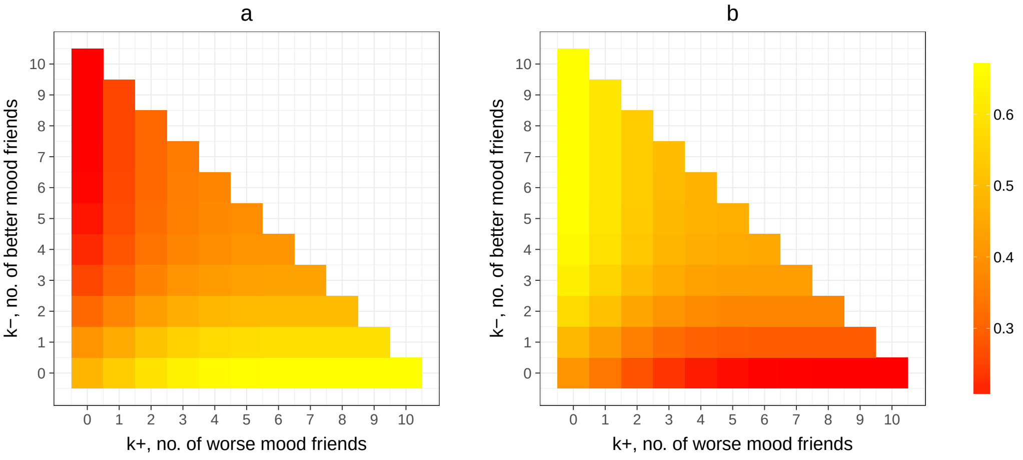


Figure 3.13: Fitted probabilities of individuals (from our sample from the Add Health dataset) worsening mood (a) and improving mood (b) for all combinations of number of worse mood (k^+) and better mood (k^-) friends the individuals have according to the sigmoidal impulse Model H. General contagion was preferred with AIC value 373.6.

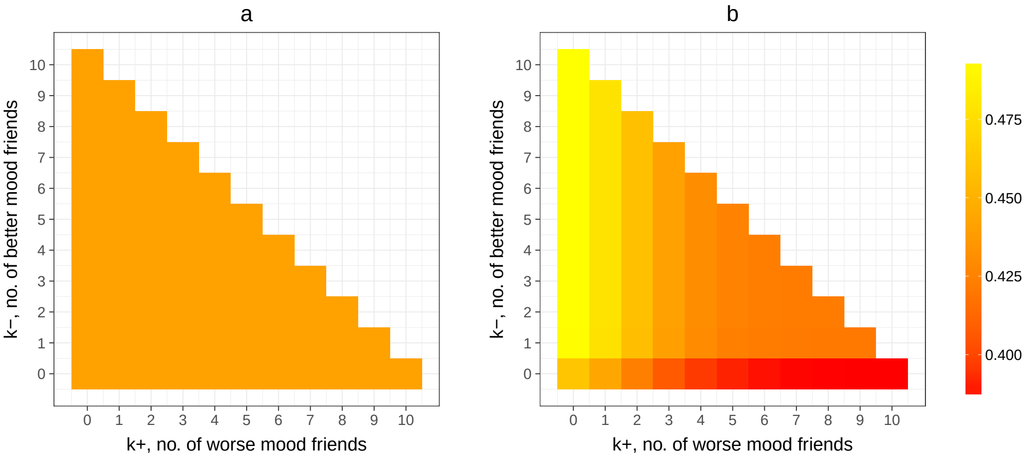


Figure 3.14: Fitted probabilities of individuals (from our sample from the Add Health dataset) worsening mood (a) and improving mood (b) for all combinations of number of worse mood (k^+) and better mood (k^-) friends the individuals have according to the sigmoidal impulse Model I. The preferred model here had only the probability of improving in mood q_k being dependent on the states of the friends, with AIC value 496.7.

between the results for most of these models appear very minor, though by AIC value Model H comes out as preferred. Model I on the other hand is the only model to have general contagion of mood not preferred, which stands in seeming contradiction to both the data and our previous results (Figure 3.14). For the remaining models, the major difference comes in comparing them to those of the equal and opposite impulse models. There we saw variations along the direction $k^+ = -k^-$, whilst here the variations curve round from the k^+ axis to the k^- one. This can be expected from the functional forms we applied, where k^+ and k^- were each allowed to make an individual impact to p and q. However, when compared to the data (Figure 3.4) these models therefore fail to capture the pattern of variation that the equal and opposite impulse models succeed in capturing.

3.5.4 Model comparison and discussion

Overall Models C and D of the equal and opposite impulse models and Model H of the sigmoidal models perform better than each of the others, though show little difference in performance from each other. However, as noted above, the sigmoidal models such as Model H fail to capture the pattern of variation in state change probability that we see in the data. Therefore, out of all these possibilities, Models C and D show the most promise as combined friend effect models for future work. This is interesting due to its various implications, such as the fact that either the magnitudes of positive and negative impulse effects, or the thresholds on K after which these effects start to take place, must be equal. Too much flexibility worsens the performance of the models in the case of mood in the Add Health dataset. It also implies that the effects from higher and lower state friends are indeed equal and opposite, which appears consistent with our previous results using separate models for higher and lower state friends effects (Eyre, 2014).

This now gives us an avenue for further developing and applying the model, allowing us to refine it to better capture the complexities of the behaviour of social contagion and ensure more robust results and research into for what health aspects social contagion does occur.

One final possible extension of the model would be in considering state changes of individual sizes rather than simply the directions of those changes. We could then model the probability of changing state as

$$\begin{aligned} p &= \Pr\left(Y\left(t+1\right) - Y\left(t\right) = \epsilon\right) \\ &= f\left(k^{+}, k^{-}, Y\left(t\right), \epsilon\right) \; . \end{aligned} \tag{3.30}$$

This would then give us a general model that would allow us to look at both binary and non-binary states, as knowing the size, direction, and starting points of each state change allows us to know when individuals cross the boundaries separating binary states. Though this is an interesting direction for future work, it would also come with great difficulties both from having to design a valid functional form and for having to obtain enough data to perform inferences on each individual possible state change.

CHAPTER 4

Gaussian process modelling of pregnancy

4.1 Adolescent pregnancy in sub-Saharan Africa

Adolescent pregnancy (i.e. pregnancy occurring at or before the age of 19) is a major social concern throughout much of the world, and in particular in large parts of the developing world. In sub-Saharan Africa adolescent pregnancy is still considered to be at unacceptably high levels (Panday et al., 2009). This is considered also true for South Africa, though it has the lowest adolescent pregnancy rates in all of sub-Saharan Africa (Kaufman et al., 2001). An overall decline in fertility has been observed in South Africa for the last several years but this decline is much slower for adolescents (Garenne et al., 2007).

There are many reasons for the high prevalence of adolescent pregnancies. There is an apparent lack of use and knowledge about contraceptives (Kaufman et al., 2001; Ehlers, 2003). In fact young individuals in sub-Saharan Africa have less knowledge of general contraceptive methods than in any other developing parts of the world (Khan and Mishra, 2008). There is also a troubling lack of access to contraception for many young people due to the social stigma of adolescent sexuality from adults. Nurses at clinics in South Africa, for instance, were found to sometimes block access to contraceptives based on the idea that the individuals requesting them were too young or needed parental permission (Wood and Jewkes, 2006; Ehlers, 2003). This social stigma from adults also results in a lack of communication on sexual matters between them and adolescents, leaving the adolescents to find out information from their peers (Mkhwanazi, 2010).

Pregnancy termination comes under a lot of social stigma throughout sub-Saharan Africa, resulting in them being quite rare (Khan and Mishra, 2008; Panday et al., 2009). In South Africa adolescent pregnancy terminations are often carried out using illegal means even though legal procedures are available, in order to try and hide the pregnancy from family members (Panday et al., 2009). This itself puts pregnant adolescent girls at great risk.

Other contributing factors are gender and age imbalances existing within the culture of the country and the sexual relationships adolescents are participating in. Adolescent women who become pregnant have often become so from a relationship with an older man. Such age gaps can result in power imbalances leading to an increased likelihood of forced sex and pregnancy occurring (Jewkes et al., 2001). Adolescent men are also often reluctant to admit paternity due to the ramifications for their education and employment prospects (Kaufman et al., 2001). This then leaves the burden of the pregnancy on the woman.

In addition to the risks of financial burden, illegal termination, and social stigma, adolescent pregnancy also results in a much greater likelihood of dropping out and not returning to education (Panday et al., 2009; Timæus and Moultrie, 2015). There is also a higher risk of poverty and unemployment (Panday et al., 2009), as well as increased susceptibility to HIV (Panday et al., 2009; Christofides et al., 2014). Beside all that there is an increased risk of neonatal mortality from an adolescent pregnancy (Ramaiya et al., 2014).

Considering all this, adolescent pregnancy, and pregnancy overall, is a pertinent context to perform analyses in. In Section 4.3 we consider the modelling of fertility patterns over different socioeconomic variables, which is important to the general study of populations including that of adolescents. As such we focus on all ages, not simply adolescents, for this analysis. In Section 4.4, following on from the results of Chapter 3, we explore the possibility of social contagion of pregnancy risk behaviour amongst adolescents. The aim of both of these analyses is to provide robust results that will help in understanding adolescent pregnancy in rural South Africa and therefore provide information for the development of policy and interventions in mitigating against it.

In both these analyses we find the need to employ methods other than parametric regression. In exploratory modelling like that performed in Chapter 3 with the aim of inferring causal relationships, using non-linear parametrisation to build further structure and assumptions justified from established knowledge in to the model enables us to more realistically represent the causal phenomena and improve our modelling of it. In descriptive modelling, where we aim to discover patterns within our data, if no established patterns exist in previous knowledge then the use of non-linear parametrisation can enforce unjustified a priori assumptions on to the results. When attempting to address complicated research questions, a justifiable parametrisation can be difficult to formulate from established knowledge. These issues necessitate the ability to employ methods other than just parametric models. As described in Section 2.2, Gaussian process regression and classification are semi-parametric Bayesian statistical methods allowing patterns in the data to become manifest without imposing too much a priori structure on the model.

The first analysis presented takes advantage of the flexibility of Gaussian processes for modelling fertility patterns over socio-economic variables where we have no a priori justification for assuming a strict parametric model of these patterns. The second involves trying to take advantage of the covariance functions of observations that result from Gaussian processes when attempting to model social contagion using data that have no social network information, seeing whether the more complex aspects of the method that help differentiate it from GLMs can increase its usefulness over such methods. Both of these analyses are performed using data from the Agincourt Health and Socio-demographic Surveillance System (HDSS) dataset, an annual update on a baseline census collecting information on the health and socio-demographic status of individuals living within a poor rural region of South Africa.

4.2 The Agincourt HDSS dataset

The Agincourt HDSS is run by the Medical Research Council / University of the Witwatersrand Rural Public Health and Health Transitions Research Unit (Agincourt). Details on its methodology have been published in greater detail elsewhere (Kahn et al., 2012, 2007).

The Agincourt HDSS is an annual update round of a baseline census performed on a poor rural South African population initially in 1992. The HDSS was set up as part of the changes in health services in South Africa post-Apartheid. Its main areas of focus include analysing health, population, and social transitions;

household responses to shocks and stresses; and the effectiveness of policy changes and interventions (Kahn et al., 2012).

The Agincourt HDSS study area is located in North-East South Africa, across the Kruger national park from the Mozambican border (Figure 4.1). It is a former 'Bantustan' area, i.e. a so-called 'homeland' for black people established under the Apartheid regime to enforce racial segregation (SAHO, 2017). The area is characterised by rudimentary sanitation, poor quality education, and poor quality land that makes agricultural farming difficult. Healthcare access is limited, with the closest hospitals being 25-60 km away and only health centres and clinics existing within the area (Kahn et al., 2012). There is high unemployment, electricity access only affordable to the minority, and a consistently rising HIV seroprevalence which is estimated to have changed from 1.7% in 1992 to 25% in 2003 (Kahn et al., 2007). Originally the study area covered 57600 people in 8900 household in 20 villages (Kahn et al., 2007), and by 2011 it had increased to 90000 people in 16000 households in 27 villages (Kahn et al., 2012). A third of the population is formed from Mozambican refugees who came over in the early to mid-1980s in response to their civil war (Kahn et al., 2007).

In each update round demographic data is collected including births, deaths, and migration. Health information is collected at regular bases and since 2006 a new system allows the linkage of census data with morbidity data at the existing Primary Health System in the study area. Various socio-economic characteristics such as education, socio-economic status, and aspects related to food security are collected amongst this health information.

The Agincourt research unit takes great care to involve the community with the HDSS. The fieldworkers who perform the census are trained individuals from the study area. In addition, the LINC (Learning, Information dissemination, and Networking with Community) office undertakes regular meetings with the Community Advisory Group (CAG) to keep the community up to date with findings and operations of the HDSS and to gauge their opinions (Kahn et al., 2012).

Research using this dataset proves valuable at the basic level in simply allowing us to affect policy and improve health for the individuals living within the Agincourt study area. Beyond that, this population can be thought of as representative of poor rural populations existing within sub-Saharan Africa as a whole (though admittedly with generally better welfare), as well as possibly other economically similar countries to South Africa such as India, China, and Russia. Though high quality research has

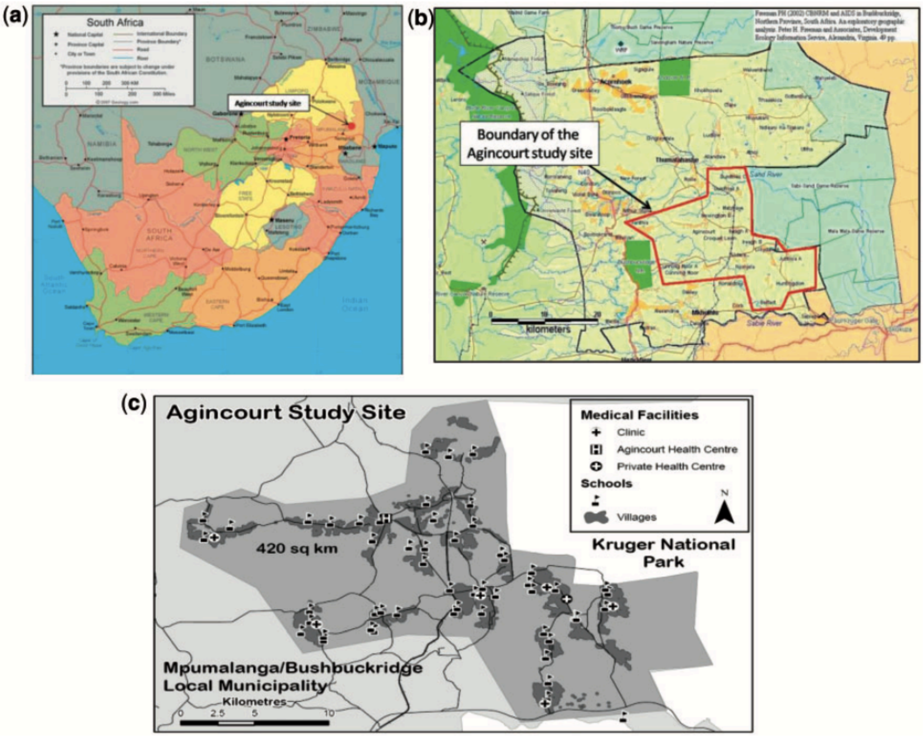


Figure 4.1: (a) Location of the Agincourt Health and Socio-demographic Surveillance System (HDSS) study area in South Africa. (b) Boundary of the Agincourt HDSS study area. (c) Agincourt HDSS study area and sub-district indicating positions of villages and health and education facilities. Reprinted from Kahn et al. (2012) with permission from Oxford University Press.

been done using this dataset, most have failed to involve the use of models suitable for analysing such complex systems as the population dynamics and interactions of household characteristics within the populations, as we shall see in the following analyses in this and the next chapter. This provides us with the opportunity to employ more suitable models to improve the robustness of current results and allow for the findings of new ones.

4.3 Modelling fertility patterns over socio-economic variables

4.3.1 Modelling of fertility rates

The measurement of fertility rates and their relationships to socio-economic variables are essential to the analysis of the population dynamics of a society. For South Africa, whose history of Apartheid has resulted in a very socio-economically diverse population, the ability to examine trends and patterns in fertility is even more important when trying to assess the development of the country. In the last few decades the country has experienced a number of health and demographic shifts including the HIV pandemic, the rise in prevalence of noncommunicable disease (Houle et al., 2014), and the decline over time of fertility itself since 1960 (Burger et al., 2012). The calculation of fertility rates from various data sources across the country and sub-Saharan Africa as a whole has proven useful in looking at the impact of HIV/AIDS (Camlin et al., 2004; Arthur et al., 2013), increased education (Kravdal, 2002), delayed marriage (Arthur et al., 2013), premarital reproduction (Garenne et al., 2000; Palamuleni and Adebowale, 2014; Arthur et al., 2013), contraceptive use (Arthur et al., 2013), and the development of refugee populations (Williams et al., 2013), as well as more administrative issues such as the evaluation of potentially unreliable Apartheid-era data (Moultrie and Timæus, 2003).

Most of this research, similar to what we have seen before in Chapter 3 and section 1.2, has relied on established statistical analysis methods of parametric and generalised linear regression (GLMs), despite more recent innovation in statistical analysis in recent years. Fertility rates are often only examined empirically, leaving the conclusions drawn vulnerable to noise that could exist within the data (Garenne et al., 2007; Kirk and Pillet, 1998; Palamuleni et al., 2007; Nilses et al., 1997; Moultrie and Timaeus, 2002; Garenne and Zwang, 2006). Linear and logistic regression techniques are commonly used, but are very constraining in their assumption of a linear relationship between fertility and (transforms of) the various covariates considered (Camlin et al., 2004; Palamuleni and Adebowale, 2014; Ayele, 2015). Often there is no reason to believe these relationships to be linear at all. A variety of non-linear models for fertility over age have been developed, such as the Hadwiger, Gamma, and Beta functions (Garenne et al., 2000; Peristera and Kostaki, 2007). However, these models fail to incorporate further covariates in anything more than a linear fashion (Burger et al., 2012), and also impose their own strong assumptions (although these are potentially much better justified than generalised linearity).

Here we present a more general method for examining the relationship between fertility and various covariates, focusing on age and socio-economic status, by combining a standard non-linear parametric model of fertility rates over age with the use of Gaussian process regression to bring in further covariates that we do not have well-established models for. In using a parametric model over age, we make sure to capture the non-linear relationship shown to exist between fertility and age in other work (Garenne et al., 2000; Peristera and Kostaki, 2007). Gaussian process regression then allows us to find potentially non-linear relationships between fertility and these other covariates without having to define a precise parametric form to the relationships that would force possibly unfounded assumptions onto the results. We then apply this method to data from the Agincourt HDSS.

4.3.2 Data

We created a dataset out of the Agincourt HDSS by selecting women who were living in a household in the HDSS dataset during the years that socio-economic status data was collected (2001, 2003, 2005, 2007, 2009, and 2011). The inclusion criteria for each year were individuals of all ages who had a recorded date of birth and no date of death proceeding the selected year, and belonged to a household that supplied enough information to calculate an absolute socio-economic status (SES) index. The total sample size was 224643, where an observation was defined as a woman in an individual year who meets all inclusion criteria. Some women were counted as multiple observations due to appearing in the census dataset in multiple years. Though we did consider the inclusion of other covariates such as education (measured in number of years of education achieved), our analysis focused on fertility (defined as the fraction of women associated with each set of covariate values who experienced a live birth) over age (measured in years) and SES (measured by Agincourt's household absolute SES index, which averages a set of quantitative measures of the amounts of different types of assets the household possesses (Collinson et al., 2009)), both of

Age	SES	Year					
		2001	2003	2005	2007	2009	2011
10-20	1-2	3063	2883	1772	1033	828	788
	2-3	5333	5784	6497	6402	8097	6769
	3-4	370	362	356	651	1156	1819
20-30	1-2	2046	1915	1183	716	578	570
	2-3	3718	4291	5214	5223	6980	6305
	3-4	254	241	279	567	1125	1832
30-40	1-2	1537	1331	827	502	445	409
	2-3	2644	2960	3439	3427	4448	3842
	3-4	187	210	178	401	678	1104
40-50	1-2	990	911	558	337	278	275
	2-3	1631	1797	2115	2392	3067	2598
	3-4	115	105	108	247	482	809

Table 4.1: Sample sizes for different ranges of age (in years) and socio-economic status for each year, given to aid comparison of the analytical results to the data. Each value is given for chosen example intervals of age and SES values taken from an overall continuous range, where the lower age/SES value of the interval is inclusive and the upper age/SES value is exclusive.

which we measured at the midyear point for each year.

In order to calculate sensible values for the empirical fertility rates so that it could be used as the dependent variable of a regression, we binned the observations to set covariate values by splitting them into quantiles and then setting their covariate values to the midpoints for the quantiles they belong to. The precise number of quantiles used for binning each covariate was chosen by a combination of cross-validation and goodness-of-fit techniques, more detail of which is given in the description of the model below. In the end, the preferred quantiles were 125-quantiles for age and 25-quantiles for SES.

Examples of sample sizes and average fertility rates in our chosen dataset for various age and SES ranges in each year are shown in Tables 4.1 and 4.2.

Age	SES	Year					
		2001	2003	2005	2007	2009	2011
10-20	1-2	0.0501	0.0501	0.0608	0.0580	0.0724	0.0516
	2-3	0.0386	0.0352	0.0504	0.0472	0.0519	0.0408
	3-4	0.0355	0.0423	0.0349	0.0310	0.0497	0.0320
20-30	1-2	0.1006	0.1140	0.1296	0.1357	0.1269	0.1252
	2-3	0.0998	0.0959	0.1059	0.1060	0.1104	0.1120
	3-4	0.0645	0.0969	0.0777	0.0919	0.1057	0.1021
30-40	1-2	0.1000	0.0795	0.0918	0.0885	0.0846	0.1281
	2-3	0.0912	0.0754	0.0944	0.0756	0.0773	0.0936
	3-4	0.0798	0.0749	0.0912	0.0897	0.0814	0.0715
40-50	1-2	0.0355	0.0310	0.0206	0.0322	0.0273	0.0191
	2-3	0.0156	0.0242	0.0157	0.0179	0.0183	0.0207
	3-4	0.0077	0.0111	0.0077	0.0229	0.0223	0.0120

Table 4.2: Average fertility rates for different ranges of age (in years) and socio-economic status for each year, given to aid comparison of the analytical results to the data. Each value is given for chosen example intervals of age and SES values taken from an overall continuous range, where the lower age/SES value of the interval is inclusive and the upper age/SES value is exclusive.

4.3.3 Method

In order to obtain insight into what is happening within our dataset, we relied on regression methods where fertility rate acted as our dependent variable and covariates such as age, SES, and education acted as independent variables. Though there are no generally accepted non-linear models of fertility over the other covariates, some have been described for fertility over age typically using functions that result in a bell curve with a peak skewed to lower ages (Peristera and Kostaki, 2007). Indeed such a relationship can be seen in both the kernel density estimate of women experiencing live births over various years (Figure 4.2) and in plots of the empirical fertility rates calculated for individuals grouped into age centiles (Figure 4.3). In order to be certain that we captured this relationship, we used a parametric model for fertility over age and incorporated further covariates by allowing the parameters of our parametric model to be dependent on the other covariates. Various work has shown the age-pattern of fertility to contain a secondary earlier age peak credited to premarital fertility (Garenne et al., 2000). However, as our data does not show significant evidence of this second peak (perhaps due to the nature of the binning we used) we therefore chose the Gamma distribution, a standard model for fertility over age (Peristera and Kostaki, 2007), as our parametric form for our fertility rate for individual i, $p_i\left(a, \mathbf{x}\right)$, over age a and dependent on further covariates $\mathbf{x}$, i.e.

$$\begin{aligned} p_i\left(a; \mathbf{x}\right) &= \Pr\left(Y_i = 1 | a, \mathbf{x}\right) \\ &= \text{Gamma}\left(a | \alpha\left(\mathbf{x}\right), \beta\left(\mathbf{x}\right)\right) \\ &= \frac{a^{\alpha(\mathbf{x})-1} e^{-a/\beta(\mathbf{x})}}{\Gamma\left(\alpha\left(\mathbf{x}\right)\right) \beta\left(\mathbf{x}\right)^{\alpha(\mathbf{x})}} \end{aligned} \tag{4.1}$$

where the fertility indicator Y_i of individual i is equal to 1 if the individual experienced a live birth for covariates a and $\mathbf{x}$ and equal to 0 otherwise, $\Gamma\left(\cdot\right)$ is the gamma function, and $\alpha\left(\mathbf{x}\right)$ and $\beta\left(\mathbf{x}\right)$ are our shape and scale parameters which depend on our other covariates. It should be noted here that the Gamma distributions used here are normalised over age for each combination of the other covariate values. Whilst the shape of the relationships inferred will be accurate, this normalisation should be taken into account when wanting to use any inferred numerical values for the fertility rates in decision making. It should also be noted that this model represents only interaction effects between each of the other covariates and age, but not any main effects from each individual covariate.

For the functional forms of $\alpha\left(\mathbf{x}\right)$ and $\beta\left(\mathbf{x}\right)$, due to the lack of established models, we

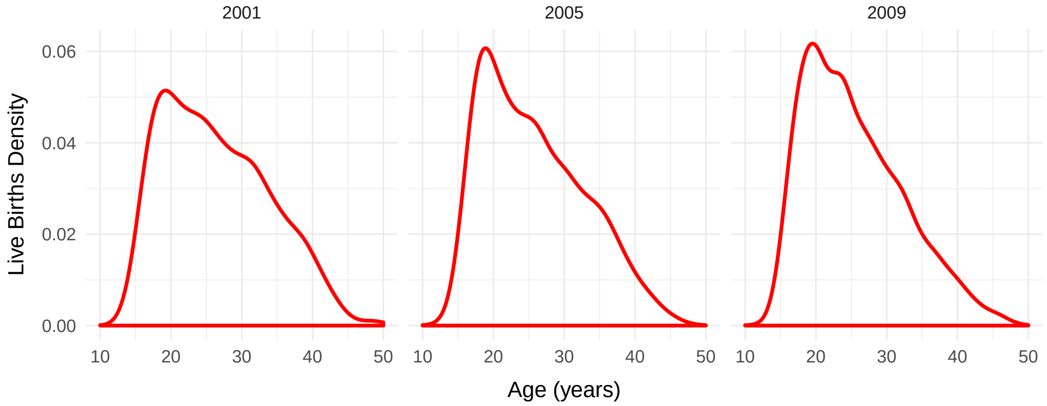


Figure 4.2: Non-parametric kernel density estimate of the distribution over age of women experiencing live births in the years 2001, 2005, and 2009 in the Agincourt health and socio-demographic surveillance system (HDSS) study area in rural South Africa. The distributions show the standard skewed hill-shaped age-pattern for fertility as found in most other work.

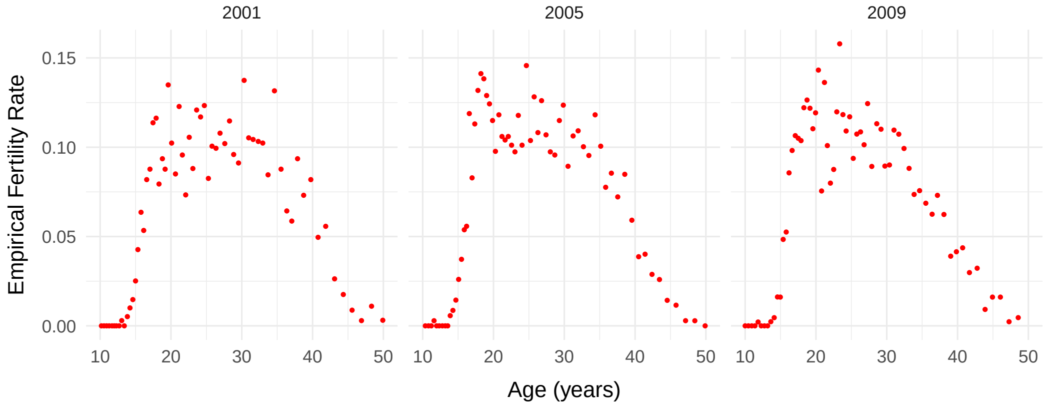


Figure 4.3: Empirical fertility rates, i.e. fractions, of the number of women experiencing live births for each age centile, calculated for individuals living in the Agincourt health and socio-demographic surveillance system (HDSS) study area in rural South Africa in the years 2001, 2005, and 2009. The empirical probabilities show the standard skewed hill-shaped age-pattern for fertility as found in most other work.

employed Gaussian process regression for its flexibility and potential non-linearity. A description of Gaussian process regression can be found in Section 2.2. As it is a standard choice that satisfies our needs of resulting in smooth continuous functional forms for $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ we use the squared exponential covariance function

$$\boldsymbol{K}(\mathbf{x}_i, \mathbf{x}_j) = \sigma_f^2 \exp\left[-\frac{1}{2}(\mathbf{x}_i - \mathbf{x}_j)^T \boldsymbol{M}(\mathbf{x}_i - \mathbf{x}_j)\right] \tag{4.2}$$

for both. The covariance function parameters (i.e. the lengthscales and signal and noise variances) were fitted by maximisation of the marginal log likelihood.

One way to think of this method of combining Gaussian process regression with parametric regression is that the Gaussian process regression smooths over the other covariates the parameters of our model for the role of age. Fitting Gaussian distributions of functions to the estimates of $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ found from parametrically fitting over age allows the data to both give an initial noisy estimate of the functional forms of $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ and then to smooth over them by defining, given these initial estimates, the magnitude and frequency of how $\alpha(\mathbf{x})$ and $\beta(\mathbf{x})$ vary over $\mathbf{x}$.

We guarded against overfitting of the Gaussian process through use of a smoothing prior for the length parameter for SES, a gamma distribution with shape parameter 6 and scale parameter 0.25. Wider and thinner priors were also tried to see what effect the prior choice had on the results, but little to no differences were found.

In order to find which combination of covariates would be best to include in the model, as well as to decide on how many quantiles should be used for binning as described in the previous section, a combination of cross validation and goodness-of-fit tests were used. To measure the predictive performance of each possible model choice 10-fold cross validation was used, where the performance was measured by their Briers score (Brier, 1950)

$$S_B = \frac{1}{N}\sum_{i=1}^{N}(Y_i - p_i(a, \mathbf{x}))^2 \tag{4.3}$$

which compares the fitted probability of fertility to the actual fertility status of each of the N observations. Due to the treatment of the problem as a regression, and therefore having to bin the data, goodness-of-fit tests to the unbinned data could not be performed. Instead we performed Kolmogorov-Smirnov (KS) tests (Wasserman, 2013) comparing the fitted marginal fertility probabilities over age (with cumulative distribution function $\hat{F}(a)$) to the empirical marginal fertility rates over age (with cumulative distribution function $F_n(a)$) using test statistic

$$D_n = \sup_a \left| F_n(a) - \hat{F}(a) \right| . \tag{4.4}$$

Model choices were rejected at a 5% significance level, where the Bonferonni method of correcting this significance level by a factor of one over the number of tests performed was used to mitigate against the possibility of rejecting by chance due to performing a large number of tests (Bland and Altman, 1995). It was found that, though it was possible to include education in addition, it was best to simply focus on age and SES. This is down to two reasons. First, there are more missing values for the Agincourt HDSS education data than the SES data. Second, that introducing more covariates leads to worse fits when using maximum likelihood on the parametric model over age. Results of the cross validation and statistical tests can be found in Appendix B.

4.3.4 Results

The resulting fitted forms for fertility rate over age and SES for a range of years between 2001 and 2011 can be seen in Figures 4.4 and 4.5. Figure 4.4 shows how fertility rate varies over age. We can see that the model has captured the standard skewed relationship, as we would expect from our choice of parametric model. Fertility increases rapidly from mid-adolescence to peak in mid- to late-20s, before more gently decreasing until early-50s. We observed little age-pattern fertility changes for different SES values. However, there is a slight increase in peak fertility over age as time increases, which goes against the fertility decline which has been observed elsewhere (Burger et al., 2012). This is an interesting detail as it appears to contradict work done on fertility trends over time that have found a significant decline in fertility over the past several decades (Burger et al., 2012; Kirk and Pillet, 1998; Moultrie and Timaeus, 2002; Moultrie and Timæus, 2003; Garenne and Joseph, 2002). This could simply be a quirk of the data (such as the result of recording errors or other such biases), or could be due to some local phenomenon happening in the study area in recent years. On the other hand it is consistent with an increase in peak fertility we see in the raw data (Figures 4.2 and 4.3).

The relationship between fertility rate and SES can be seen in Figure 4.5 for a variety of ages and years. It must be first noted that fertility rate varies little over SES for any year or age, which would be consistent with how homogeneous

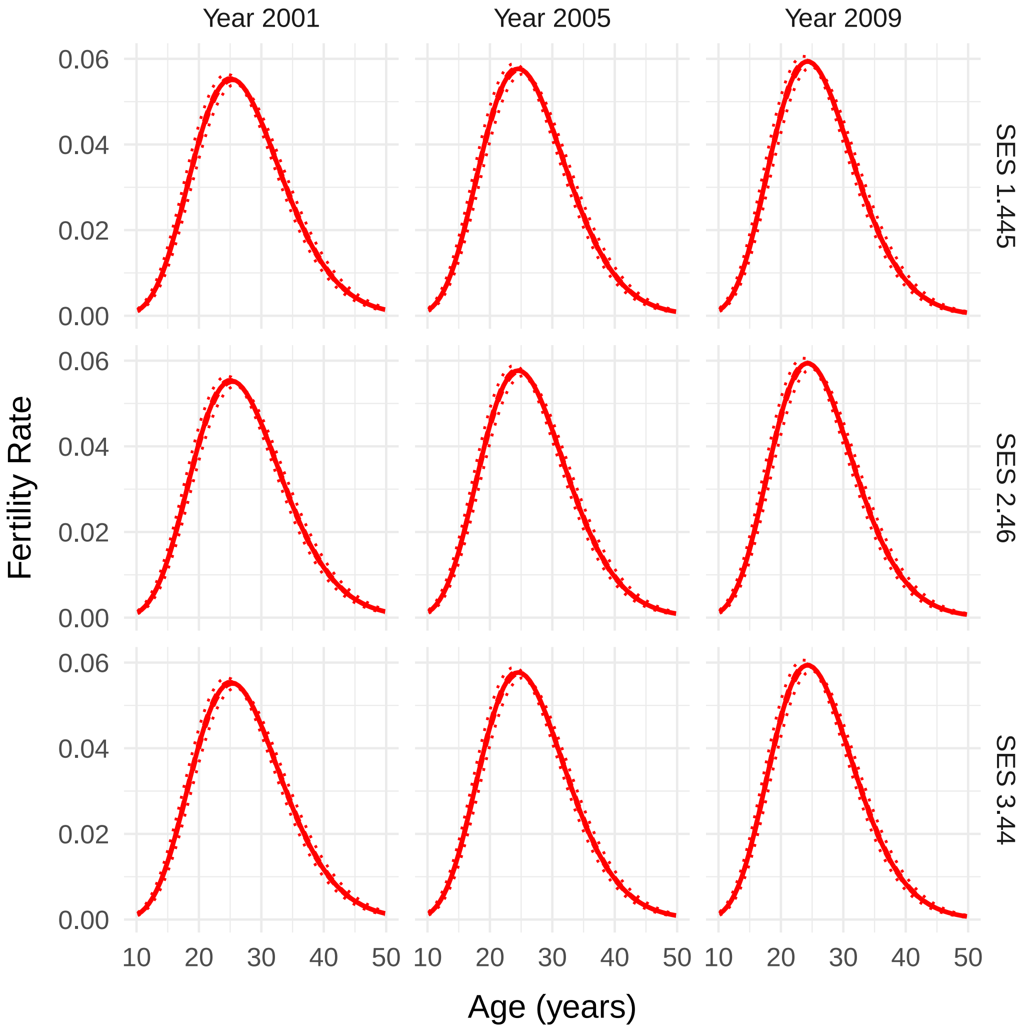


Figure 4.4: Fertility rate over age as fitted by our combined parametric and semi-parametric model, for socio-economic status values of 1.445, 2.46, and 3.44, and years 2001, 2005, and 2009. Parametrically bootstrapped confidence intervals (from 1000 samples of the model) are shown for the 50% level (dashed lines) and 95% level (dotted lines). The model has managed to capture the standard skewed hill-shape of the age-pattern as found in the raw data and in many fertility age-patterns in the literature.

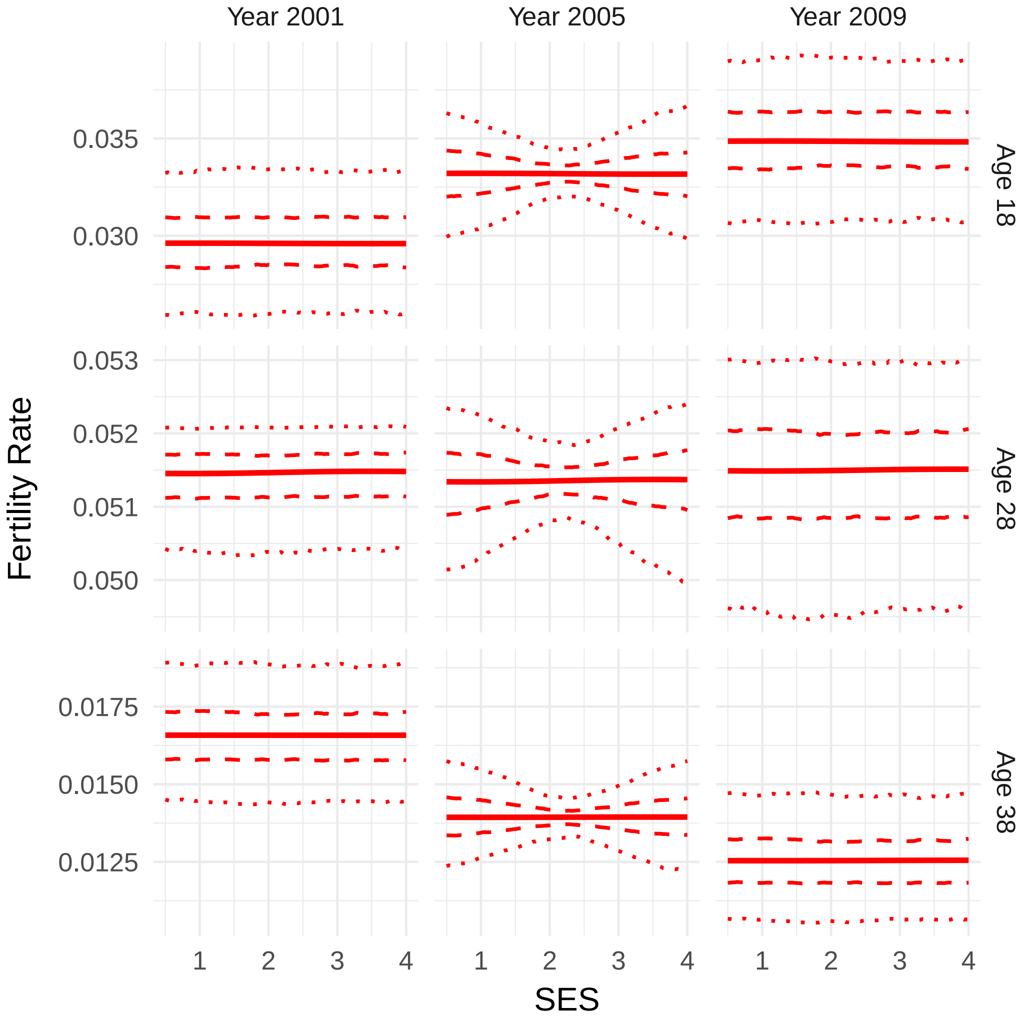


Figure 4.5: Fertility rate over socio-economic status (SES) as fitted by our combined parametric and semi-parametric model, for age values of 18, 28, and 38, and years 2001, 2005, and 2009. Parametrically bootstrapped confidence intervals (from 1000 samples of the model) are shown for the 50% level (dashed lines) and 95% level (dotted lines). The model manages to capture various details of the fertility-SES pattern, such the time trends of fertility for the different age groups.

we know the individuals in the study area to be. Overall it would appear to be almost constant over SES for the individual age-year combinations. The fertility-SES relationship varies over time but differently for different ages, where the plots suggest that adolescent pregnancy has actually increased over time and late-30s pregnancy has actually decreased over time. Late-20s pregnancy, which is closer to the peak in the fertility age-pattern, has a very small magnitude of variation over time with no consistent trend.

We also performed the same analysis (using the same covariates and quantiles) on the individual refugee and non-refugee populations, achieved by splitting the dataset into refugee and non-refugee sub-datasets according to whether the individuals were marked out as refugees in the dataset or not. Significant differences have been shown in fertility levels between the Mozambican refugee population who came over to the study area in response to the civil war, and the South African population, though it has been shown that the populations have converged in recent years (Williams et al., 2013). The results here (Figures 4.6 and 4.7) do show some differences, though indeed the fertility patterns of each population appear to have become quite similar. The non-refugees show the same increase in peak fertility as the overall population, whilst the refugees have lower fertility at both earlier and later years. The fertility patterns over SES remain constant, and for individuals in their late teens and late-30s we see the same patterns over time as we have before. However, there is more variation over time for individuals in their late-20s for both populations, though in opposite directions, with refugee fertility increasing substantially in 2005 before settling back down again by 2009 and non-refugee fertility varying in the opposite direction before coming back as well. It should be noted that there is a severe overlap in the confidence intervals of the fertility SES-patterns for most ages and years. Overall this implies that there is not a great deal of variation from the overall population when differentiating by refugee status, consistent with the convergence of fertility in the refugee and non-refugee populations.

By combining a parametric regression of fertility rate over age with the use of Gaussian process regression to bring in further covariates such as SES, we produce an improvement in robustness to the modelling of fertility. The parametric part of our model successfully captures the well known skewed hill relationship between fertility and age that can be seen both in empirical plots of our own data shown in Figures 4.2 and 4.3, as well as in many other research papers that have used empirical calculations or other similar parametric models to model the fertility age-patterns of sub-Saharan Africa (Camlin et al., 2004; Garenne et al., 2007, 2000; Kirk

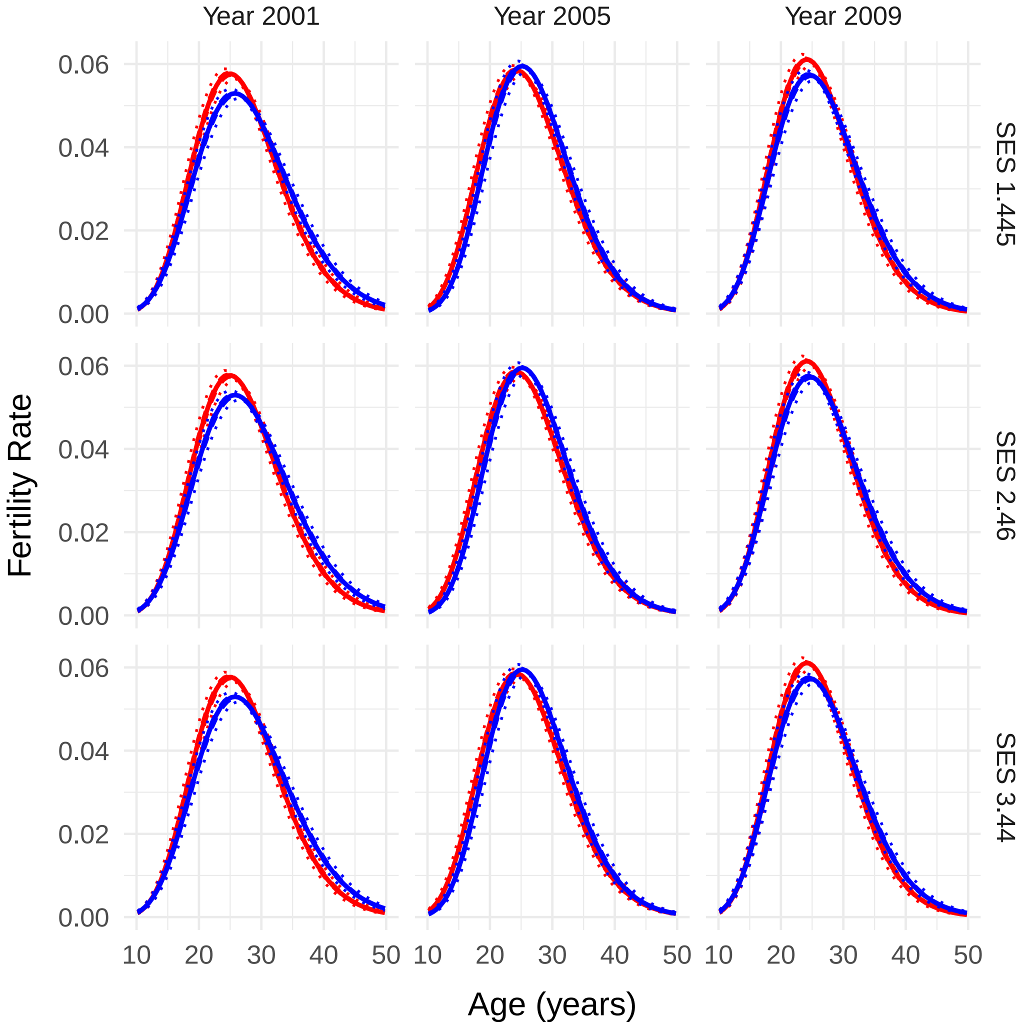


Figure 4.6: Fertility rate over age as fitted by our combined parametric and semi-parametric model, for socio-economic status values of 1.445, 2.46, and 3.44, and years 2001, 2005, and 2009, for the refugee (blue) and non-refugee (red) populations of Agincourt. Parametrically bootstrapped confidence intervals (from 1000 samples of the model) are shown for the 50% level (dashed lines) and 95% level (dotted lines). These fertility age-patterns only show slight differences between the populations and the overall population, reflecting the convergence of fertility between them.

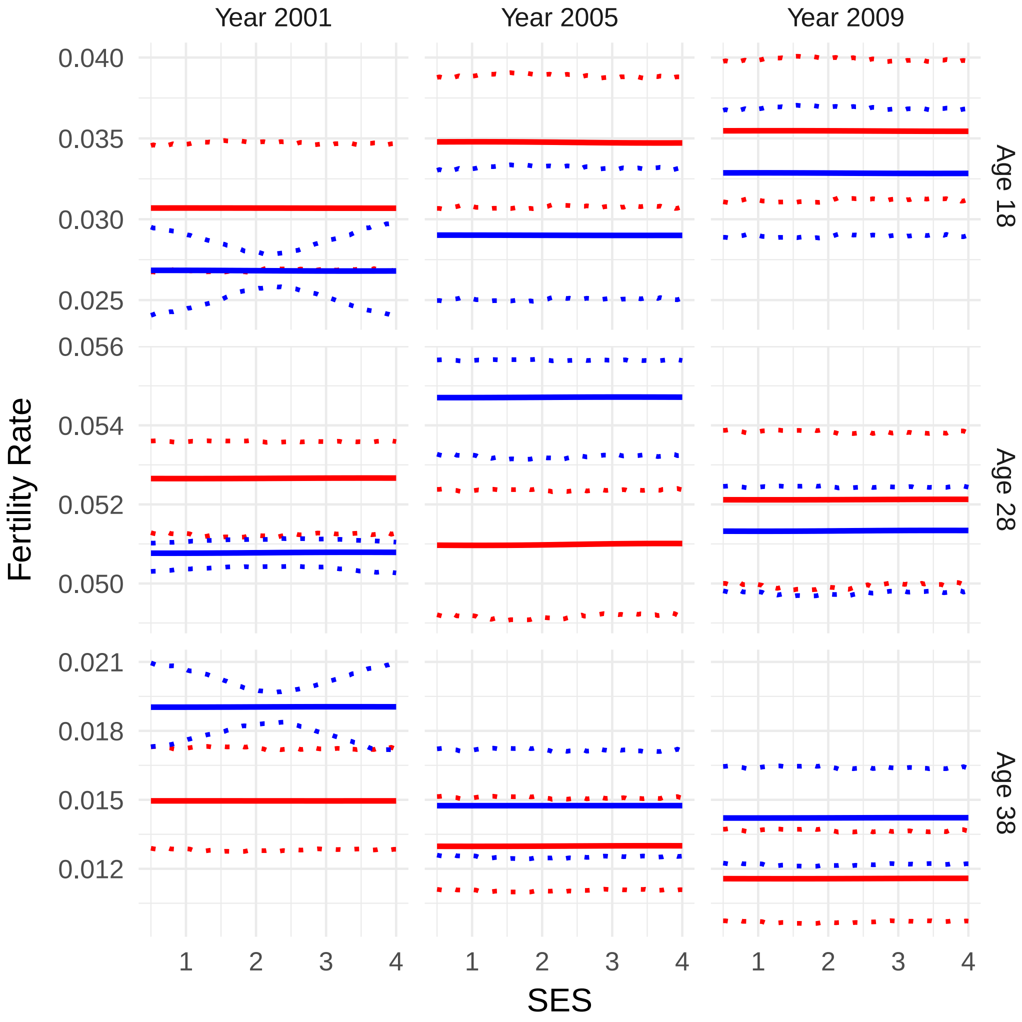


Figure 4.7: Fertility rate over socio-economic status (SES) as fitted by our combined parametric and semi-parametric model, for age values of 18, 28, and 38, and years 2001, 2005, and 2009, for the refugee (blue) and non-refugee (red) populations of Agincourt. Parametrically bootstrapped confidence intervals (from 1000 samples of the model) are shown for the 95% level (dotted lines). Some slight differences between the refugee and non-refugee populations are shown, particularly in variation of fertility over time for 28 year old individuals. However, these differences are slight, reflecting the convergence of fertility between the two populations.

and Pillet, 1998; Peristera and Kostaki, 2007; Moultrie and Timaeus, 2002; Garenne and Zwang, 2006).

The semi-parametric part of our model, using Gaussian process regression over other covariates, successfully manages to model the SES pattern of fertility without simply assuming the relationship to be linear as other work has done (Burger et al., 2012; Camlin et al., 2004; Ayele, 2015). This gives the potential to capture more detail within the relationship and provide greater insight in to what has been happening to fertility in the Agincourt study area between 2001 and 2011. We found that the magnitude of variation of fertility over SES is quite small, suggesting that SES does not have as big an impact on fertility as we would think, reflecting the homogeneity of the population built into the model by the smoothing prior. In fact it would appear to be almost constant, and certainly quite linear. Though this means that a GLM could have been used in this case, the incorporation of this modelling technique into fertility modelling is still useful. We had no a priori justification for a linear model, and using one would definitely have restricted our results such that we would have no chance of capturing possible non-linearities. It is also bad practise to justify heavy assumptions a posteriori on the relationships we are modelling. For other similar modelling problems where linear models are commonly used, and for further fertility modelling itself in other types of populations, this method allows for much more relaxed assumptions about relationships where we have no a priori justification for stricter assumptions. Otherwise, our results have also shown that adolescent fertility does appear to have increased over time, whilst later life fertility appears to have decreased. Overall, the flexibility and non-linearity of the method allows for the potential capture of much more information than a single linear coefficient can show, and therefore increases the robustness of the results.

4.4 Social contagion of adolescent pregnancy risk

4.4.1 Social contagion without social network data

Research on social contagion of pregnancy risk behaviours is scarce. Instead of considering populations where pregnancy has a more negative impact, such as adolescents in rural South Africa, most of the few studies that exist in this area have considered only young couples in central Europe (Lois and Becker, 2014; Bühler and Fratczak, 2007; Bernardi et al., 2007; Keim et al., 2009) or women in the work place (Pink et al., 2014). Much of the work uses either qualitative or mixed methods approaches and focus on specific mechanisms of the spread of behaviour such as social learning (i.e. an individual forming decisions based on information gained from their social network) and social influence (e.g. social norms, peer pressure, social conformity etc) rather than attempting to build a quantitative model of social contagion in general (Bernardi et al., 2007; Bernardi, 2003; Keim et al., 2009). The work done on implementing quantitative models commonly uses generalised linear models similar to the research into mood and weight contagion discussed in Chapter 3 and are therefore subject to the same criticisms (Lois and Becker, 2014; Bühler and Fratczak, 2007; Pink et al., 2014; Montgomery and Casterline, 1996).

Considering this gap in the literature concerning social contagion of pregnancy risk in rural South African, or sub-Saharan African, communities, as well as the severe consequences adolescent pregnancies can have for individuals living in such communities (as discussed in Section 4.1), it is of great interest to see if the social contagion model presented in Chapter 3 can be used to find whether evidence of social contagion of pregnancy risk does or does not exist within this data. The difficulty with achieving this is in the lack of social network data present within the Agincourt HDSS dataset. Whilst the Add Health study (Chapter 3) collected information on the friends of each individual, the Agincourt HDSS contains only basic information on who an individuals parents are, who their partner is, birth events, household membership, and their relationship to the head of their household. Such information could be used to construct a social network but there is no ground to assume such a network would be complete. Also such a network would largely not be useful for modelling social exposure of adolescents to each other, only to their family members. This problem is not unique to Agincourt. It is generally not commonplace to collect social network data outside of the environment of online social networks (where network information can be very easily extracted).

The social contagion model discussed in Chapter 3 considered state change transition probabilities (which in this case would be the probability of becoming pregnant between two time steps) stratified by the number of contagious state friends k (in this case the number of pregnant friends) at the first time point of the transition. In the following analysis we attempt to overcome the issue of no social network data by replacing the parametric model used previously with a Gaussian process classification. We then replace k with a contagion variable measuring an individuals likelihood of exposure to the contagious behaviour. This is done so by taking advantage of the covariance matrix present in the Gaussian process classification. We then apply this model to pregnancy amongst adolescents in the Agincourt HDSS dataset.

In doing so we aim to gain insight into the social mechanisms behind adolescent pregnancy in the Agincourt study area, which can then inform interventions against adolescent pregnancy. We also aim to show the usefulness in applying Gaussian processes in situations where designing a precise functional form is difficult, therefore generally showing how the use of methods with more complex features opens up the possibility of researching questions we could not before.

4.4.2 Pregnancy risk contagion model

Due to the lack of social network data in the Agincourt HDSS dataset, which makes it difficult to assign a precise functional form to the transition probability of becoming pregnant, we employ the method of Gaussian process classification as detailed in Section 2.2. For an individual i at a given time point t we model their probability of becoming pregnant over t

$$\pi_i\left(t\right) = \Pr\left(Y_i\left(t\right) = 1 \mid \mathbf{x}_i\left(t\right)\right) \ , \tag{4.5}$$

where $Y_i\left(t\right) \in \{0, 1\}$ is their pregnancy status at t and $\mathbf{x}_i\left(t\right)$ are their values for whatever covariates are included in the model, as a latent function f modelled by a Gaussian process then passed through a link function $\sigma : \mathbb{R} \mapsto [0, 1]$ (in this case an inverse probit function) as is standard for Gaussian process classification. We do this in two different ways. First with dependency only on a set of established pregnancy risk factors, i.e. the covariates $\mathbf{x}_i$, such that Model 1 is

$$\pi_i\left(t\right) = \sigma\left(f\left(\mathbf{x}_i\left(t\right)\right)\right) \ . \tag{4.6}$$

Second with dependency on both these covariates and a new pregnancy contagion variable $C_i\left(t\right)$, i.e. Model 2 is

$$\pi_i\left(t\right) = \sigma\left(f\left(\mathbf{x}_i\left(t\right), C_i\left(t\right)\right)\right) \ , \tag{4.7}$$

where we define the contagion variable either as

$$C_i\left(t\right) = \sum_{i \neq j} Y_j\left(t-1\right) \boldsymbol{K}_{t-1}\left(\mathbf{x}_i\left(t\right), \mathbf{x}_j\left(t-1\right)\right) \tag{4.8}$$

for global contagion, i.e. someone can be affected by a pregnancy occurring anywhere within the study area depending on the similarity to that individual, or as

$$C_i\left(t\right) = \sum_{i \neq j} \mathbb{1}\left(V_i = V_j\right) Y_j\left(t-1\right) \boldsymbol{K}_{t-1}\left(\mathbf{x}_i\left(t\right), \mathbf{x}_j\left(t-1\right)\right) \tag{4.9}$$

for local contagion, where we restrict the contagion variable for individual i to influences from pregnancies occurring only within the same village as them V_i with

$$\mathbb{1}\left(V_i = V_j\right) = \begin{cases} 1 & \text{if } V_i = V_j \\ 0 & \text{if } V_i \neq V_j \end{cases}. \tag{4.10}$$

Here $\boldsymbol{K}_{t-1}\left(\cdot,\cdot\right)$ is the covariance function for Model 1 fitted from the previous time point. Therefore $C_i\left(t\right)$ effectively measures the similarity of individual i in terms of pregnancy risk to all the pregnant individuals from the previous time point either throughout the study area (global contagion) or in their village (local contagion).

This is a very precise and constraining definition of contagion, i.e. that an individual is most likely to be influenced by those in their community exhibiting similar risk traits to them, but it is also the most practical one. It enables us to take advantage of the covariance matrix from the Gaussian process classification therefore preventing us from having to invent some more definite measure and a parametric functional form for $\pi_i\left(t\right)$ to go with it. Doing that would involve making a great deal of assumptions about what is going on within the data, whilst this method requires us to make minimal assumptions beyond what we already make when modelling social contagion (that the probability of becoming pregnant would be stratified by the level of influence from other pregnancies if contagion is indeed occurring). It is worth seeing how this method performs as it presents an innovative combination of two already innovative methods in the form of Gaussian processes and the social contagion model we have used up until now. Using pregnancy risk variables when considering similarity is largely a practical measure due to the fact that such risk factors are already established in the literature (see Section 4.4.3) whilst alternatives such as risk factors of exposure to behaviour are not so well established.

The modelling procedure involves first fitting Model 1 to the data from some initial time point, and then proceeding to fit both Models to the data from each time point, using the resulting fit of Model 1 from the previous time point to form the contagion variable for Model 2 in the current time point. We consider two different time periods between our time points: one year (annual fit) and two years (biennual fit). The biennual fit may be using a time period too long for causal effects to take place. On the other hand, due to the continuous occurrences of pregnancy events in the Agincourt HDSS dataset, there is a risk of overlap between the pregnancy of individual i at time t and pregnancies that occurred for other individuals at the previous time point. Trying both of these time periods will help safeguard against the issues both present.

Due to the rather experimental nature of the methodology we are using here, we expand beyond simply using the standard choice of a squared exponential covariance function

$$\boldsymbol{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sigma_f^2 \exp\left[-\frac{1}{2}\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \boldsymbol{M}\left(\mathbf{x}_i - \mathbf{x}_j\right)\right], \tag{4.11}$$

instead fitting and comparing the performance of four contending covariance functions. These include the squared exponential, as well as the Matern 3/2

$$\boldsymbol{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sigma_f^2 \left(1 + \sqrt{3\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \boldsymbol{M}\left(\mathbf{x}_i - \mathbf{x}_j\right)}\right) \exp\left(-\sqrt{3\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \boldsymbol{M}\left(\mathbf{x}_i - \mathbf{x}_j\right)}\right), \tag{4.12}$$

the Matern 5/2

$$\begin{aligned}\boldsymbol{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) =& \sigma_f^2 \left(1 + \sqrt{5\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \boldsymbol{M}\left(\mathbf{x}_i - \mathbf{x}_j\right)} + \frac{5}{3}\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \boldsymbol{M}\left(\mathbf{x}_i - \mathbf{x}_j\right)\right) \\ &\times \exp\left(-\sqrt{5\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \boldsymbol{M}\left(\mathbf{x}_i - \mathbf{x}_j\right)}\right),\end{aligned} \tag{4.13}$$

and the rational quadratic

$$\boldsymbol{K}\left(\mathbf{x}_i, \mathbf{x}_j\right) = \sigma_f^2 \left(1 + \frac{\left(\mathbf{x}_i - \mathbf{x}_j\right)^T \boldsymbol{M}\left(\mathbf{x}_i - \mathbf{x}_j\right)}{2\alpha}\right)^{-\alpha}. \tag{4.14}$$

As each of these are isotropic they will result in smoothly varying continuous functions, increasing the interpretability of both the resulting functions and contagion variables. Definitions of the various covariance hyperparameters can be found in Section 2.2. The hyperparameters are fitted to the data using the Laplace approximation for Gaussian process classification, detailed in both Section 2.2 and Rasmussen and Williams (2006). We then choose the preferred covariance function by comparison of their Bayesian Information Criterion (BIC) values

$$\mathrm{BIC} = \nu \log N - 2 \log \hat{L} \tag{4.15}$$

where ν is the number of hyperparameters, N is the sample size, and $\hat{L}$ is the fitted marginal likelihood (Schwarz et al., 1978).

In order to compare between Model 1 and Model 2 fitted to each time point we measure the performance of each method in predicting pregnancies occurring in the following time point. A standard method of doing this would be by calculating the area under the ROC (receiver operating characteristic) curve (AUROC). However, it has been suggested that as the AUROC value does not take into account the predicted risk values it is therefore a poor choice for examining the contribution of covariates to the predicted risk (Cook, 2007), which is essentially what we are attempting here in examining the impact of adding a contagion variable. We also avoid using a likelihood based measure such as information criteria due to our use of the BIC in comparing covariance functions. We therefore resort again to using the Brier's score (Brier, 1950)

$$S_B\left(t\right) = \frac{1}{N}\sum_{i=1}^{N}\left(Y_i\left(t\right) - \pi_i\left(t\right)\right)^2 . \tag{4.16}$$

The combination of this and the contagion variable lengthscale (recalling this measures the relevance of the contagion variable to the fitted pregnancy probability) give us a satisfactory indicator as to whether the contagion variable contributes a significant enough increase in predictive performance to indicate evidence of contagion of pregnancy risk occurring within the dataset.

4.4.3 Data

Common pregnancy risk factors for adolescents include low socioeconomic status (SES), low education, and belonging to a minority group (Pradhan et al., 2015), as well as whether they have had any previous pregnancies (implied by expert knowledge from members of the Agincourt research unit). Based on the relationship observed between age and fertility rates in Section 4.3, we also include age as a risk factor. These then act as our five covariates.

Our dataset was comprised of data from the Agincourt HDSS dataset captured in the years between 2001 and 2011 inclusive, restricted to women aged 15-19 years and living in the study area in their respective year of inclusion. SES and age were calculated from the data in the same way as described in Section 4.3.2. Education was measured in years completed, and in the census was collected in 2002, 2006, and 2009. We used the closest previous years data for education for each individual in each respective year of analysis. Whether the individual had South African national

Year	Sample size
2001	796
2002	3909
2003	4194
2004	4223
2005	4251
2006	4212
2007	4368
2008	4888
2009	5126
2010	4935
2011	4639
Total	45541

Table 4.3: Sample sizes for each year.

status or not was taken from their refugee status listed in the data. As for their pregnancy status (i.e. the outcome) at the year of analysis t, to have value $Y_i = 1$ they must have experienced a pregnancy event with any outcome (i.e. live birth, still birth, abortion etc) listed as coming to fruition any time within the year of analysis, otherwise $Y_i = 0$. For their status of ever having had a previous pregnancy, this had value 1 if they had experienced any pregnancy events of any outcome that had come to fruition any time previous to the start of the year of analysis, and was otherwise 0. Individuals were only included if they had complete data on age, SES, and refugee status, as well as having provided education data at some point either up to or including the year of analysis.

This resulted in a sample size of 45541 split up amongst the different years, a breakdown of which can be seen in Table 4.3.

4.4.4 Results

The preferred covariance function by BIC in the majority of cases was found to be the rational quadratic (4.14) (Appendix C), and using an annual contagion period produced an increase in predictive performance above that found when using a biennual contagion period, so the results presented here are for these specific cases. Further results for a biennual contagion period can be found in Appendix D.

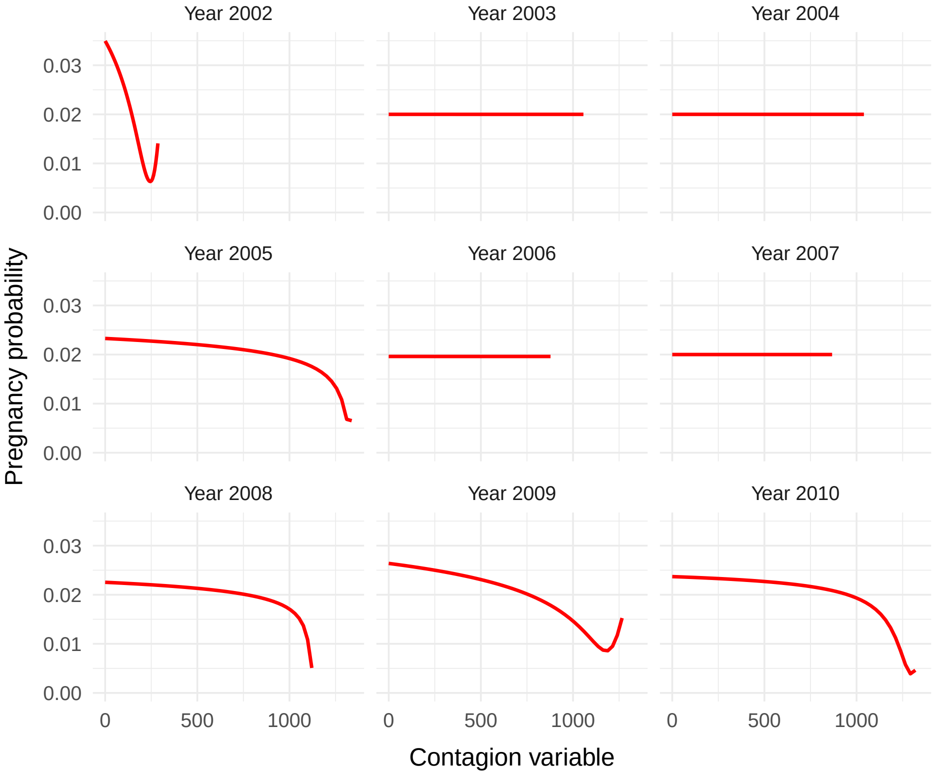


Figure 4.8: Fitted marginal pregnancy probabilities over contagion variable value for each year from the global contagion analysis (with the contagion variable ranging up to the maximum value within the data). Few years show much variation, but those that do show an increase in influence from past pregnant individuals appearing to cause a decrease in the likelihood of becoming pregnant.

Year	No contagion S_B	Contagion S_B	Contagion lengthscale
2002	0.0771	0.0775	3.512
2003	0.0728	0.0728	2137
2004	0.0732	0.0732	4362
2005	0.0877	0.0882	0.6110
2006	0.0750	0.0750	384.8
2007	0.0773	0.0776	4251
2008	0.0864	0.0866	0.7619
2009	0.0869	0.0871	1.568
2010	0.0735	0.0737	3.683
Average	0.0789	0.0791	3.683

Table 4.4: Comparison between Model 1 (no contagion) and Model 2 (contagion) from the global contagion analysis for each year shown by the Briers scores (S_B) from predicting the pregnancies for the next timestep. Contagion lengthscale values from the Gaussian process classification for Model 2 are also included. There is little difference in performance between Model 1 and Model 2. The lengthscales vary greatly, but are mostly greater than 1 implying the contagion variable has little relevance to the shape of the pregnancy probability. Note the average lengthscale is the median, whilst for Briers scores they are means.

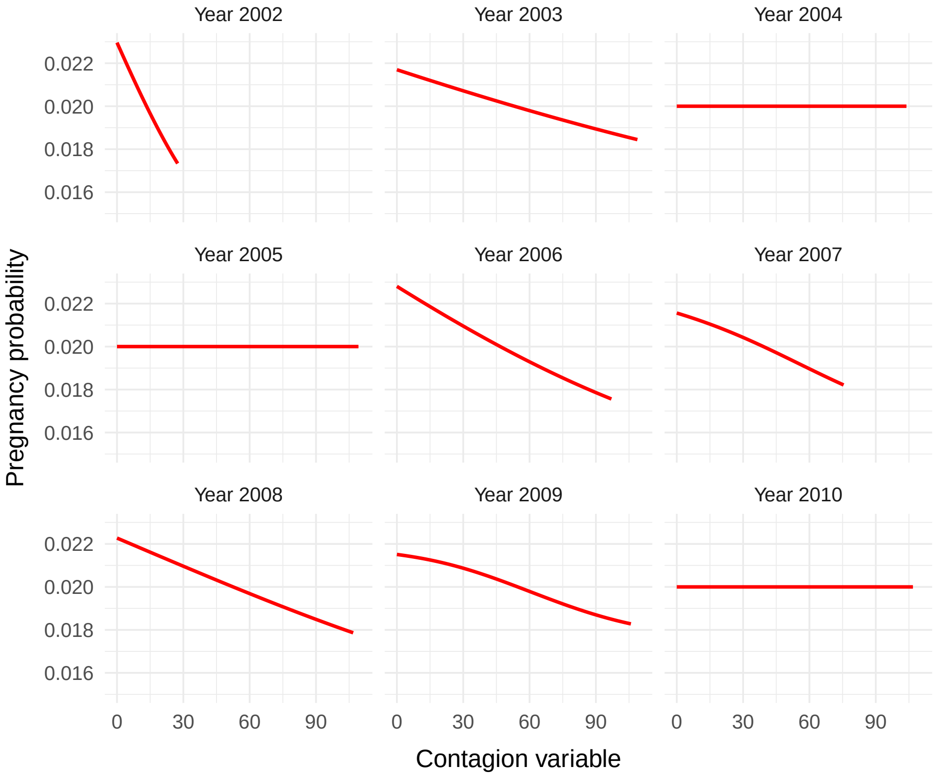


Figure 4.9: Fitted marginal pregnancy probabilities over contagion variable value for each year from the local contagion analysis (with the contagion variable ranging up to the maximum value within the data). The variation of pregnancy probability over the contagion variable is decreased compared to the case of global contagion (Figure 4.8).

Year	No contagion S_B	Contagion S_B	Contagion lengthscale
2002	0.0771	0.0771	8.897
2003	0.0728	0.0729	15.37
2004	0.0732	0.0732	9365
2005	0.0877	0.0877	3252
2006	0.0750	0.0757	7.094
2007	0.0773	0.0771	3.205
2008	0.0864	0.0864	8.722
2009	0.0869	0.0864	1.665
2010	0.0735	0.0735	6512
Average	0.0789	0.0789	8.897

Table 4.5: Comparison between Model 1 (no contagion) and Model 2 (contagion) from the local contagion analysis for each year shown by the Briers scores (S_B) from predicting the pregnancies for the next timestep. Contagion lengthscale values from the Gaussian process classification for Model 2 are also included. The contagion variable performs better here than in global contagion (Table 4.4), though the values of the lengthscales are more commonly large enough to imply the contagion variable is irrelevant. Note the average lengthscale is the median, whilst for Briers scores they are means.

The results (Figure 4.8 and table 4.4) for global contagion were unfortunately a little ambiguous. Model 1 performed better for most years, but the difference between it and Model 2 was minimal. The contagion variable lengthscale in Model 2 was found to leap to extremes presumably based on characteristics of the data for each year. However, for the majority of years the lengthscales were larger than 1 (the variation in the standardised data), implying that it was rare for the contagion variable to have much contribution to the changes in pregnancy probability. In the years when it did the impact was counter-intuitively negative. Overall the model does not provide evidence for pregnancy risk behaviour contagion amongst adolescents in Agincourt, but the ambiguity of the results means we cannot conclude we have convincing evidence that contagion definitely does not occur.

Restricting the contagion variable to only local contagion does little to reduce this ambiguity (Figure 4.9 and table 4.5). We find a marginal increase in predictive performance, and actually see Model 2 being preferred for some years. However, we still observe counter-intuitive negative effects from the contagion variable on the pregnancy probabilities, and the lengthscales are more commonly greater than 1. As such we still see no evidence for the existence of contagion, though if contagion is occurring it seems more likely it is occurring at the village level rather than across the study area. This makes sense due to the localised nature of the lives of the individuals living there.

Overall these results give no evidence of pregnancy risk behaviour contagion occurring amongst adolescents in the Agincourt study area. The ambiguity produced by the method employed is unfortunate, but we can still see the benefits in using a more complex method such as Gaussian process classification over the more simplified methods commonly applied in public health. The flexibility of the method is what allowed us to attempt an analysis like this in the first place. Relying on a simple method for an obviously complex problem at best results in weak conclusions, as shown in our previous analyses, but here would simply have prevented the analysis from even being attempted.

4.4.5 Possible future directions

Though there was great merit in showing how using Gaussian process classification provides increased flexibility to address more difficult problems, the ambiguity in the results found here provide incentive to try alternative methods for examining social contagion in datasets devoid of social network data.

The biggest limitation of the method presented here was in defining the contagion variable in terms of the other covariates. One part of future work in this area must include some more in-depth assessment of what covariates should be included. In this particular case of pregnancy risk contagion in Agincourt it may be more sensible to define the contagion variable in terms of variables associated with the likelihood of individuals being exposed to each others behaviours rather than on the other pregnancy risk covariates. However, such variables would be difficult to determine and in most cases are unlikely to be included in the provided dataset.

Another possible limitation might be in the choice of a semi-parametric method such as Gaussian processes which may have provided too much flexibility making it hard to distinguish between the contagion and no contagion models (particularly if contagion is not actually occurring). A non-linear parametric model similar to those presented in Chapter 3 would include strict enough assumptions making it much easier to distinguish the contagion from the no contagion model when comparing model performance. To get around the issue of having no social network data we could use a similarity graph (Aggarwal, 2015), i.e. a network where the nodes represent the observations in our data and weighted links are formed between them based on the similarity of certain covariate values between them, which in this case would be either pregnancy risk factors or behaviour exposure risk factors. The disadvantage of using a parametric model would be in having to design a justified functional form for the model, an issue that Gaussian processes allowed us to bypass.

CHAPTER 5

Bayesian belief network modelling of household food security

5.1 Food security in rural South Africa

The Food and Agriculture Organization of the United Nations defined food security as “access of all people at all times to sufficient, nutritionally adequate, and safe food, without undue risk of losing such access” (FAO, 2006). The lack of such security has been a concern of governments over the world for many years now. Large scale surveys have been performed in many developed countries including the UK (Holmes, 2008; Pilgrim et al., 2011), USA (Anater et al., 2014), and Canada (Tarasuk et al., 2014) in order to assess the severity and impact of food insecurity, and in many cases to inform the development of new policies.

It is also a current concern of the citizens living within the Agincourt Health and Socio-demographic Surveillance System (HDSS) study area. It is this concern that lead to a meeting between the Agincourt Community Advisory Group (CAG) and the author on 29th July 2015 to discuss food security in the area. The CAG are individuals who live in the Agincourt study area and act as liaisons from the community to the research unit. The perception of individuals within the CAG was of food insecurity being a persistent and widespread issue in the area, with high financial contraints, intermittent water access, poor land quality, and the unfashionable status of subsistence farming amongst young people making it difficult to attain enough food each month to live a healthy and active lifestyle.

Results from national surveys showed that the number of food insecure individuals in South Africa decreased between 1999 and 2008 (Labadarios et al., 2011), and the United States Department of Agriculture International Food Security Assessment 2014-2024 forecasts that the number of food insecure individuals throughout sub-Saharan Africa will fall by 13% in that time (Rosen et al., 2016). However, previous to this it was found that around 42% of individuals in South Africa lived below the food poverty line (Rose and Charlton, 2002), and even during the time between 1999 and 2008 over a third of children still had low dietary diversity (Labadarios et al., 2011). Pertinent to the individuals living in the Agincourt study area the food insecurity rate in rural areas of South Africa has been shown to be twice that of urban areas (Rose and Charlton, 2002). Therefore it is understandable that food insecurity should be a concern of the CAG, and there is value in pursuing research that addresses concerns of the individuals we are performing research on.

Many studies have already been performed on household food security in the Agincourt study area, including on the impacts of factors such as adult mortality (Twine and Hunter, 2011; Hunter et al., 2007) and the food retail sector (Pereira et al., 2014), the quality of different food security indicators (Kirkland et al., 2013), and how household food security in the area varied over the 2008 financial crisis (Nawrotzki et al., 2014). All these studies share a reasonably common set of methodologies including linear models, summary statistics, and statistical tests, similar to many of the studies we referenced in the previous chapters. However, in this chapter we do not argue that there is something fundamentally wrong with these methods that our more complex methods fixes, such as how the lack of assumptions opened up causal linear models to confounding in Chapter 3 and how the strict linear assumptions of linear regression models introduced unfounded a priori assumptions on the relationships between our variables in Section 4.3. Instead here we aim to show how complex methods can improve upon the limitations of traditional methods, similar to how the flexibility of Gaussian processes allowed us to attempt analyses in Section 4.4 that would be impossible with traditional methods. The methods used so far in the analysis of household food security in the Agincourt study area are limited in what they are able to tell us about what is arguably a complex system. Statistical tests are undermined by the many issues discussed in Section 1.2, summary statistics have limited ability to differentiate between different sets of data, and most linear models are only able to provide us with simplistic measures of direct effects from a set of independent to one dependent variable.

When analysing the role of sugar in UK food security, Barons et al. (2014) applied the method of Bayesian belief networks, the basics of which we described in Section 2.3. Such a network gives a graphical representation of the probabilistic dependencies and independencies in a system. In doing so it presents more of a many-to-many view of the system where each variable is represented as a random variable with the potential to have a direct relationship with any other variable. Therefore though we are interested in food security, we can technically make observations of any variables in the model to make inferences on any of the other variables. This is more representative of complexity within a system than the many-to-one view of many other models (including both linear models and those we have used in previous chapters) where we always have one set dependent random variable being inferred from a group of set observed independent variables. It also allows us to reason probabilistically about the system, and by applying a probabilistic structure to the system it gives us the ability to ask questions of how each variable probabilistically impacts on each other variable. As such it gives us a flexible tool that can be used in a variety of analyses, overcoming many of the limitations of the models inferred in previous studies. Bayesian belief networks have already seen successful use in many areas, including for example semantic search (Koumenides and Shadbolt, 2012), information retrieval (de Campos et al., 2004), analysis of gene expressions (Friedman et al., 2000), medical diagnosis (Jiang and Cooper, 2010), and filtering, smoothing, and prediction (Weiss et al., 2010). Many machine learning techniques, including naive Bayes classifiers and hidden Markov models, are types of Bayesian belief networks (Murphy, 2012). Due to their structure providing directed relationships amongst variables, and to the inbuilt modelling of conditional independence, Bayesian belief networks are also well suited to act as causal models of systems, which then enables us to make interventional, counterfactual, and other more complicated queries of our model (Pearl, 2013).

Though Barons et al. (2014) have already built a Bayesian network model of food security, they have only done so for a developed country at an infrastructural level. An attempt to use this method for analysing household food security in sub-Saharan Africa has yet to be made. In this chapter we build a Bayesian belief network model with the specific aim of providing a causal representation of the interrelations between various characteristics of households in the Agincourt study area within the context of food security. Though we aim for the resulting model to provide causal insights into the system, we note later that validating such a causal model is difficult. However, such a model forms at least a starting point in achieving this aim (perhaps to be later improved by interventional trials), and can act as a basis for the development of future Bayesian network models of household food security in other similar populations. In the rest of this chapter we give details of the construction of our specific Bayesian network from choosing the variables, through learning the structure and parameters, to attempting quantitative validation of the model. Finally, we give some elementary applications of the networks in interventional inferences.

5.2 Building the Agincourt food security belief network

5.2.1 Variable selection and data

As the initial justification for developing the Agincourt food security Bayesian belief network came from members of the community (see Section 5.1), it is of interest to see whether using expert knowledge from members of the community can aid in learning the structure of the network. This not only introduces a strong element of community involvement in the work, if the expert knowledge forms an accurate representation of the system it can also improve the possibility of discovering an actual causal structure which is difficult to achieve when learning from data (Smith, 2010). This is based on the assumption that there are no latent variables not covered in the variables considered, and that the knowledge of the experts is expansive enough.

We perform variable selection with the aim in mind of learning a causal structure from expert elicitation. When performing the expert elicitation, we follow a protocol stated in Smith (2010) designed specifically for eliciting a causal structure from the knowledge of our experts. The actual protocol is detailed in Section 5.2.2.1. Here we focus on the first step, which involves a literature search to discover the variables.

The literature search was performed over literature based on the Agincourt HDSS dataset (for further details of which, see Section 4.2) for household characteristics related to food security. The search involved finding variables in steps. The first step involved searching the literature for household characteristics that form possible causes of change in household food security, which is our utility variable. The second step then involved searching the literature for household characteristics that form possible causes of changes in the characteristics found in step one. The third step looked for characteristics that form possible causes of those from step two, and so on and so forth. This is repeated for as many steps as we are practically able to complete, though we were also limited by two other factors. First, we could only select variables that are practically measurable using data in the Agincourt HDSS dataset, or any other relevant datasets available to us. Second, the number of variables were limited to those of greatest importance and relevance to simulation of food security in order to avoid placing too much of a burden on the experts when eliciting the structure of the network.

Household food security in the Agincourt study area was found in the literature to be influenced by whether households participate in subsistence or commercial farming, forage for food from the local environment, receive aid from the community, or receive government child welfare grants; as well as on their socio-economic status, how well educated the household members are, what the gender of the household head is, how many individuals in the household are too young or old to work, and how many working age adults live in the household (defined as aged 15-59 as individuals make their greatest economic contribution to the household from the age of 15 (Twine and Hunter, 2011), and the age at which individuals in South Africa can claim an older persons grant is 60 (SASSA, 2017a)) (Hunter et al., 2007; Kirkland et al., 2013; Nawrotzki et al., 2014; Pereira et al., 2014; Twine and Hunter, 2011; Kaschula, 2011; Lemke et al., 2003; Sedibe et al., 2014). Whether a household forages for food from the environment was found to be influenced by the numbers of individuals in the household too young or old to work, as well as the number at working age (Dovie et al., 2007), and is obviously also affected by the level of local vegetation in the local area. Whether a household is likely to receive communal aid (e.g. food or money) was found to be influenced by their socio-economic status, the gender of their household head, and whether the household contains refugees (i.e. individuals not defined in the dataset as South African nationals, usually due to being Mozambican refugees) (Kaschula, 2011; Pronyk et al., 2008). Households selling crops and livestock is obviously influenced by whether they grow or rear them in the first place. Whether households claim child support grants was found to be influenced by how well employed the household members are and their socio-economic status (which is unsurprising as the grant is means tested (SASSA, 2017b)), as well as their refugee status and obviously by whether the household has any children (Pereira et al., 2014; Kaschula, 2011; Twine et al., 2007). Whether the household grows crops and livestock was found to be influenced by their socio-economic status, how good their access to water is, and whether they have enough household members to perform the farming (Hunter et al., 2007; Pereira et al., 2014). Household socio-economic status was found to be influenced by the household levels of employment and education, as well as their refugee status and the gender of their household head (Pereira et al., 2014).

Variable	Definition	States	Data source
Food insecure	Whether the household has or has not had enough to eat over the previous year.	0 - had enough to eat. 1 - did not have enough to eat.	NotEatYear variable in the Agincourt Food Security module.

Table 5.1: Agincourt food security belief network variables on level 1 of the causal ordering. The data source specifies how the variable is calculated from the data, and the states specify how the discrete variable states relate to the values resulting from the calculations described in the data source.

Variable	Definition	States	Data source
Use of wild foods	Whether the household uses wild foods.	0 - do not use wild foods. 1 - gather wild foods OR wild herbs. 2 - gather wild foods AND wild herbs.	Sum of the SupGather and SupGatherHerbs variables from the Agincourt Food Security module.
Receipt of communal aid	Whether the household gets financial aid from friends, neighbours, or family.	0 - receive no aid. 1 - receive aid.	SupDonate variable from the Agincourt Food Security module.

Table 5.2: Agincourt food security belief network variables on level 2 of the causal ordering. The data source specifies how the variable is calculated from the data, and the states specify how the discrete variable states relate to the values resulting from the calculations described in the data source.

Variable	Definition	States	Data source
Local vegetation level	Amount of vegetation that grows in the local area of the household.	1 - up to 0.161. 2 - 0.161-0.302. 3 - 0.302-0.443. 4 - 0.443-0.584. 5 - 0.584-0.725. 6 - greater than 0.725.	Sum of the average NDVI values from the MODIS Terra satellite data in a 2000km by 2000km area around the household divided by the number of households in that area.
Selling of crops and livestock	Whether the household sells crops or livestock to supplement their income.	0 - sell no crops or livestock. 1 - sell crops or livestock.	Sum of the SupSellGoods and SupSellStock variables from the Agincourt Food Security module.
Child grant status	Number of child grants the household receives.	State is equal to number of grants received that year.	Count of grants received using the ReceivedYear variable from the Agincourt Child Grant module.

Table 5.3: Agincourt food security belief network variables on level 3 of the causal ordering. The data source specifies how the variable is calculated from the data, and the states specify how the discrete variable states relate to the values resulting from the calculations described in the data source. Value intervals are inclusive of the upper values.

Variable	Definition	States	Data source
Number of dependents	Number of household members aged less than 15 or greater than 59.	1 - up to 9.5. 2 - 9.5-19.0. 3 - 19.0-28.5. 4 - 28.5-38.0. 5 - 38.0-47.5. 6 - greater than 47.5.	Count of household members (membership counted at midyear) with the appropriate age (calculated at midyear) in the Agincourt dataset.
Use of crops and livestock	Whether the household grows its own crops or keeps its own livestock.	0-7+.	Sum of the MaizePlot, MaizeField, OtherCropsPlot, and OtherCropsField variables from the Agincourt Food Security module with the Cattle, Goats, Poultry, and Pigs variables from the Agincourt Asset Status module.

Table 5.4: Agincourt food security belief network variables on level 4 of the causal ordering. The data source specifies how the variable is calculated from the data, and the states specify how the discrete variable states relate to the values resulting from the calculations described in the data source. Value intervals are inclusive of the upper values.

Variable	Definition	States	Data source
Water access	Availability, stability, and quality of the water supply to the household.	1 - up to 10.5. 2 - 10.5-16.0. 3 - 16.0-21.5. 4 - 21.5-27.0. 5 - 27.0-32.5. 6 - 32.5-38.0. 7 - 38.0-43.5. 8 - greater than 43.5.	Product of the WaterDistMetre, WaterAvail, and WaterSup variables from the Agincourt Asset Status module.

Table 5.5: Agincourt food security belief network variables on level 5 of the causal ordering. The data source specifies how the variable is calculated from the data, and the states specify how the discrete variable states relate to the values resulting from the calculations described in the data source. Value intervals are inclusive of the upper values.

Variable	Definition	States	Data source
Socio-economic status	Socio-economic status of the household.	1 - up to 1.262. 2 - 1.262-1.648. 3 - 1.648-2.033. 4 - 2.033-2.419. 5 - 2.419-2.804. 6 - 2.804-3.190. 7 - 3.190-3.575. 8 - greater than 3.575.	SESAbsolute variable from the Agincourt SES index module.

Table 5.6: Agincourt food security belief network variables on level 6 of the causal ordering. The data source specifies how the variable is calculated from the data, and the states specify how the discrete variable states relate to the values resulting from the calculations described in the data source. Value intervals are inclusive of the upper values.

Variable	Definition	States	Data source
Employment level	Number of household members with current employment.	0-6+ employed household members.	Count of household members (membership counted at midyear) who replied yes to the CurrentlyWorking variable in the Agincourt Labour Status module.
Household head gender	Gender of the head of the household.	0 - male. 1 - female.	Household head gender as specified in the Agincourt dataset.
Refugee status	Whether any household members are refugees or non-South African nationals.	0 - no refugee members. 1 - one or more refugee members.	Refugee variable from the Agincourt dataset.

Table 5.7: Level Agincourt food security belief network variables on level 7 of the causal ordering. The data source specifies how the variable is calculated from the data, and the states specify how the discrete variable states relate to the values resulting from the calculations described in the data source.

Variable	Definition	States	Data source
Number of working age adults	Number of household members aged 15 to 59.	1 - up to 13.29. 2 - 13.29-26.57. 3 - 26.57-39.86. 4 - 39.86-53.14. 5 - 53.14-66.43. 6 - 66.43-79.71. 7 - greater than 79.71.	Count of household members (membership counted at midyear) with appropriate age (calculated at midyear) from the Agincourt dataset.

Table 5.8: Agincourt food security belief network variables on level 8 of the causal ordering. The data source specifies how the variable is calculated from the data, and the states specify how the discrete variable states relate to the values resulting from the calculations described in the data source. Value intervals are inclusive of the upper values.

Variable	Definition	States	Data source
Education level	Average years of completed education by the household members.	1 - up to 3. 2 - 3-6. 3 - 6-9. 4 - 9-12. 5 - greater than 12.	Average of the EducationInYears variable from the Agincourt Education module for household members (membership counted at midyear) for that year.

Table 5.9: Agincourt food security belief network variables on level 9 of the causal ordering. The data source specifies how the variable is calculated from the data, and the states specify how the discrete variable states relate to the values resulting from the calculations described in the data source. Value intervals are inclusive of the upper values.

These relationships gave us a causal ordering of the different household variables, as shown by the breaking up into levels of the variables in Tables 5.1 to 5.9. The variables of each level can only be caused by a variable in a higher numbered level, and can only cause those in a lower numbered level, according to this ordering. For example, water access in level 5 (Table 5.5) is possibly a cause of variables such as use of crops and livestock, use of wild foods, and food security itself which are all in lower numbered levels, but is disallowed from causing variables such as socio-economic status, refugee status, or education level which are in higher numbered levels. This ordering is important for attempting to ensure a causal structure when eliciting the network structure from experts.

Each variable was calculated as specified in Tables 5.1 to 5.9. For most, this involved a suitable combination of appropriate variables in the Agincourt HDSS dataset, calculated at or aggregated to the household level. There are two variables worth providing further details for.

Food security was calculated from a single variable in the Agincourt dataset specifying whether the household had enough food to eat over the past year. This admittedly lacks the nuance of most definitions of food security commonly used in past studies. However, most other variables related to food security in the Agincourt HDSS dataset feature large amounts of missing data, whilst this variable does not. It also provides a strong relation to actual food security despite being simplistic. When considering the results of applying the completed Bayesian network, we must simply remember that food security in this case is defined as having had enough food to eat over the past year.

Following Nawrotzki et al. (2014), we calculate local vegetation level from the Normalised Difference Vegetation Index (NDVI) (Tucker, 1979). NDVI data was obtained from the MODIS/Terra sensor satellite images, which contains NDVI values for 250m by 250m areas averaged over 16 day periods (Didan, 2015). As per Nawrotzki et al. (2014), local vegetation level is calculated for a particular household as the NDVI over the 2000km by 2000km region around the household (not including land within villages so as to avoid including privately owned land) averaged over the households within that region. Though Nawrotzki et al. (2014) averaged these values over the preceding three years of the year of analysis, we only average them over the year of analysis to avoid removing detail of differences between each year.

Variables that were calculated from the Agincourt food security module were done so for the years 2007, 2010, and 2013, i.e. the years in which the module was performed during the census. For the other variables, their value from the nearest previous year when the data to calculate them was collected was used for each household at each of the three food security module years. Our sample was formed by households that provided complete enough data to calculate values for each of the variables for at least one of the food security module years. For each household, the latest entry from either 2007, 2010, or 2013 was taken, as it was found that allowing households to have multiple entries for different years reduced the performances of the models seen in the model validations and comparisons. The sample size for our dataset was 11739 households.

Though it would be possible to construct a dynamic Bayesian network with nodes for the different variables at each of the different time points (Koller and Friedman, 2009), it was decided instead to produce only a simple static network due to each year only providing a small amount of data, and only three years worth of data being available with a significant gap of three years between each year. Eliciting dynamic as well as static relationships from the experts would also create an even greater burden upon them. Construction of a dynamic network, which would be more suitable for the causal modelling we aim for, should be an immediate concern of future work where greater resources and access to experts are available.

Several of the variables were discretised, as detailed in the states shown in Tables 5.1 to 5.9, in order to reduce the number of possible states and simplify the calculations involved in finding the parameters of the model and performing inferences. This was done in such a way in order to still coarsely approximate the distribution of the variable, either by combining neighbouring states that occurred rarely, or by binning the data for that variable to an appropriate number of histogram bins.

5.2.2 Structure learning

5.2.2.1 Expert elicitation

As mentioned in Section 5.2.1, one method of learning a causal structure is via expert elicitation. Here we rely on the knowledge of a set of experts who can inform us of how our network should be structured based on how changes in each variable affect their beliefs about the states of other variables. We performed an expert elicitation to find a possible causal structure for the Agincourt food security belief network, relying on the knowledge of members of the Community Advisory Group

(CAG).

To perform this elicitation we relied on the protocol described in chapter 7 of Smith (2010). The first step, in which we established a set of variables and a causal ordering for them, was described in Section 5.2.1. The next step involved the actual elicitation. The design of the elicitation was piloted on various academic colleagues (including some from a non-mathematical background) to attempt to make the process more understandable for the experts. The elicitation was performed by Rhian Twine of the Agincourt LINC office following our design and instructions in the MRC/Wits Agincourt offices on 28th October 2016. The experts were guided through a list of the variables, in the order defined by the causal ordering starting with the highest level variables (which our ordering specifies have no possible causes within the other variables, but are possible causes for all the other variables) and finishing with the lowest level variable food security (which is specified to cause none of the other variables, and can be caused by all other variables). For each variable, we attempted to establish based on the local knowledge of our experts which of the previous variables in the list would have an effect on the beliefs of the experts on what state the variable in question would take for a generic household.

The difficulty in this is in trying to capture conditional independencies, i.e. that we only want the experts to tell us of direct effects between the variables rather than indirect effects where variable A affects our belief of variable B because it actually impacts on C which impacts on B. As such, we carefully designed our questions to attempt to capture this by asking them to first consider the hypothetical situation where they already know the states for a household of all the variables previous in the list except one, and then whether learning the state of that one extra variable would provide any extra impact on their belief of the variable of interest. Say the experts have gone through variables A, B, C, and D, and are now on variable E. Formally, they were shown a list of A, B, C, and D and were asked the following.

> Consider each of the characteristics in the list below separately and answer the following question. If you already knew a households status for the other characteristics listed, would knowing the households status for this characteristic have any additional impact on your estimate of their status for E (given you are unable to ask them)?

Often this was reframed in more informal language to make it easier for the experts to understand.

We know B, C, and D for the household (sometimes picking example states), and want to know E. Would knowing A help you know E if you already know the others?

In order to suitably prepare the experts for the elicitation, they were first told the purpose of the work and what outcomes are hoped for, as well as a brief laymans explanation of Bayesian networks. They were guided through a set of example questions on a different smaller system taken from the Agincourt study area in order to help them understand how to answer the questions, and to get them used to answering them before having to answer the ones we cared about. They were also reassured that none of their answers would be incorrect in order to avoid any biases such as adjusting their answers to try and get to the 'right one'. The elicitation was performed with the experts as a group rather than individually to further reduce the burden on them.

The resulting network from the expert elicitation (Figure 5.1) already shows some interesting discrepancies from findings in the literature. For instance, the experts determined that household head gender has no impact on any of the other variables. Refugee status is also shown to have minimal direct impact. Also water access has no impact on the growing of crops, supposedly due to individuals relying on rainwater rather than the water supply to water their crops and gardens.

5.2.2.2 Learning from data

Whilst expert elicitation has the advantage of being able to learn the structure in such a way as to maximise the chances of learning an actual causal structure, it also relies entirely on the knowledge of the experts which could be narrowed by their life experiences. Unlike the experts used by Barons et al. (2014) who came from a variety of backgrounds and who had studied the problem, ours were experts in living in the locality. An alternative to expert elicitation is learning the structure algorithmically from the data. There are many different algorithms that can be used to learn the structure of a network, broadly broken up in to two groups known as constraint-based and score-based structure learning where constraint-based algorithms are generally considered more suitable for causal learning due to their aim of discovering local independencies between variables rather than simply maximising some information theoretic measure of the model (Koller and Friedman, 2009).

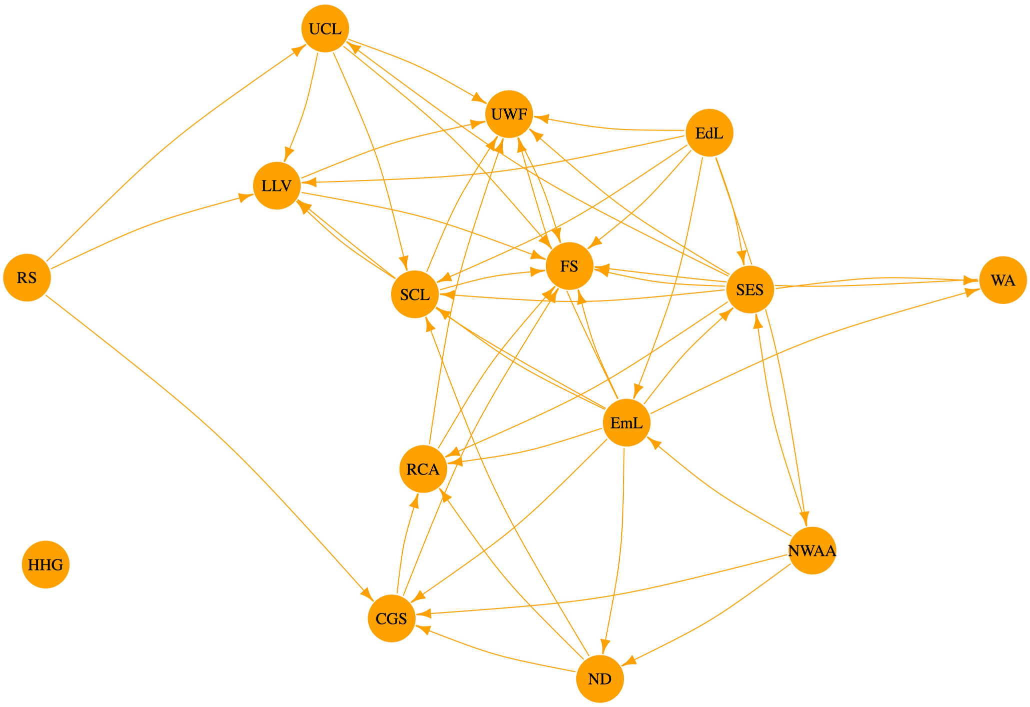


Figure 5.1: Agincourt food security belief network structure learned from the expert elicitation. CGS - child grant status. EdL - education level. EmL - employment level. FS - food security. HHG - household head gender. LLV - level of local vegetation. ND - number of dependents. NWAA - number of working age adults. RCA - receipt of communal aid. RS - refugee status. SCL - selling of crops and livestock. SES - socio-economic status. UCL - use of crops and livestock. UWF - use of wild foods. WA - water access.

In order to learn an alternative possible causal network for our system we used a constraint-based algorithm known as Max-Min Parents and Children (MMPC) (Tsamardinos et al., 2003, 2006). For each node X, the MMPC algorithm attempts to discover the set of parents and children of X. It first builds a candidate set of parents and children (CPC). This is done by performing conditional independence tests between X and each node not yet in the CPC set given all possible subsets of the nodes already in the CPC set. Any node found by the test to be independent of X given any subset of the CPC set are discluded for the rest of the iterations. From the remaining, the node with the highest minimal association with the node of interest (measured by the independence test p-values, where minimal refers to the smallest association between X and the node found by conditioning on every possible subset of the CPC set) is then added to the CPC set. This is repeated, increasing the CPC set by one each time (and also possibly increasing the set of nodes that have been found to be independent and can therefore no longer be considered for the CPC set), until the CPC set can no longer be increased. To check for false positives of the test, a backwards run is then performed through the CPC set where nodes are removed if they can be rendered d-separated from X given some subset of the other nodes in the CPC set.

The MMPC algorithm can be performed for each node to find all the local structures and construct a skeleton of the network (i.e. with no directions). This was performed for our Agincourt household food security dataset using the asymptotic normal Jonckheere-Terpstra test for conditional independence (which tests the null hypothesis of the distribution of e.g. A given B and C being equal for all values of B against the alternative hypothesis of the distribution having a stochastic ordering over B) to take into account the ordinal nature of the variables (Jonckheere, 1954). Links that would go against the causal ordering provided by the levels described in Section 5.2.1 were disallowed from the start. The order of the variables used in the expert elicitation was used to gain directions in the skeleton network achieved by the algorithm.

The resulting network (Figure 5.2) shows some substantial differences to the expert elicited network (Figure 5.1), which is unsurprising given the very different sets of information each network was built upon. Most different is the much greater level of sparseness in the data-learned network. Household head gender is also linked in to the data-learned network, unlike the expert elicited one, but selling of crops and livestock is now disconnected. Child grant status has no directed path to food security either, meaning that neither of these nodes would be effective means of

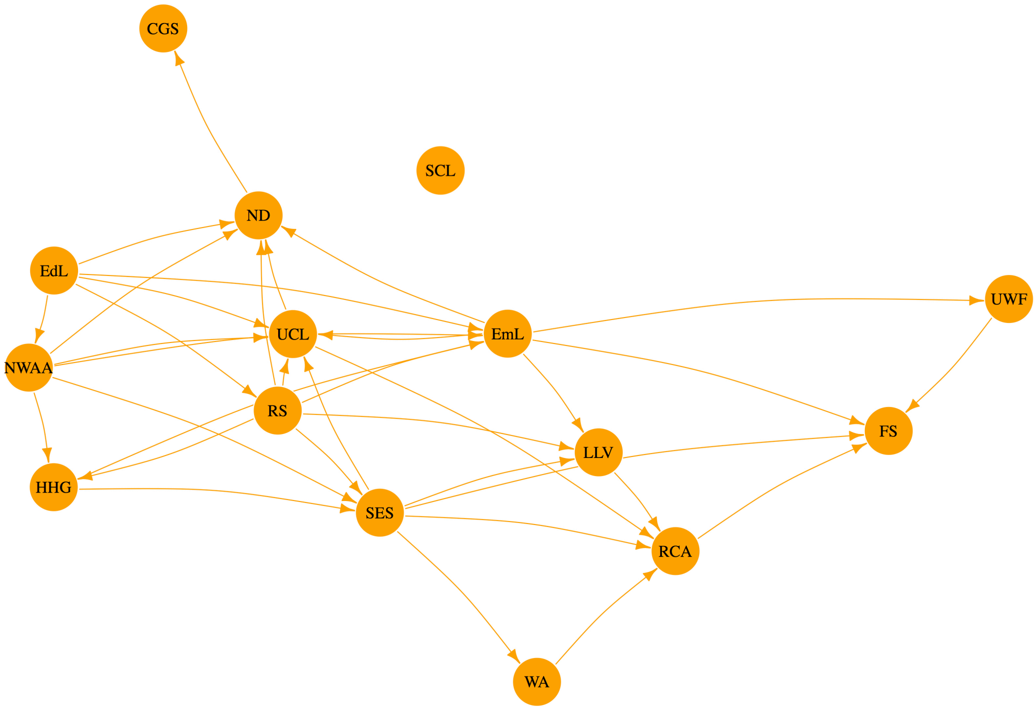


Figure 5.2: Agincourt food security belief network structure learned from the data. CGS - child grant status. EdL - education level. EmL - employment level. FS - food security. HHG - household head gender. LLV - level of local vegetation. ND - number of dependents. NWAA - number of working age adults. RCA - receipt of communal aid. RS - refugee status. SCL - selling of crops and livestock. SES - socio-economic status. UCL - use of crops and livestock. UWF - use of wild foods. WA - water access.

intervention according to this network.

5.2.2.3 Combining expert and data learning

One final means of constructing a causal network is to combine both the expert results and data-based learning by using the expert network as a prior for algorithms to use the data to build upon. Whilst this is difficult to achieve using constraint-based algorithms, it is attainable quite easily using score-based methods where the expert network can be used as a starting point for the search. For this we relied on a standard algorithm that has shown good performance in the past known as Tabu search (Bouckaert, 1995; Tsamardinos et al., 2006).

Tabu search, in the case of Bayesian networks, is a score-based algorithm where a greedy search over the possible network structures is performed (by either adding or deleting edges as necessary) in order to maximise some score assigned to each structure. The score we used was the Akaike information criterion (AIC) (Akaike, 1974)

$$\mathrm{AIC} = 2\nu - 2\log L\left(\left\{\hat{\theta}_i^{jk}, n_i^{jk}\right\}\right) \tag{5.1}$$

where $\hat{\theta}_i^{jk}$ are the fitted network parameters (defined as the local conditional probabilities that variable i is in state j given its parents are in states k as stated in (2.18)), ν are the number of such parameters that exist within the particular network structure, n_i^{jk} are the number of observations in the data with variable i in state j given its parents (as defined by this structure) are in states k, and the likelihood is

$$L\left(\left\{\theta_i^{jk}, n_i^{jk}\right\}\right) = \prod_{i=1}^{N} \prod_{j \in \mathbb{S}_{X_i}} \prod_{k \in \mathbb{S}_{\mathbf{pa}(X_i)}} \left(\theta_i^{jk}\right)^{n_i^{jk}} \tag{5.2}$$

where $\mathbb{S}_X$ is the set of states of variable X. The AIC avoids penalising non-sparse networks in ways that other information criteria such as the Bayesian information criterion would.

Tabu differs from a regular hill climbing algorithm in that once it has found a local optimum it attempts to check for other possible better local optima nearby. This is achieved by allowing the algorithm to take steps that slightly reduce the score for a set number of steps in which it can attempt to find a better solution. Once the set number of steps have been taken the algorithm is allowed to return to the original solution if no better ones have been found.

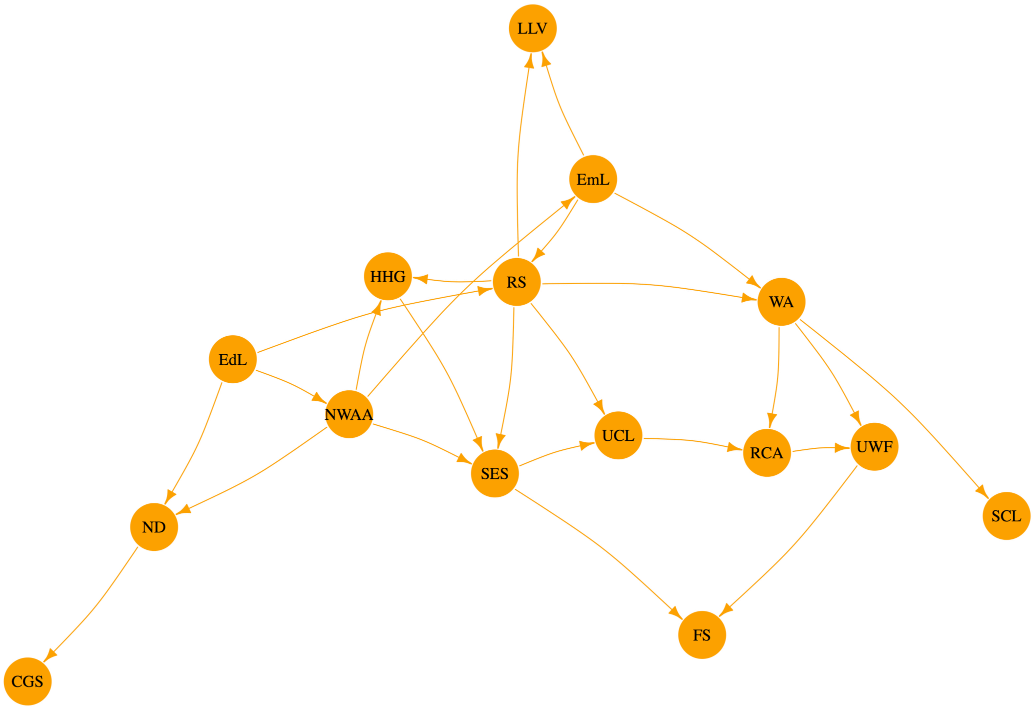


Figure 5.3: Agincourt food security belief network structure learned from the data with the expert network as a prior. CGS - child grant status. EdL - education level. EmL - employment level. FS - food security. HHG - household head gender. LLV - level of local vegetation. ND - number of dependents. NWAA - number of working age adults. RCA - receipt of communal aid. RS - refugee status. SCL - selling of crops and livestock. SES - socio-economic status. UCL - use of crops and livestock. UWF - use of wild foods. WA - water access.

This final resulting network (Figure 5.3) appears sparser than the other two but also has all nodes linked into the network. There are also more leaf nodes in the network with no children such as selling of crops and livestock, level of local vegetation, and child grant status again. As before, these nodes would therefore be definitely ineffective for any interventions against food insecurity according to this network. Despite the efforts to ensure relationships going against causality were avoided, some seemingly counter-intuitive ones still appeared such as the number of working age adults and refugee status having a causal impact on household head gender. However, perhaps these relationships are not as counter-intuitive as they appear. Refugee households are perhaps far more likely to have household heads of one gender, whilst non-refugee households are more likely to have household heads of the other.

5.2.3 Parameter estimation

Though it is possible to elicit the network parameters θ_i^{jk} from experts as well, we instead estimated them from data. This both avoids the many cognitive biases that the experts could be prone to, and avoids placing an additional and immense burden on the experts on top of what we have already asked of them.

The parameters were estimated by their maximum a posteriori (MAP) values found by maximising the product of the likelihood (5.2) with a Dirichlet prior

$$\pi\left(\left\{\theta_i^{jk}, \alpha_i^{jk}\right\}\right) = \frac{1}{B\left(\left\{\alpha_i^{jk}\right\}\right)} \prod_{i=1}^{N} \prod_{j \in \mathbb{S}_{X_i}} \prod_{k \in \mathbb{S}_{\mathbf{pa}(X_i)}} \left(\theta_i^{jk}\right)^{\alpha_i^{jk}-1} \tag{5.3}$$

where $B(\cdot)$ is a Beta function. This results in MAP estimate values of

$$\hat{\theta}_i^{jk} = \frac{n_i^{jk} + \alpha_i^{jk} - 1}{\sum_{j' \in \mathbb{S}_{X_i}} \left(n_i^{j'k} + \alpha_i^{j'k} - 1\right)} \tag{5.4}$$

where the values are being normalised over the $J = |\mathbb{S}_{X_i}|$ possible states of the variable i. We used a prior with $\alpha_i^{jk} = 2 \, \forall \, i, j, k$ resulting in parameter values of

$$\hat{\theta}_i^{jk} = \frac{n_i^{jk} + 1}{\sum_{j' \in \mathbb{S}_{X_i}} n_i^{j'k} + J} \tag{5.5}$$

which is equivalent to add-one smoothing (Murphy, 2012). This adjusts for cases where no occurrences of a particular combination of i, j, and k appear in the data.

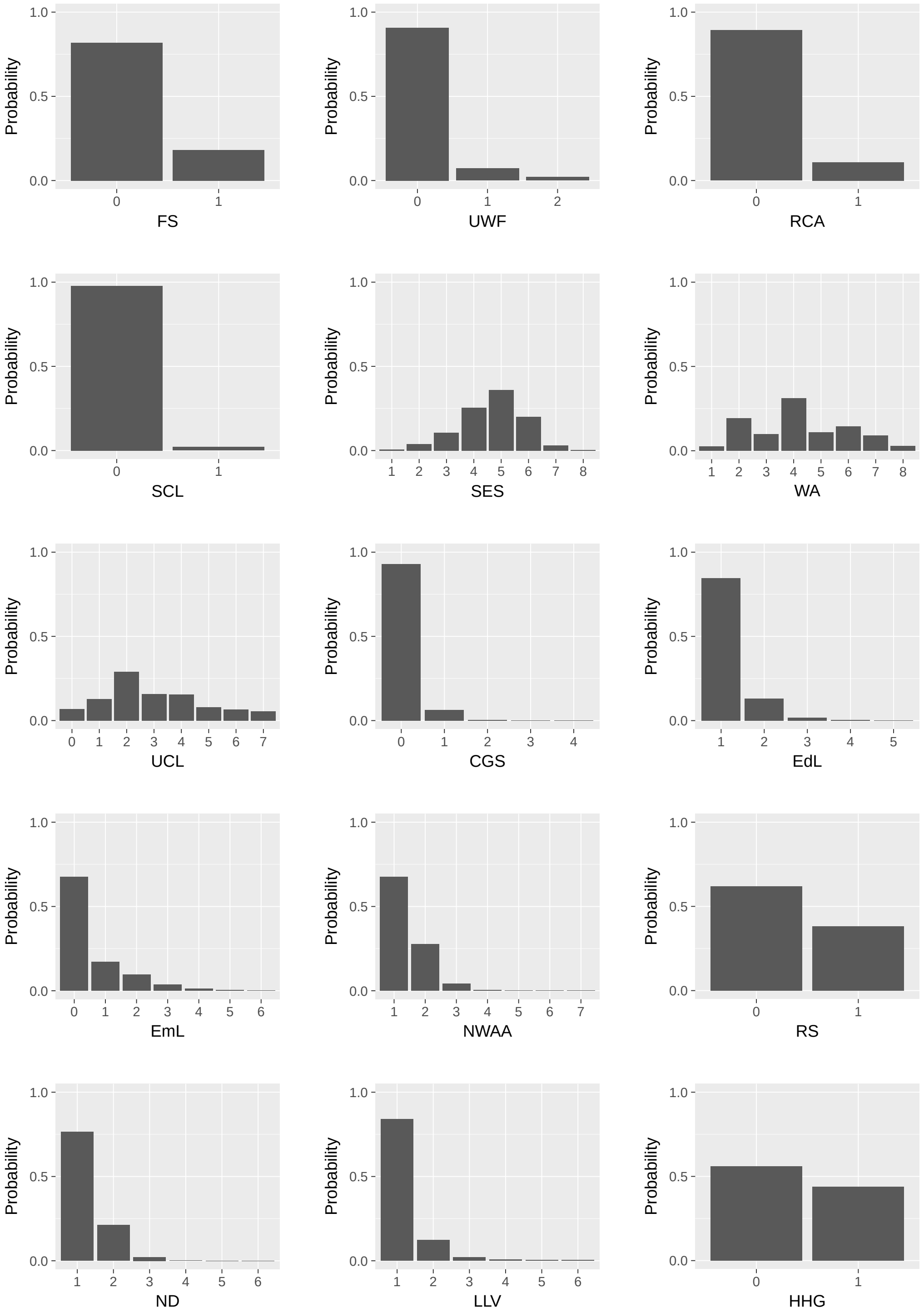


Figure 5.4: Marginal probability distributions for each of the variables in the Agincourt food security belief network with structure learned by expert elicitation. The variable acronyms are described in Figure 5.1.

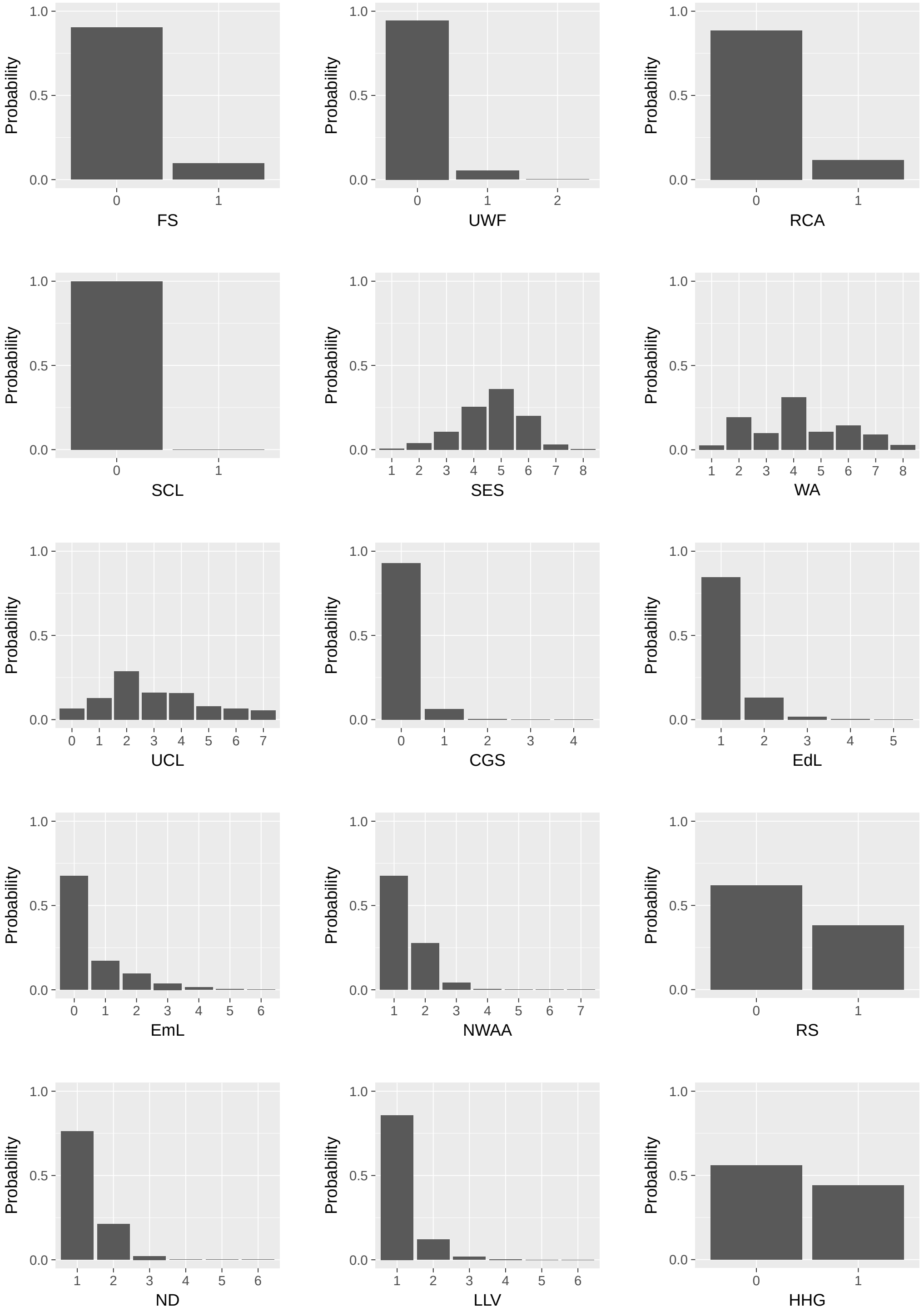


Figure 5.5: Marginal probability distributions for each of the variables in the Agincourt food security belief network with structure learned from data. The variable acronyms are described in Figure 5.2.

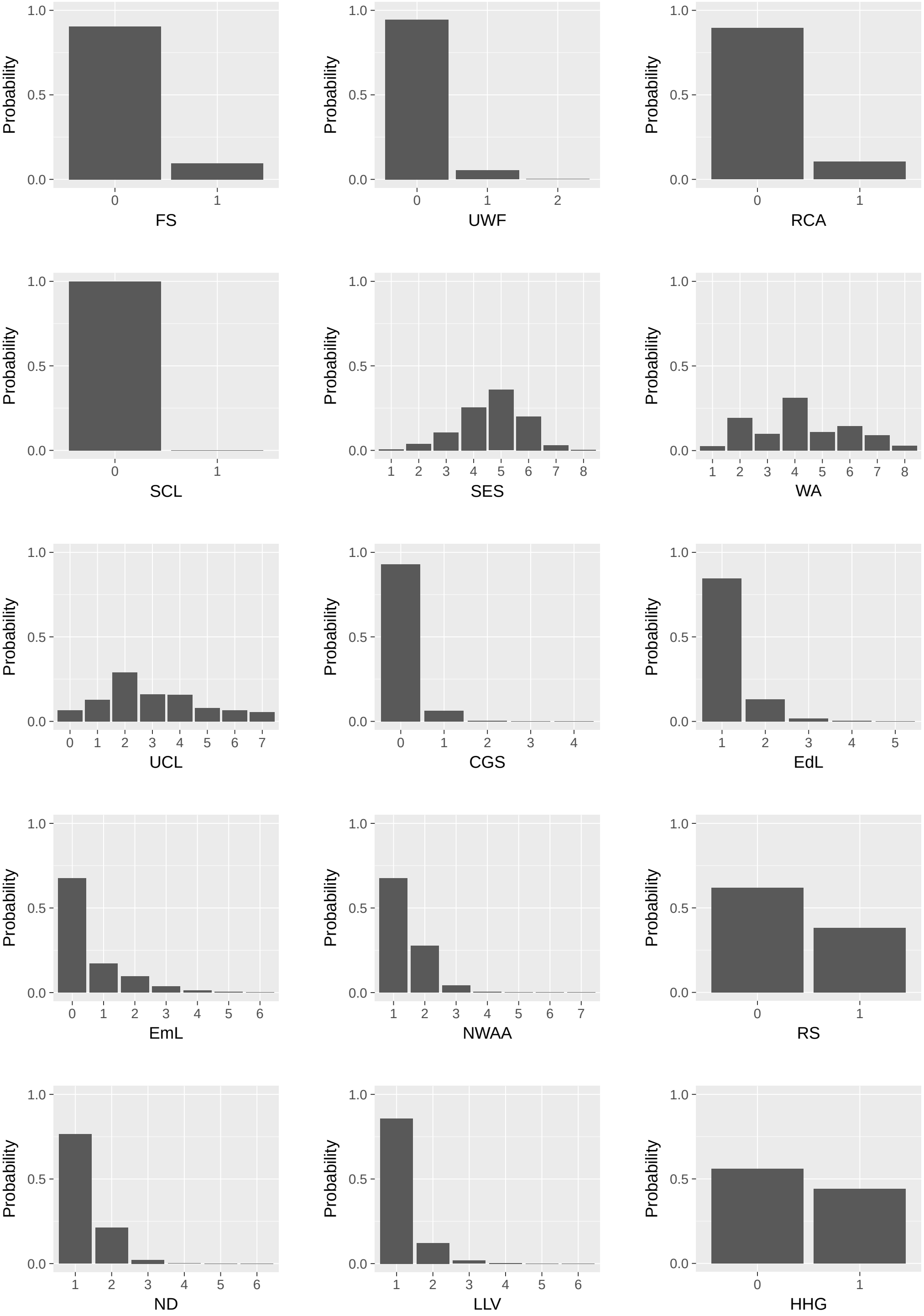


Figure 5.6: Marginal probability distributions for each of the variables in the Agincourt food security belief network with structure learned from data with the expert network as a prior. The variable acronyms are described in Figure 5.3.

Though the amount of parameters is too vast to show all their values here, the marginal probability distributions of the variables for each possible network structure are shown in Figures 5.4 to 5.6 in order to give an indication of the probabilistic values within each network. There is no substantial different between the marginals of each network, which is unsurprising as the marginals should be close to the empirical frequencies observed in the data despite the networks structure.

5.2.4 Model validation

5.2.4.1 D-separation checks

The expert elicitation protocol defined in chapter 7 of Smith (2010) specifies one last step. Independence relationships such as $A \perp B \mid C$ implied by the network structure can be found using d-separation (see Section 2.3.2). Smith (2010) suggests finding such relationships between nodes not directly connected in the network, and checking them against the beliefs of the experts to see if the implications of the network match the knowledge of the experts. If not, then the network should be adjusted accordingly to remove these relationships.

As we were unable to perform more than one extra elicitation, it was decided that this d-separation check would be used as a test of validation against the beliefs of the experts and that no further changes would be made to the structure of the expert network (Figure 5.1). Changing the structure would result in new independence relationships that would have to be rechecked against the experts, which we were unable to do.

As over 8000 different independence relationships are implied by the networks structure, the number was severely reduced in a structured fashion. In order to make it simpler for the experts, only relationships of the kind $A \perp B \mid C$ where A and B only contained one variable each and B came before A in the causal ordering described in Section 5.2.1 were questioned. From these, examples where the fact that A is at all independent of B appears counter-intuitive were selected as these seemed the most pertinent to check. Finally, for each possible pair of A and B in the remaining relationships, the relationship $A \perp B \mid C$ where C contained the smallest set of nodes was selected, again to make it easier for the experts. This resulted in 34 relationships to check against the beliefs of the experts, which were phrased in the following way.

A tells us nothing about B if we already know C. True or false?

This elicitation was performed with a new set of experts, still taken from the CAG. It was again performed by Rhian Twine of the Agincourt LINC office following our design and instructions, located in the MRC/Wits Agincourt offices in March 2017. It must be admitted that there was possible confusion from the experts as to the meaning of the questions. It is possible that they often ignored C when considering whether A should be independent of B. This possibly undermines the results.

That being said, it appears that the initial elicited structure proved quite robust against these checks. Only 11 out of 34 of the relationships were deemed to be false (though 2 were unanswered). Typically this was due to the new set of experts thinking that household head gender does impact on certain variables, Mozambican refugees are now indistinguishable from South African nationals, and that both age and education impact on attitudes which in turn impact on things such as the willingness to farm, forage for food, and claim welfare.

5.2.4.2 Predictive validation

When validating the three possible network structures it was most important to try and measure their performance in approximating the actual causal relationships occurring within the system. Unfortunately this is a difficult thing to measure. An easier aspect to measure is the performance of each possible network when attempting to predict the food security status of a test set of households. Though this does not give us any indication of being a good causal model, it does tell us how useful a tool the networks are in at least one respect and gives us information about what the networks can tell us about the food security of observed households.

To achieve this we performed a 10-fold cross validation for each network, where the network structures were kept constant but parameter values were allowed to vary dependent on the data subset they were being fitted to. Due to the fact that predictive performance does not capture what we really need, we used three different measures of predictive performance for probabilistic classifiers to allow us to best distinguish between the networks.

The first is the Briers score, as used in Sections 4.3 and 4.4, defined as

$$S_B = \frac{1}{M} \sum_{i=1}^{M} (Y_i - \hat{p}_i)^2 \tag{5.6}$$

where M is the number of observations in the cross validation set being predicted for, $Y_i \in \{0, 1\}$ is the food security status of household i as defined in Section 5.2.1, and $\hat{p}_i \in [0, 1]$ is the fitted marginal probability of $Y_i = 1$ (Schwarz et al., 1978). Briers score gives a heavier penalisation the lower the predicted probability of the correct food security state of the household, with severe penalisation given for particularly extreme incorrect results.

The second is information reward, defined as

$$S_{IR} = \frac{1}{M} \sum_{i=1}^{M} (1 + \log_2 \tilde{p}_i) \tag{5.7}$$

where

$$\tilde{p}_i = \begin{cases} \hat{p}_i & \text{if } Y_i = 1\,, \\ 1 - \hat{p}_i & \text{if } Y_i = 0\,. \end{cases} \tag{5.8}$$

Information reward prefers models that estimate probabilities of the correct state that are better than random, i.e. 0.5, and penalises ones that are worse (Good, 1952).

The final measure is Bayesian information reward, defined as

$$S_{BIR} = \frac{1}{M} \sum_{i=1}^{M} \left(1 - \frac{\log \tilde{p}_i}{\log \tilde{\pi}_i}\right) \tag{5.9}$$

with

$$\tilde{\pi}_i = \begin{cases} \pi & \text{if } Y_i = 1\,, \\ 1 - \pi & \text{if } Y_i = 0 \end{cases} \tag{5.10}$$

where π is some chosen prior probability of any given household being food insecure. Bayesian information reward prefers models that estimate probabilities of the correct states that are better than the chosen prior probabilities π and $1 - \pi$, and then penalises ones that are worse (Hope and Korb, 2004). We take π to be equal to the empirical frequency of food insecure households in the training data. This measure then indicates whether a given network structure allows for better prediction of food insecurity than what we can tell by simply looking at the data.

Learning method	S_B	S_{IR}	S_{BIR}
Expert elicitation	0.1321	0.3908	-1.9004
Data-learned	0.0851	0.5577	-0.0270
Data-learned with expert prior	0.0846	0.5590	0.0057

Table 5.10: Model comparison of predictive performance for the different network structures learned as described in Section 5.2.2 from a 10-fold cross validation. Briers score, information reward, and Bayesian information reward indicate that learning from data massively outperforms learning from experts, but using expert knowledge as a prior provides an additional small increase in performance.

Unsurprisingly learning the structure from data gives much greater predictive performance than relying on the expert elicitation (Table 5.10). The data gives a greater population view of what is happening than the experts. It is surprising, though, that the data gives no improvement on performance over simply relying on the empirical data frequencies. Interestingly, relying on data-learning with the expert network as a prior results in a small increase in performance over even the data-learned method. It is also the only method to improve upon relying on the empirical frequencies.

5.2.4.3 Causal validation

There is no established and widely accepted way to measure the faithfulness of a model to the causal relationships of the system it represents using observational data. Indeed it is arguably impossible to do so. The best way to measure the causal validity of these three networks would be to perform interventional studies in the Agincourt study area, examining how implementing interventions focused around particular variables whilst other variables are controlled for impacts upon the food security of households. Beyond the ethical difficulties of such experiments, they would also come at great expense and would take a great deal of time and are therefore well beyond the scope of this work.

In order to attempt some kind of causal validation we relied upon an imperfect method of simulating past food security interventions performed on similar populations on the models and see which (if any) successfully capture these results. To do this we performed a thorough search of the literature, looking for interventions aimed at food security in populations similar to that of the Agincourt study area.

It was found that several agricultural interventions have been performed in the form of community agricultural projects, homestead food production, and the development of gardens in countries and regions such as Bangladesh, South-East Asia, Latin America, Rwanda, South Africa itself, and other low to middle income countries (Helen Keller International, 2010; Bushamuka et al., 2005; Galhena et al., 2013; Gaihre et al., 2016; Rose, 2008; Nsabuwera et al., 2016; Faber et al., 2011). Various financial interventions have also been performed, such as providing public sector employment and cash transfers in Latin America and Ethiopia (Rose, 2008; Coll-Black et al., 2012), as well as micro-loans and micro-credit in sub-Saharan Africa (Stewart et al., 2010; Van Rooyen et al., 2012). In addition to these, we also found interventions on water quality throughout Asia and Africa (Gaihre et al., 2016).

These past interventions gives us six possible simulations, where we can simulate setting the values for the 'use of crops and livestock' (UCL), 'child grant status' (CGS), 'selling of crops and livestock' (SCL), 'socio-economic status' (SES), 'employment level' (EmL), and 'water access' (WA) variables individually and look at the impact on the probability of the household being food secure. If the networks capture these interventions, then increasing the state of each of these variables should increase the probability of a household being food secure.

It should be noted that whilst some of these interventions were parts of controlled studies, others were reports of government strategies that therefore may not have successfully controlled for confounding variables which may undermine any causal implications we can take from them. However, ignoring these interventions would severely reduce the already small number that can be checked. As there are only six simple interventions to check, there is already a severe courseness to this test. Also though the populations may be similar, they may also be different enough that interventions that work on them may not work on the Agincourt population. One notable difference between South African populations in general and those from less economically developed countries is their access to welfare. However, there is still value in seeing whether our networks capture these past interventions, and if not then why. This is the closest we can get to a true test of causal faithfulness without performing actual studies and experimentation in the field. Even if we cannot consider it a test of such faithfulness, if we assume this faithfulness to be true then it highlights possible differences between the system we study and other similar systems.

The queries were performed using the inference techniques described in Section 2.3.3. For each possible query where we wish to find the probability of FS $= 0$ given

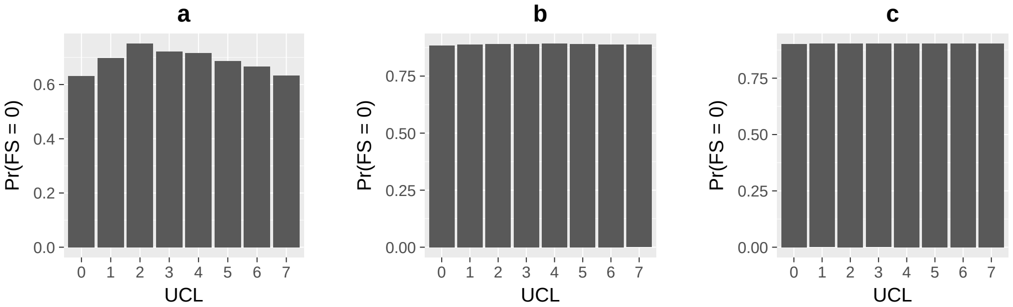


Figure 5.7: Simulations of setting the state of 'use of crops and livestock' (UCL) in order to alter the probability of a household being food secure ($FS = 0$) on the different possible Agincourt food security belief networks. a - expert elicited structure. b - data-learned structure. c - data-learned structure with the expert network as a prior.

that some variable E is set to state e, we must calculate the marginal conditional probability $\Pr(\text{FS} = 0 \mid \text{do}(E = e))$. Note here that $\text{do}(E = e)$ indicates we have gone into the system and artificially set the value of E regardless of the values of its parent nodes. Therefore we find that often $\Pr(\text{FS} = 0 \mid E = e)$ is different from $\Pr(\text{FS} = 0 \mid \text{do}(E = e))$ as the former is merely the probability of observing a household that is food secure given it has $E = e$. Calculating this observational marginal probability is as simple as taking the complete network and performing the inference on it as described, with E set to e. To calculate $\Pr(\text{FS} = 0 \mid \text{do}(E = e))$ we must first remove the links from the parents of E to E, representing how our interference with the system has rendered E now independent of its parents, before performing the same inference as described with E set to e. This small difference is what differentiates interventional queries from observational ones (Koller and Friedman, 2009).

The simulation results (Figures 5.7 to 5.12) show that all three networks fail to capture the literature interventions. The expert elicited network even shows very counter-intuitive results for employment level, child grant status, and selling of crops and livestock, where increases in these lead to decreases in the probability of being food secure. This is also observed for employment level in the data-learned network. As the expected result of increasing the states of these variables, which is that the frequency of food secure individuals should increase, largely appears in the data we must conclude that the other links existing between the interventional nodes and food security are what lead to these unexpected relationships. It is possible that what we expect to see may not actually be true and these other links are helping to

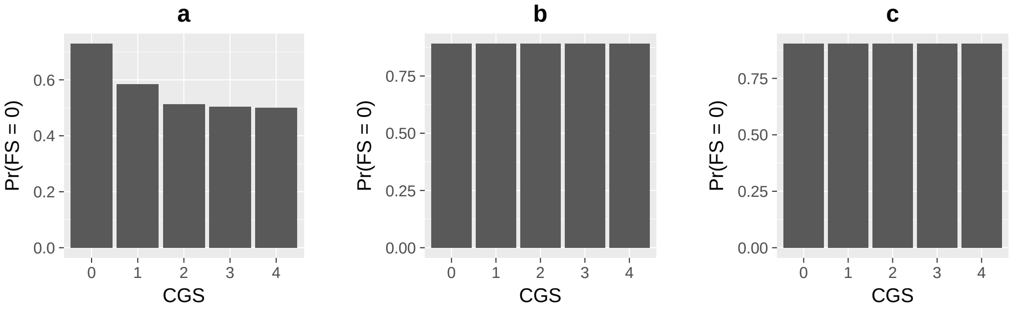


Figure 5.8: Simulations of setting the state of 'child grant status' (CGS) in order to alter the probability of a household being food secure ($FS = 0$) on the different possible Agincourt food security belief networks. a - expert elicited structure. b - data-learned structure. c - data-learned structure with the expert network as a prior.

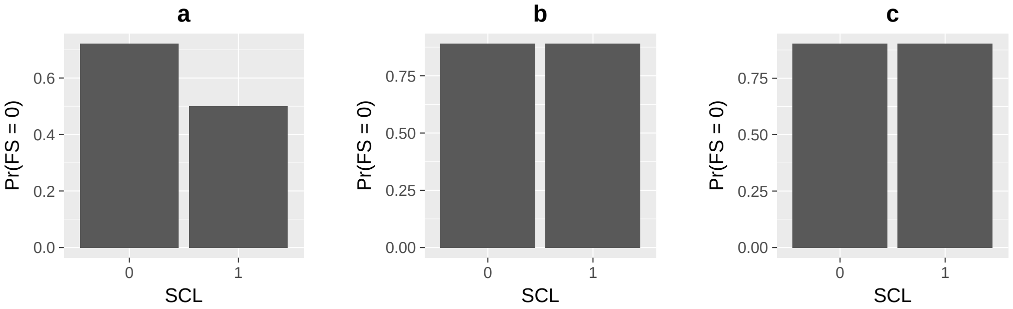


Figure 5.9: Simulations of setting the state of 'selling of crops and livestock' (SCL) in order to alter the probability of a household being food secure ($FS = 0$) on the different possible Agincourt food security belief networks. a - expert elicited structure. b - data-learned structure. c - data-learned structure with the expert network as a prior.

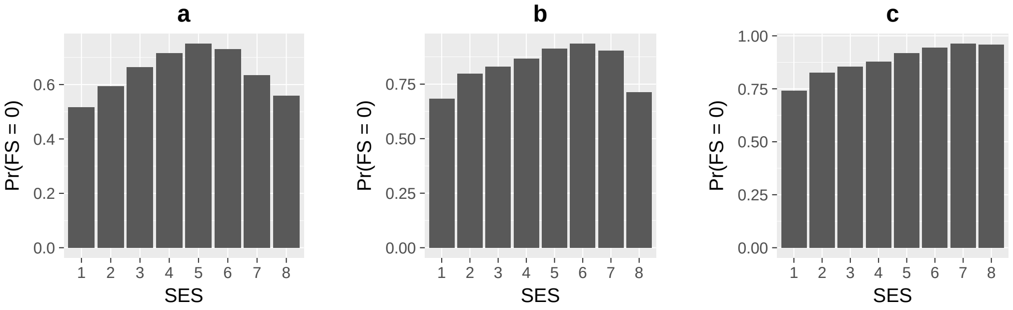


Figure 5.10: Simulations of setting the state of ‘socio-economic status’ (SES) in order to alter the probability of a household being food secure ($FS = 0$) on the different possible Agincourt food security belief networks. a - expert elicited structure. b - data-learned structure. c - data-learned structure with the expert network as a prior.

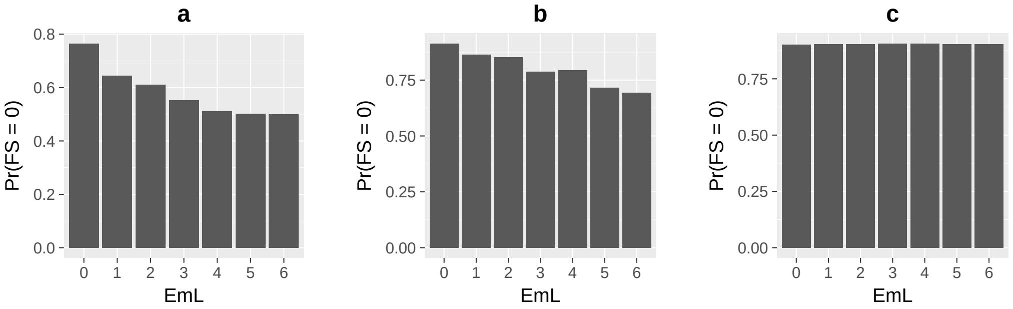


Figure 5.11: Simulations of setting the state of ‘employment level’ (EmL) in order to alter the probability of a household being food secure ($FS = 0$) on the different possible Agincourt food security belief networks. a - expert elicited structure. b - data-learned structure. c - data-learned structure with the expert network as a prior.

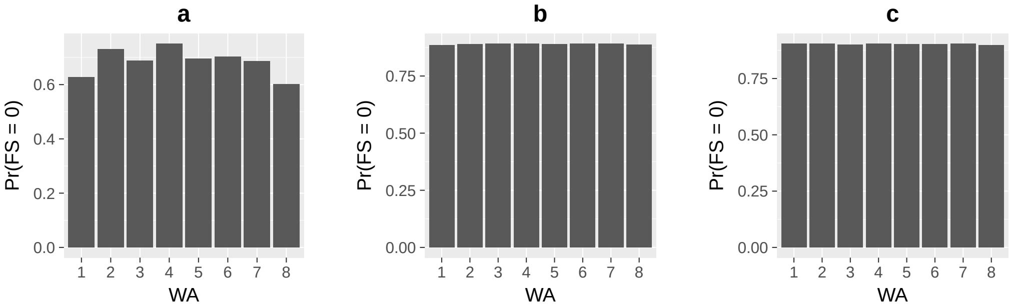


Figure 5.12: Simulations of setting the state of 'water access' (WA) in order to alter the probability of a household being food secure ($FS = 0$) on the different possible Agincourt food security belief networks. a - expert elicited structure. b - data-learned structure. c - data-learned structure with the expert network as a prior.

control for confounding, or it could also be that they are introducing further levels of confounding.

Another explanation for the failure of expected patterns to appear could be the courseness in our definition of the food security variable. Food security may change in more subtle ways throughout the year, so only considering lack of food over an entire year may fail to capture these subtleties. However, as mentioned before the food security Agincourt module has enough issues with missing data to make the calculation of a more intricate food security measure difficult.

For the data-learned networks, with and without expert priors, we see some nodes have no impact at all upon food security. In particular, variations in use and selling of crops and livestock, child grant status, and water access lead to no change in the probability of being food secure in both these networks, as does employment level for the data-learned network with expert prior alone. For many this is largely unsurprising due to the much sparser structures of the networks. If we step in and control these nodes, via the mechanisms of do-calculus as described above we should expect the probabilistic relationships between many of them and food security to be cut off.

For use of crops and livestock and water access, the expert elicited network shows some more unusual variations in the food security probability. However, for water access in particular this does not lead to any meaningful pattern of variation. On the other hand, use of crops and livestock does show some increase in food security probability for low values which then turns in to a decrease at higher values. This is not entirely what we expect to see, but is at least closer than many of the other results.

The only intervention that performs roughly as expected is socio-economic status, probably due to how well connected it is to food security in all three networks. This implies that socio-economic status is as important to food security as we expect it to be, perhaps more so than any other household characteristic. Given how it relates so heavily to many household characteristics it is not surprising that it could capture much of what makes a household food secure or not.

As mentioned before, the issues with what is arguably a naive test mean that these results to not negate the possibility of these networks being good causal models. They do, however, stress that much more must be done to test the causal validity of these networks and work towards a true causal representation of the system which would give the greatest utility in informing interventions against food insecurity. As said, such work should involve interventional studies, but perhaps could also consider a more indepth search through current methods of causal validation from observational data with perhaps the aim of developing a widely accepted (though still obviously inferior) alternative to experimentation. Unfortunately this is well beyond the scope of this current work. However the results presented here may give possible focuses for these future experiments.

5.3 Applications of the Agincourt food security belief network

5.3.1 Sensitivity analysis

Though the validation tests did not enable us to confirm the causal validity of the different possible networks (Section 5.2.4.3), it is still of interest to look at possible applications of them. One such interesting application is in using them to find which of the variables have the greatest probabilistic impact on food security, namely through a sensitivity analysis. This was performed following Barons et al. (2014) by calculating values for mutual information reduction and expected change of belief.

The mutual information reduction is defined as

$$I\left(\mathrm{FS}\right)=H\left(\mathrm{FS}\right)-H\left(\mathrm{FS}\mid X\right)=\sum_{\mathrm{fs}\in\mathbb{S}_{\mathrm{FS}}}\sum_{x\in\mathbb{S}_X}\log_2\frac{\Pr\left(\mathrm{fs},x\right)}{\Pr\left(\mathrm{fs}\right)\Pr\left(x\right)}\tag{5.11}$$

where the summations are taken over all the possible states of food security FS and node X, and

$$H\left(X\right)=-\sum_{x\in\mathbb{S}_X}\Pr\left(x\right)\log_2\Pr\left(x\right)\tag{5.12}$$

is the information of node X. This essentially measures how learning the state of X changes our uncertainty about the food security FS of the household, and can also be expressed as a percentage of $H\left(\mathrm{FS}\right)$ i.e. our uncertainty about FS. The expected change of belief is defined as

$$S^2\left(\mathrm{FS}\right)=\sum_{\mathrm{fs}\in\mathbb{S}_{\mathrm{FS}}}\sum_{x\in\mathbb{S}_X}\Pr\left(\mathrm{fs},x\right)\left[\Pr\left(\mathrm{fs}\mid x\right)-\Pr\left(\mathrm{fs}\right)\right]^2\tag{5.13}$$

and measures the change in the posterior probability of food security upon learning the state of node X. Both of these measures essentially give an indication of which nodes have the most impact on food security.

The expert elicited network shows food security having a greater sensitivity to a wider range of variables, including child grant status, vegetation and employment levels, use of wild foods, and selling of crops and livestock (Table 5.11). This could perhaps be due to the experts exhibiting a more nuanced understanding of food security that is beyond the relative simplicity of our network building methodology and variable definitions. The data-learned network, being much sparser, shows smaller effects (Table 5.12) though socio-economc status, use of wild foods, receiving communal aid, and employment level all show reasonable impact. Finally, the data-learned network using the expert network as a prior is sparser still, and therefore unsurprisingly most of the variables have only very small impact on food security. Socio-economic status is the only one to still have some substantial effect, with use of crops and livestock and use of wild foods both having slightly smaller impacts that are at least bigger than those of all the other variables.

These results indicate possible avenues for interventional studies if these networks are taken with a causal interpretation. Otherwise it shows us the variables with the greatest probabilistic relationships with food security, which may aid in streamlining efforts to identify potential food insecure households which may be helpful given how rarely food security information is collected compared to other information. The

Variable	I	I/H	S^2
CGS	0.0099	0.0115	0.0030
EdL	0.0078	0.0091	0.0023
EmL	0.0177	0.0205	0.0052
HHG	0.0000	0.0000	0.0000
LLV	0.0264	0.0307	0.0079
ND	0.0021	0.0025	0.0006
NWAA	0.0007	0.0008	0.0002
RCA	0.0083	0.0097	0.0025
RS	0.0000	0.0000	0.0000
SCL	0.0108	0.0126	0.0034
SES	0.0067	0.0078	0.0020
UCL	0.0039	0.0046	0.0011
UWF	0.0334	0.0388	0.0101
WA	0.0049	0.0056	0.0014

Table 5.11: Sensitivity analysis results showing mutual information reduction I (also as a percentage of food security information H) and expected change of belief S^2 for food security given each of the other variables for the expert elicited network. CGS - child grant status. EdL - education level. EmL - employment level. HHG - household head gender. LLV - level of local vegetation. ND - number of dependents. NWAA - number of working age adults. RCA - receipt of communal aid. RS - refugee status. SCL - selling of crops and livestock. SES - socio-economic status. UCL - use of crops and livestock. UWF - use of wild foods. WA - water access.

Variable	I	I/H	S^2
CGS	0.0002	0.0004	0.0000
EdL	0.0001	0.0001	0.0000
EmL	0.0082	0.0165	0.0013
HHG	0.0000	0.0001	0.0000
LLV	0.0028	0.0056	0.0005
ND	0.0008	0.0016	0.0001
NWAA	0.0005	0.0011	0.0001
RCA	0.0066	0.0133	0.0010
RS	0.0003	0.0005	0.0000
SCL	0.0000	0.0000	0.0000
SES	0.0114	0.0228	0.0017
UCL	0.0008	0.0017	0.0001
UWF	0.0066	0.0133	0.0012
WA	0.0001	0.0003	0.0000

Table 5.12: Sensitivity analysis results showing mutual information reduction I (also as a percentage of food security information H) and expected change of belief S^2 for food security given each of the other variables for the data-learned network. CGS - child grant status. EdL - education level. EmL - employment level. HHG - household head gender. LLV - level of local vegetation. ND - number of dependents. NWAA - number of working age adults. RCA - receipt of communal aid. RS - refugee status. SCL - selling of crops and livestock. SES - socio-economic status. UCL - use of crops and livestock. UWF - use of wild foods. WA - water access.

Variable	I	I/H	S^2
CGS	0.0000	0.0000	0.0000
EdL	0.0000	0.0000	0.0000
EmL	0.0001	0.0002	0.0000
HHG	0.0002	0.0004	0.0000
LLV	0.0000	0.0000	0.0000
ND	0.0001	0.0001	0.0000
NWAA	0.0007	0.0014	0.0001
RCA	0.0000	0.0001	0.0000
RS	0.0007	0.0016	0.0001
SCL	0.0000	0.0000	0.0000
SES	0.0112	0.0244	0.0014
UCL	0.0011	0.0025	0.0001
UWF	0.0043	0.0093	0.0007
WA	0.0000	0.0000	0.0000

Table 5.13: Sensitivity analysis results showing mutual information reduction I (also as a percentage of food security information H) and expected change of belief S^2 for food security given each of the other variables for the data-learned network with the expert network as a prior. CGS - child grant status. EdL - education level. EmL - employment level. HHG - household head gender. LLV - level of local vegetation. ND - number of dependents. NWAA - number of working age adults. RCA - receipt of communal aid. RS - refugee status. SCL - selling of crops and livestock. SES - socio-economic status. UCL - use of crops and livestock. UWF - use of wild foods. WA - water access.

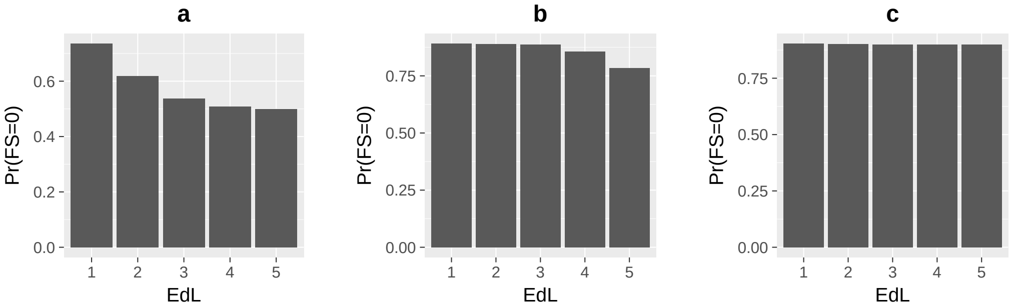


Figure 5.13: Interventional inference on the impact of 'education level' (EdL) on the probability of being food secure ($\Pr(\text{FS} = 0)$). a - expert elicited network. b - data-learned network. c - data-learned network with the expert network as a prior. Depending on the network, education has either a negative impact or none at all on food security.

expert elicited network results also give an indication of what variables have the greatest impact on the beliefs of a member of the community when considering what other households may be food insecure.

5.3.2 Interventional inferences

The most useful application of a causal Bayesian network is in simulating possible interventions on it to see whether they should be attempted in actuality. As such, for the final part of this chapter we show some possible interventional inferences based on findings from the literature. Unlike in Section 5.2.4.3 these are not based on any established interventions, but simply on what the literature has determined should have some effect on food security in the Agincourt study area.

For instance, various observational studies imply that education, communal aid, local vegetation, and being a refugee should each individually have a substantial effect on food security (Leyna et al., 2007; Kaschula, 2011; Nawrotzki et al., 2014). However, our inferences imply that these variables actually have only negative, if any, impact on food security (Figures 5.13 to 5.16) dependent on the network used. This only matches the literature for refugee status, where it implies we should expect a negative impact. However, we must note that the networks show only at best a small effect from refugee status.

Dovie et al. (2007) found that use of wild foods is prevalent in the Bushbuckridge area. Performing an inference on it finds that increased use of wild foods leads

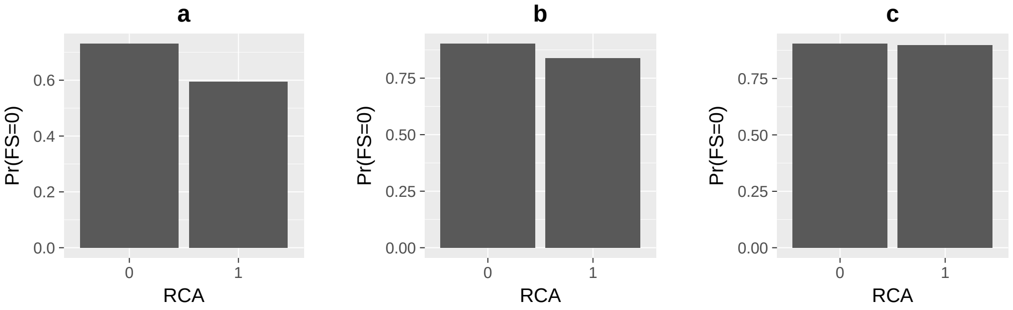


Figure 5.14: Interventional inference on the impact of 'receipt of communal aid' (RCA) on the probability of being food secure $(\Pr(\text{FS} = 0))$. a - expert elicited network. b - data-learned network. c - data-learned network with the expert network as a prior. Depending on the network, receipt of communal aid has either a negative impact or none at all on food security.

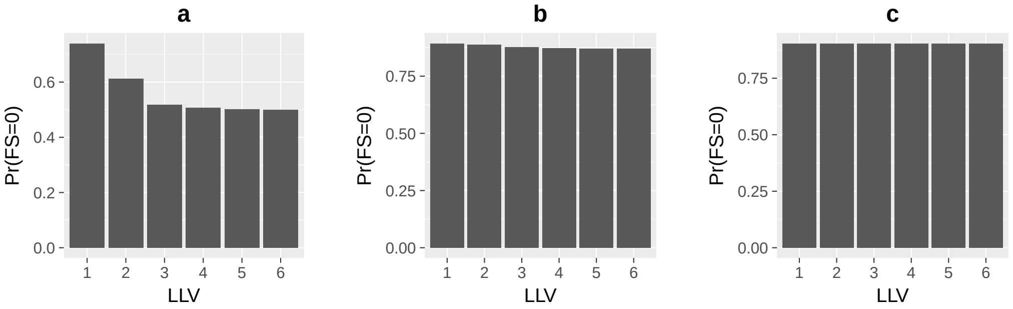


Figure 5.15: Interventional inference on the impact of 'level of local vegetation' (LLV) on the probability of being food secure $(\Pr(\text{FS} = 0))$. a - expert elicited network. b - data-learned network. c - data-learned network with the expert network as a prior. Depending on the network, local vegetation has either a negative impact or none at all on food security.

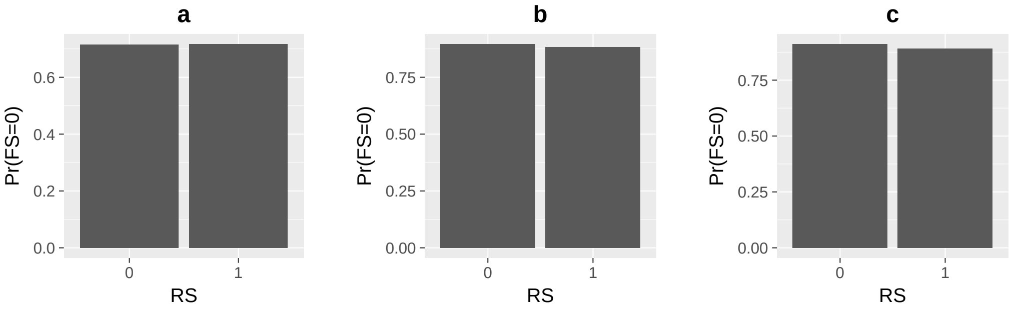


Figure 5.16: Interventional inference on the impact of 'refugee status' (RS) on the probability of being food secure $(\Pr(\text{FS} = 0))$. a - expert elicited network. b - data-learned network. c - data-learned network with the expert network as a prior. Depending on the network, refugee status only a slight negative impact on food security.

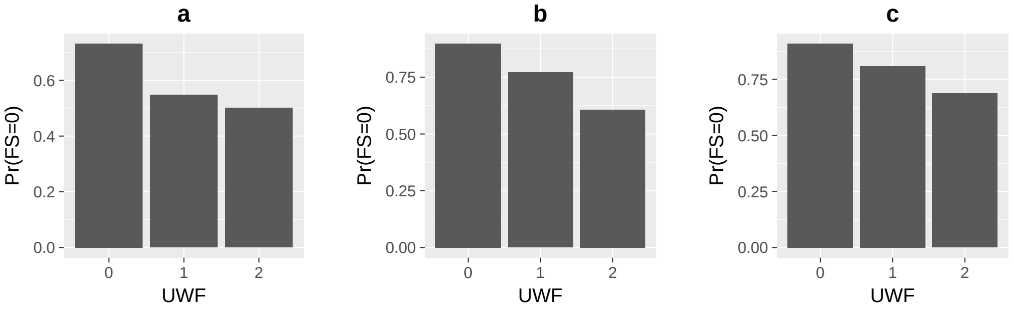


Figure 5.17: Interventional inference on the impact of 'use of wild foods' (UWF) on the probability of being food secure $(\Pr(\text{FS} = 0))$. a - expert elicited network. b - data-learned network. c - data-learned network with the expert network as a prior. Use of wild foods consistently has a negative impact on food security irrespective of the network.

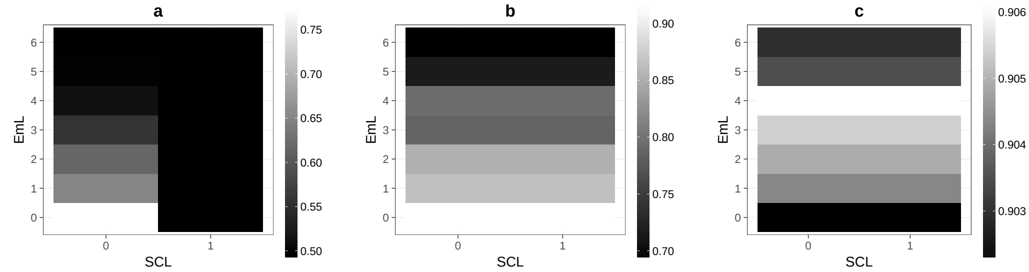


Figure 5.18: Interventional inference on the impact of both 'employment level' (EmL) and 'selling of crops and livestock' (SCL) simultaneously on the probability of being food secure $(\Pr(\mathrm{FS} = 0))$. a - expert elicited network. b - data-learned network. c - data-learned network with the expert network as a prior. Only the data-learned network with expert prior shows a positive impact on food security, whilst the others show only negative impacts.

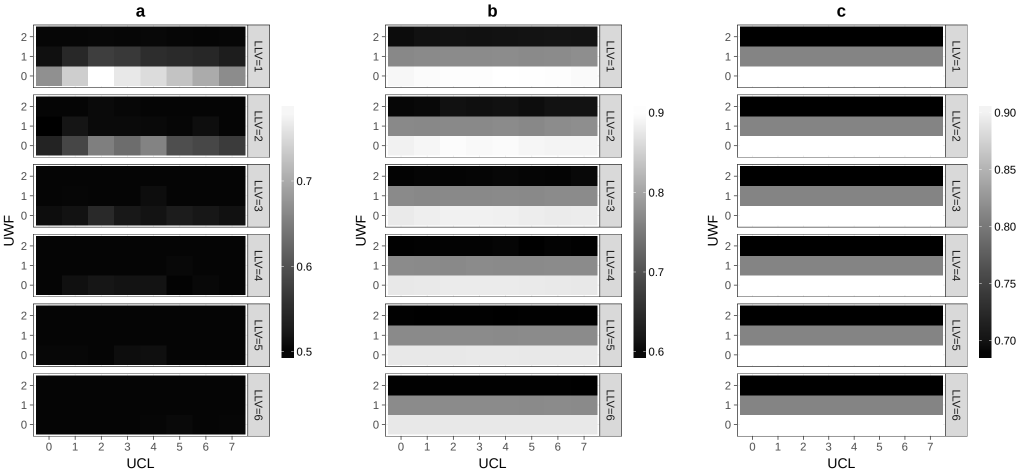


Figure 5.19: Interventional inference on the impact of 'level of local vegetation' (LLV), 'use of wild foods' (UWF), and 'use of crops and livestock' (UCL) on the probability of being food secure $(\Pr(\mathrm{FS} = 0))$. a - expert elicited network. b - data-learned network. c - data-learned network with the expert network as a prior. Only the two data-learned networks show substantial variation of food security probability.

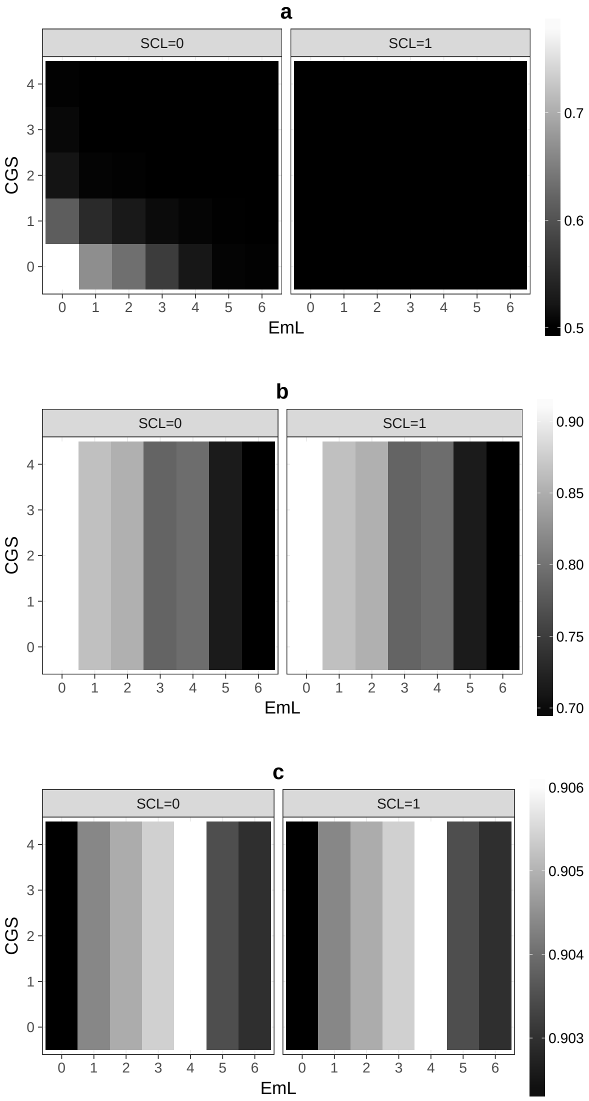


Figure 5.20: Interventional inference on the impact of 'child grant status' (CGS), 'employment level' (EmL), and 'selling of crops and livestock' (SCL) on the probability of being food secure ($\Pr(\text{FS} = 0)$). a - expert elicited network. b - data-learned network. c - data-learned network with the expert network as a prior. Only the two data-learned networks show substantial variation of food security probability, which is only positive for the one with the expert prior.

to lower probability of food security (Figure 5.17) completely irrespective of the network. It is possible here that the direction of the relationship is incorrect, as perhaps more food insecure households use foraging as a coping strategy so that use of wild foods only occurs after food insecurity has been attained.

Mabuza et al. (2016) found that non-farming income provides for greater food security than farming income. If we look at a combined inference of selling of crops and livestock and employment level at the same time (Figure 5.18) we find that selling crops and livestocks makes largely no difference. Controlling for it though does allow for an impact from employment level, though this is only positive for the data-learned network with the expert prior where the effect is also quite small.

MKaibi et al. (2015) examined the impact of both agriculture and the environment on food security. We can do the same by performing a combined inference on level of local vegetation, use of wild foods, and use of crops and livestock (Figure 5.19). Both vegetation level and use of crops and livestock have little to no impact, but holding them fixed again allows an impact from use of wild foods (though not so much for the expert elicited network). As we have seen before, this impact is still negative for both data-learned networks which perhaps lends greater support to our earlier hypothesis of the relational direction between food security and use of wild foods being opposite to that in the models.

Finally, Pereira et al. (2014) examined the impact of income on food security. We can look at different income sources by performing a combined inference on selling of crops and livestock, child grant status, and employment level (Figure 5.20). As expected from the sensitivity analysis, child grant status and selling of crops and livestock have little impact. Therefore we conclude that actual income is more important to food security than welfare or commercial farming, though it is only a positive impact for the data-learned network with expert prior again. This implies that it is perhaps the paths going through these other nodes in the network that leads to the switching of the relationship between employment and food security to being negative.

Though we have failed to definitively determine a causal model for our system, we have shown the great advantages of Bayesian networks over linear models. There is far more utility in Bayesian networks, where the inbuilt aspects of representing conditional independence, encoding both direct and indirect causal relationships, and modelling a many-to-many set of relationships allows us to perform far more useful simulations such as these interventional inferences. In addition, Bayesian networks are far more intuitively understandable for non-technical individuals such as medical practitioners and government policy-makers who need to rely upon the models we build for making decisions. This is also achieved whilst encoding a great amount of detail and complexity. Instead of strict coefficients showing precise effects (which is an unrealistic result to attain from any model) we have probability distributions allowing us to measure the impacts of different actions on the likelihood of something happening. Overall, we find much more usefulness and flexibility in these models than in what has been used before. They can now stand as justifications for interventional studies and funding decisions, as well as starting points for the building of further models of food security in similar populations.

CHAPTER 6

Discussion

In this research we have attempted to address the need to move beyond generalised linear models in the statistical modelling of various different health aspects dependent on socio-economic variables. This has included using non-linear parametric models to introduce greater assumptions needed to overcome basic confounding in the causal modelling of social contagion, the use of the semi-parametric method of Gaussian processes to overcome unjustified a priori assumptions of linearity in the modelling of fertility patterns, the additional use of the more complicated aspects of Gaussian processes to enable us to analyse social contagion with datasets lacking in social network variables, and finally modelling household food security using Bayesian belief networks to provide more realistic and flexible models that allow us to consider indirect causal relationships when modelling interventions.

In this final chapter we consider the research as a whole, in the context of our initial aims as outlined in Section 1.3. We first present the concept of a complex system which gives a core justification for the overall aim of the research. We then explore both the strengths of the work as well as the weaknesses, with specific focus given to the importance of community involvement from those the research is about, the lack of use of more powerful modern black box models, and the difficulty in capturing causality. We then consider how well the analyses performed have allowed us to achieve our aims, both in terms of the individual contexts as well as within public health as a whole, before looking towards where this research can lead in the future. Finally, we briefly discuss a side benefit of the work in showing the importance of collaboration between scientists of different disciplines.

6.1 Public health as a complex system

A complex system is commonly defined as "a dynamic network of agents acting in parallel, constantly reacting to what the other agents are doing, which in turn influences behaviour and the network as a whole" (Foundation, 2010). The key aspect of this definition is that a complex system is one formed of many parts that either have many complex interactions or exhibit some kind of complex dynamics. For social systems that consider populations, the complexity usually arises from the large amount and complicated nature of the interactions between the various members and factors of the population at different scales alongside the various structural, geographical, environmental, historical, cultural, and other such elements that influence the population. There are many common characteristics of complex systems, such as emergence of large scale phenomena resulting from the small scale dynamics and interactions, nestedness within other systems, high levels of connectivity between the different elements of the system, inherent non-linearity, and self-organisation (Plsek and Greenhalgh, 2001; Foundation, 2010). It is these characteristics that lead to the difficulties of effectively modelling and applying interventions to such systems.

Now many are calling for the consideration of complexity and complex systems in the analysis and implementation of many aspects of healthcare. This includes the considerations of healthcare infrastructure as complex systems, including financial transactions (Bar-Yam, 2006) and transferral of knowledge (Smith and Joyce, 2012), as well as different aspects of healthcare such as nursing (Chaffee and McNeill, 2007), clinical care (Wilson and Holt, 2001), primary care (Ellis, 2010), and GP consultations (Love and Burton, 2005). It has also been argued that complexity must be considered in both the management of healthcare systems (Plsek and Greenhalgh, 2001), as well as the education of healthcare professionals (Fraser and Greenhalgh, 2001). Health in general has been proposed as a complex system as it is dependent on many different aspects from the biological to the social at a continuous variety of scales (Topolski, 2009).

The word 'complex' has also become a central part of the design of health interventions, where complex interventions have risen up as an intervention design that includes multiple components designed to address the different aspects of the problem (Campbell et al., 2007; Craig et al., 2008; Hawe et al., 2004). It should be noted there is a definite distinction between complex interventions and complex systems (Shiell et al., 2008). However, that does not take away from the fact that designing complex interventions allows us to more effectively address the needs of complex systems.

It is clear that the systems that public health research deal with, which tend to be on the population scale, are indeed complex. According to Sawyer, "societies are complex configurations of many people engaged in overlapping and interlocking patterns of relationships with one another" (Sawyer, 2005). Due to the many and varied interactions we have with each other, any social system must be treated as complex in order to be effectively modelled (Poli, 2013; Ramos, 2016; Goldspink, 2000). All of this commentary has lead to the call from many for a re-evaluation of the methods of thinking about and modelling our health and healthcare systems, including in public health research (Sterman, 2006; Smith and Joyce, 2012; Rutter et al., 2017).

The consideration of public health systems as complex systems provides much further support to the need to introduce the methods we have used to the library of those available to public health researchers, on top of the improvements we have seen in the actual analyses. That is not to say that all of our methods have completely addressed the complex nature of the systems. We have still tended to focus on single aspects of the system, which admittedly is mostly a practical necessity caused by restrictions in both the data available and the need to avoid the mathematics getting more complicated than is useful. Overall the use of Bayesian belief networks probably most closely captures the spirit of complex systems thinking and whollistic approaches. Our analyses do, however, address a large factor of complex systems thinking: the need to employ models that can better address the subtleties introduced by the complex nature of the systems being modelled.

6.2 Advantages of the research

6.2.1 Methodological advantages

The analyses presented here come with a great many advantages over previous work. The analyses were performed using methods that provided great improvements over past work, and show a general improvement over the statistical methodology that is common to public health research (see Section 1.2). Here we summarise these improvements, many of which have already been stated within the individual chapters. In summarising them here, we hope to fully demonstrate the benefit of having used more complex methods in these analyses, and the importance of implementing and developing such methods further in future research.

Chapter 3 dealt with the difficulty in making robust causal inferences in the face of highly probable confounding in the area of social contagion. The method involved comparing how well two competing models of behaviour state transition probabilities performed, one where the transition probabilities were stratified by friends of the contagious behaviour state in a form suitable for complex contagion and the other where no such stratification existed. As described in Section 3.2, most previous studies such as those of Christakis and Fowler (2013) have implemented generalised linear models where the state of one individual is modelled on the state of one friend, with controlling variables in the form of the states of the two individuals at the previous time points in order to control for confounding. Such studies are essentially examining the state probabilities of the individuals, where contagion cannot be distinguished from homophily or shared context. By abandoning generalised linear models and instead examining whether the transition probabilities are stratified by friends we have overcome this issue, as at least for the basic confounding phenomena of homophily and shared context we would not expect such stratification to occur, as we argued heuristically in Section 3.2 (a more formal explanation can be found in the electronic supplementary material of Hill et al. (2015)). It may do so for more complicated confounding variables, which certainly presents directions for future improvements, but it is arguable that if any simpler explanations can be ruled out then contagion, as the next simplest, may very well be the most likely.

For studies that did implement methods similar to ours, such as that of Hill et al. (2010a), they still modelled the contagious transition probabilities as linear functions of the number of contagious state friends. This ignores the established status of social contagion as a type of complex contagion (Centola and Macy, 2007; Centola, 2010; Valente, 1996). Also, by modelling social contagion as a complex contagion, we have used a functional form with greater structure which makes it more distinguishable in terms of performance from the non-stratified alternative. The many criticisms levelled against p-values and small linear coefficient values with no errors provided, as described in Section 1.2, also undermine the methods used in these previous studies. The maximum likelihood fitting, combined with information theoretic model comparison measures, that we used are more robust and trustworthy than measures of statistical significance.

In the more general context of public health modelling, the social contagion analyses presented in Chapter 3 show how going beyond generalised linear models into the realm of parametric non-linear models fitted using maximum likelihood methods result in much more robust findings. This is achieved by the introduction of greater assumptions in the analysis, which as we state in Section 6.3.3 is necessary to really be able to build an effective causal model. Whilst previous studies attempted to build in such assumptions via the controlling variables, this does not do so in a strict enough way to really be able to avoid confounding.

In Chapter 4 we presented two analyses linked by their use of Gaussian processes and the context of pregnancy in rural South Africa. The first fitted a descriptive model of fertility patterns dependent on age and socio-economic status using a combined non-linear parametric and semi-parametric method. Previous studies again relied on generalised linear models, only introducing non-linear terms for age where non-linear parametric models are already established (Camlin et al., 2004; Palamuleni and Adebowale, 2014; Ayele, 2015; Burger et al., 2012). This meant that fertility patterns were consistently assumed before any analysis was performed to be linear over any variables such as socio-economic status where no set model is established in the literature. By bringing in further variables via a Gaussian process regression of the fertility-age parameters over them, we have avoided any unjustified a priori assumption of linearity. Our method has allowed for building in structure where it has been justified, and allowing the data to dictate the structure when no set structure is known. Gaussian process regression involves few assumptions, so is perfect for modelling potentially non-linear relationships where all we require is for the relationship to be smooth and continuous (Rasmussen and Williams, 2006). It has to be admitted that we found the fertility patterns over socio-economic status to be linear, but this does not undermine our method as we had no a priori justification to assume this would be so. Future research involving the building of descriptive models for public health should take this into consideration, and implement Gaussian process regression to avoid the unfounded assumption of linearity that could lead to the failure to capture more subtle and complicated relationships that could be occurring.

The second analysis examined the possibility of social contagion of pregnancy risk behaviour amongst rural South African adolescents. Like many datasets, the Agincourt HDSS dataset used does not include social network data (unlike the Add Health dataset used in Chapter 3). If we relied on standard methods, we would quite possibly not be able to perform this analysis. By using more complex methods we overcame this issue, demonstrating how the use of more complex methods such as Gaussian process classification opens up many further avenues in public health research. The results we found were admittedly ambiguous, so future research should focus on the building of a pregnancy contagion measure for data lacking social network variables and a first principles parametric model for such a measure. However, the more complicated features of Gaussian processes, in this case the covariance matrix, allowed us to attempt to bypass the difficulty of establishing such a measure and model.

Finally Chapter 5 considered the building of three different potential causal Bayesian belief network models of household food security in rural South Africa. Like with each of the other analyses, this is another area where the primary research has been performed using generalised linear models, statistical tests, and summary statistics (Twine and Hunter, 2011; Hunter et al., 2007; Pereira et al., 2014; Kirkland et al., 2013; Nawrotzki et al., 2014). This has produced many worthwhile results, but there are limitations with what can be inferred from, and what actions can be justified by, these models. The Bayesian network models we presented better reflect the complexity inherent within the system, as argued in Section 6.1. This is achieved by the inbuilt modelling of indirect probabilistic relationships. In this way we end up with more of a 'many-to-many' model rather than the 'many-to-one' model that linear regression gives us. Accompanied by a causal explanation, this enables us to more effectively model different possible actions, events, and interventions on the system.

On top of this better reflection of complexity, network models also come with a higher level of interpretability. They better reflect what we see in the world around us, where many things cannot be simply considered as having direct and linear relationships with each other. The resulting models are therefore much more satisfying in both their utility and their ability to represent the system.

Though we were unable to test the causal faithfulness of our models to the system, this is a problem that is shared by all the methods that came before. Our models are still much more suitable for causally modelling the system, as they have the directional structure that is necessary in order to do so (Pearl, 2009). Even structural equation models that model linear one-to-one relationships are accompanied by a path diagram displaying the directional structure. The lack of such structure undermines the ability to make any causal inferences from the models of previous studies. Also on the topic of causality, with both this analysis and the social contagion analysis we always considered probabilistic causality, stopping us from making much more precise inferences that would be much less robust.

6.2.2 Significance of results

Though our core focus is on how these methods have provided improvements on those that have come before, it is also important to consider the significance of what each analysis found. Combined with the methodological advantages these show the full benefit of the research performed.

The results found in Chapter 3 can inform public health policy and the design of interventions against both depression and obesity in adolescents. Sub-threshold levels of depressive symptoms in adolescents is an issue of great current concern as they have been found to be very common, to cause a reduced quality of life, and to lead to greater risk of depression later on in life than having no symptoms at all (Bertha and Balázs, 2013; McLeod et al., 2016; Klein et al., 2013). Understanding that components of mood, as well as mood in general, can spread socially suggests that while the primary target of social interventions should be to increase friendship because of its benefits in reducing the risk of depression, a secondary aim could be to reduce spreading of negative mood. This is something that was not shown when only considering binary depression states (Hill et al., 2015). On the other hand, as the findings suggest that though negative mood spreads it does not do so to such an extent to push most individuals over into depression, we therefore find evidence to support that depression is not simply a case of low mood. Depression must clearly operate via different mechanisms from general mood and individual depressive symptoms.

Whilst we found general contagion of mood previously (Eyre, 2014), we did not find strong enough evidence for the spreading of lower weight, only higher. This lends support to the conclusion of obesity spreading, whilst weight changes in general do not. Interventions should therefore focus on the reduction of the spreading of obesity, rather than on the benefits of friendship in reducing the risk of obesity. As described in Section 3.1, obesity in children and adolescents has been a concern for a long time now, so these findings are significant in terms of improving the health of adolescents in order to ensure their health in later life.

The analysis in Section 4.3 on fertility patterns was entirely focused on methodological improvements rather than providing any important findings. As such, we simply re-establish the fertility age-pattern we have seen in both the data and the literature (Garenne et al., 2000; Peristera and Kostaki, 2007). More interestingly, our results imply that socio-economic status has no significant impact on the fertility patterns of the population, which stands in contradiction of past research that only examined empirical frequencies (Williams et al., 2013). It also raises implications of differences between the Agincourt population and other similar populations (Pradhan et al., 2015). This could perhaps be due to the homogeneity of the Agincourt population in terms of wealth compared to the wider South African population. Overall this implies that socio-economic status should not be a significant factor when choosing populations for actions or interventions focused around pregnancy in the Agincourt study area.

Given further work to more strongly reinforce the results, the findings of Section 4.4 could inform public health policy and intervention design against adolescent pregnancy in the Agincourt study area as well as possibly other similar populations. Though ambiguous, the findings do imply a lack of support for contagion of pregnancy risk behaviour amongst the adolescent population. Considering the high number of risks involved with adolescent pregnancy (Panday et al., 2009; Timæus and Moultrie, 2015; Christofides et al., 2014; Ramaiya et al., 2014), as well as the unacceptably high numbers of them occurring in sub-Saharan Africa (Panday et al., 2009), such interventions are of great concern. These findings imply that social effects should not be of great concern in designing these interventions. Therefore focus should be applied to the great many other factors which have been shown to contribute to increased risk of adolescent pregnancy (Pradhan et al., 2015).

If the causal interpretation of any one of the three possible household food security models presented in Chapter 5 can be confirmed, then that model will be a valuable tool in designing and simulating basic interventions against food insecurity in the Agincourt study area. In particular, the model can be evolved via an iterative process of using it to inform intervention trials and then using the results of these trials to further develop the model. Though food insecurity is decreasing in sub-Saharan Africa (Labadarios et al., 2011; Rosen et al., 2016) it is still a great concern of the Agincourt community, as discovered by our meetings with the Agincourt Community Advisory Group described in Section 5.1. Demonstrating actions and improvements in this area would obviously be of great importance to them, and may therefore further reinforce their support of the research unit (though support is already great due to the benefits the census already provides the area). Beyond Agincourt, the models could also form the basis for the design of similar models in similar populations across the world.

In addition, the results from the expert elicitation produced some very counter intuitive findings, as described in Section 5.3.2. The findings may be heavily related to the nature of the experts we used, as our experts were experts in life in the locality. Other studies, such as that of Barons et al. (2014), used experts that had a greater overview of a wide range of evidence on the topic in question and great understanding of that evidence. This is an obvious limitation of our work, though one borne from the desire for greater community involvement in research, but it raises the interesting possible future research area as to the effect of using different expert groups. We thought that perhaps the people living within the study area might be experts on what causes what within the area, but life experience may not be sufficient for these types of expert elicitations. Despite this it may be that the model can give insight into local views and assumptions within the study area population.

6.2.3 Community-driven research

Though the results we obtained in Chapter 5 from our expert elicitations are quite possibly biased by the locality of their knowledge, the community-driven nature of that particular analysis can still be viewed as a great advantage. It is important in public health, considering the complexity of the systems analysed, the public nature of research funding, and the possibility of biases in our research due to lack of knowledge or perspective, to involve the individuals who our research is based on into that research.

Avoiding community involvement comes from a research paradigm where we as researchers like to think of ourselves as objective observers separated from the system (Israel et al., 1998). However, a complex systems view of our world contradicts this idea. Due to factors ranging from politics and economics to culture and gender we may be limited in our own interpretations of the systems we study and the importance of which parts of these systems to study.

Staniszewska and Denegri state that community-driven research “aims to enhance the quality, appropriateness, acceptability and relevance of research, ensuring it addresses issues of importance to patients and the public” (Staniszewska and Denegri, 2013). By bringing in the perspective of the CAG, we gained greater insights into what is important to members of the community and accessed greater resources in terms of their local knowledge than we could have done if we did not involve them. Many medical researchers support community-driven research due to the benefits it provides their research (Goldberg-Freeman et al., 2010). Above all, community-driven research increases the relevance and usefulness of public health research to those who are the intended end users.

There are many difficulties with community-driven research (Israel et al., 1998; Staniszewska and Denegri, 2013), not least of which for us was the language difficulties in conveying complicated mathematical concepts to non-mathematical individuals (Colquhoun et al., 2013). This specific challenge was overcome using careful design of the expert elicitation questions as described in Section 5.2.2.1. An additional challenge, as has been mentioned many times before, is the possible bias inherent in the local nature of the expert knowledge used in elicitation. This can be overcome in future with further elicitations incorporating a wider range of expert knowledge. Despite this, involving the community was still advantageous in the initial proposal and justification of the research.

Community-driven research is a key part of the Agincourt HDSS (Kahn et al., 2012, 2007). Though this aspect is not the focus of this research, we feel it adds substantially to the part of the analysis where it was used.

6.3 Limitations of the research

6.3.1 Practical limitations

All research comes with limitations, and the analyses presented here are certainly no exception. Though a lot of effort is undergone by the Agincourt research unit to ensure the reliability of the Agincourt HDSS data, as detailed by Kahn et al. (2012), there are some errors, misreporting, and missing data that we are unable to account for. The dataset is of a size and quality though that these only produce minimal issues and do not seriously undermine the results presented here (Fottrell et al., 2008). Similar issues must also exist for the Add Health dataset, and due to our need to subset the data to only the students from so-called saturated schools this could possibly undermine the work. However, such subsetting was necessary in order to ensure the most complete and detailed social network could be analysed, as the saturated schools are the only ones where all students were allowed to list more than only one friend of each gender.

As just mentioned, there is certainly missing data within each dataset which we treated simply by ignoring observations with missing variables. It is possible that use of methods such as multiple imputation (Gelman et al., 2014) or expectation maximisation (Murphy, 2012) could have mitigated for this. However, when performing initial exploratory data analysis no obvious missing data patterns appeared that would have strongly implied the use of these methods to be necessary. We do admit though that performing a thorough missing data analysis of both the Agincourt and Add Health datasets would have been useful in order to establish exactly what missing data patterns exist.

There also exist various methodological limitations for each of the analyses. For the social contagion work of Chapter 3, as noted in Section 3.2 we were not able to formulate from first principles a fully general model for the individual's state as a function of friends' states, though efforts were made towards this in Section 3.5. We were also unable to learn such a model from data due to the sample size of the study being constrained by the necessity of constructing as complete a friendship network as possible, as mentioned above. The friendship network itself may also not be complete. However, as the majority of individuals failed to list the maximum number of friends allowed, this implies that the network may, in fact, approach completeness. One other issue lies in the measure of BMI used to measure body fat in the weight change contagion analysis (Section 3.3) as evidence exists to suggest that it is a poor measure of how someone's physique is related to their health (Tomiyama et al., 2016). It is arguably still a good course measure at the population level, so the use of it does not necessarily negate the results of this analysis, but further work should consider other possible measures.

The methodology used in Section 4.3 to analyse fertility patterns also comes with some limitations, principally produced by the use of a parametric model and the decision to use regression techniques. The parametric regression must be performed for each combination of values for the non-age covariates, which means that introducing further covariates reduces the performance of the regression, a situation that can only be mitigated by using more data or not relying on a parametric model. However, the use of the parametric model allows us to definitely capture the age-pattern shown in our empirical data and in fertility age-patterns for many populations in the literature (though admittedly prevents us from being able to capture possible details such as a second fertility peak). In order to use regression techniques we have to bin the observations in the data to quantiles, which results in the removal of information. We mitigated against this by using cross validation and goodness-of-fit techniques to choose between different numbers of quantiles to use. The use of regression techniques also ensures we can produce visualisations of the relationships that can give us insight in to what is going on, and not just predictions alone. The age-pattern model also includes normalisation terms, resulting in accurate inferences for the shapes of the relationships but inaccuracies in the inferred numerical values. The model also results in interaction effects from each other covariate with age, but no main effects from the individual covariates.

Similar to our inability to formulate a general first principles model of an individual's state given their friends' states in the earlier social contagion work, when analysing possible contagion of pregnancy risk behaviour in Section 4.4 we were also unable to formulate a first principles model for social contagion that could be applied to datasets lacking any social network data. Unlike the earlier work, we did have a large enough sample size to attempt to learn such a model from the data, though as far as the data was concerned the contagion variable as defined had little to no impact on the pregnancy probability. That fact highlights another limitation in the choice of covariates. Though these covariates were chosen in an informed fashion based on known pregnancy risk factors, it is still true that for this method the contagion variable is entirely dependent on this choice. Instead of defining contagion in terms of similarities in risk, it may be better to look at the risk of exposure to each others' behaviours. However, this comes with much greater difficulties as there is less research into what factors make individuals more likely to note each others' pregnancy risk behaviours, and such factors are also not as commonly available in the data. We also noted an ambiguity in the results, leading to the conclusion of no evidence of social contagion but also not enough evidence to completely rule it out. A parametric approach, though more difficult to formulate, may overcome this ambiguity. Such a formulation is beyond the scope of this work, and despite the ambiguous results this does not undermine the advantages in using a more flexible model like Gaussian processes. If such methods did not exist, we would have much greater difficulties even attempting research like this.

Finally, limitations in the Bayesian network models of household food security presented in Chapter 5 focus mainly around the difficulties in statistically modelling a causal system. The lack of incorporating time certainly undermines the modelling of causality, as that is arguably an essential part of doing so. This could have been overcome by designing a dynamic Bayesian network, but doing so would have severely limited the amount of data that could be used to estimate each parameter as well as requiring the uncertain assumption that causal effects last over three years due to the gap between collection of the food security census module. It would have also presented a much greater additional burden to our experts in the elicitation, which we were expressly unable to do due to our limited resources. There are also issues in the learning of the networks. On one side, the experts showed possible substantial bias which is unsurprising given the similarities in their life experiences to one another, though this does present the possibility of the network as a useful model of said biases. On the other side, the algorithms used to learn the networks incorporate assumptions as to the completeness of the data in terms of its ability to cover all variables needed to fully causally model the system without the need of any latent variables. This means that again the choice of variables is a great limitation as it is possible that latent variables may have been missed, though this was again mitigated by choosing the variables in an informed way. Overall, these limitations do not undermine the usefulness of developing these models, though do emphasise the need to perform more research with much greater resources before the models can be fully put into practise to model interventions.

6.3.2 More complex methods

Though research within the areas of public health and epidemiology still rely heavily on generalised linear models and hypothesis testing for all statistical modelling, researchers from the areas of machine learning and computer science have been making in-roads into many areas including medicine with the use of more modern methodology. Over the last few years, we have seen the rise of deep learning as a very powerful statistical technology with great successes seen across science and technology (LeCun et al., 2015). Deep learning refers to neural networks where variables in the data (such as the pixel values of an image) are passed through many levels of linear combination and non-linear transformation before outputting a result (usually a classification). In doing so, the algorithm manages to capture deeply complex patterns that exist within the data (such as combinations of edges in images). Alongside many further developments, such as convolution, recurrency etc., such algorithms are currently allowing for dramatic leaps in the realms of prediction, data representation, computer vision, and much more (Goodfellow et al., 2016).

Given all this, it seems a reasonable question to ask why we relied on what can be argued as possibly predictively weaker methods such as non-linear parametric regression, Gaussian processes, and Bayesian networks. Essentially there is a very different purpose behind the methods of deep learning and what we have tried to achieve in the analyses presented here.

The core interest of social science is in being able to discover what phenomena are occurring within the systems our data come from. In a way, we are attempting to naively model parts of the actual data generating process. In doing so we can attempt to understand the systems, and better develop actions and interventions to influence them. As such we are more interested in explanatory or descriptive models, rather than predictive models which are typically the aim of deep learning. There is a distinct difference between such models, as emphasised by Shmueli (2010). Explanatory models are "statistical models for testing causal explanations", such as those we used to model social contagion and household food security. Descriptive models are "aimed at summarising or representing the data structure in a compact manner", such as our method for modelling fertility patterns. Predictive modelling, on the other hand, is "the process of applying a statistical model or data mining algorithm to data for the purpose of predicting new or future observations" which describes the most common purpose of deep learning. In this way a predictive algorithm has a strict technological purpose, whether that be telling a company what customers to aim adverts at or helping a car to drive itself, but does not concern itself with the causality involved.

This is not to say that predictive models are not useful in science. They can aid in many ways, from distinguishing patterns that can direct the formation of new hypotheses to providing more effective methods of comparison between competing explanatory models as well as perhaps even helping us pick out the individuals who could most benefit from the interventions designed via causal modelling (Shmueli, 2010). By examining versions of the system undergoing different interventions, invariance in predictions can also be used to identify causal predictors of an outcome of interest (Peters et al., 2016). However, they still rarely give results that can be deemed interpretable, which is especially true of deep learning. There is some argument as to what 'interpretable' means (Lipton, 2016). Many consider deep learning to be 'black box models' where we cannot tell what is going on inside (Lou et al., 2012), but mathematically they are not that complicated. Here we deem the lack of interpretability to be due to the fact that deep learning models take the variables to such a level of mathematical abstraction that it is impossible to discover the actual causal or structural relationships between variables from these algorithms in any human interpretable way. Some researchers are attempting to discover interpretations as to why deep learning is so powerful (Voosen, 2017), but for most applications the algorithm makes its predictions in a way unrelated to the world around us in a way most individuals can understand.

It is possible that we should focus on predictive ability over the construction of explanatory or descriptive models because of the immense difficulties in inferring causality from observational data (Breiman, 2001). Doing so, however, would present a great restriction on the research we are able to perform which is impractical given the needs of public health, medicine, and the design of public policy. As we have shown in our analyses here, we can focus on improvements in explanatory and descriptive models that can therefore lead to more worthwhile results rather than abandoning them completely.

Aside from the argument between predictive and explanatory/descriptive models, there are also further reasons for using less complicated methods than deep learning. The individuals who have to use the results of this research (including medical practitioners) are arguably more likely to trust a method that produces results in a way they can better understand or interpret themselves (Vellido et al., 2012). This is an area where Bayesian networks have the potential to thrive due to their structural interpretability, where modelling directed relationships between components of a system allows us to build a narrative for the system that users can appreciate.

Finally, there is also the argument that due to the size of our datasets we simply do not need algorithms as powerful as deep learning (Leek, 2017). Deep learning has especially come in to focus due to the rise of 'big data' (i.e. data of a great enough size or complexity that standard computational resources are not enough to process them, for instance data from internet services that are able to collect vast arrays of variables on global populations or biological data on complicated structures such as genomes). Though our datasets are large by public health standards they do not come under this definition of 'big data' so there is no need to introduce unnecessary complexity in to our modelling particularly when it risks reducing interpretabillity. Throughout these analyses we have attempted to introduce enough complexity to deal with subtle issues that arise in public health modelling, and to increase the flexibility in such modelling, but in doing so we have attempted to maintain the interpretability of the models to further aid their usefulness in public policy and intervention design.

6.3.3 Causal modelling

Modelling causality is a subject that comes with great controversy, as we saw in Chapter 3. At the same time, modelling causality is something that researchers within public health will always need to do. The core aim of public health is the design and implementation of interventions to improve the health of populations.

Interventions involve some action or combination of actions that influence the system of interest, and to be able to choose these actions in an effective way we must know how different parts of the system affect each other. We want to know for instance that if we vaccinate this portion of the population, it will definitely prevent the spread of a disease throughout the entire population.

Unfortunately, due to the ethical and resource issues involved in experiments on health aspects and behaviours of large populations, such causal modelling will often need to be performed on observational data. It is technically impossible to observe causality in observational data (Schield, 1995; Lyons, 2011). It can only be inferred and therefore we can only claim inferred causality in our results here.

By modelling causality in observational data we inevitably introduce a weakness in our analyses and results that we cannot avoid, which further demonstrates the need we have argued throughout for moving beyond simpler methods like generalised linear models. Such models can only infer associations, not causal effects. Pearl (2009) states that causal models must have assumptions built in, usually in the form of a structure that can show the directionality of relationships and more importantly the lack of effects between variables. This incorporation of assumptions and structure forms a large part of many of the methods we have used here, especially for Bayesian network models (Pearl, 2009). All of this does mean, however, that we must come to modelling with some knowledge and assumptions of what causal relationships actually exist a priori. This is a limitation we have attempted to overcome by relying on the established knowledge base contained within the literature (often developed from experimentation) to inform our assumptions.

The approach of Pearl (2009) is not the only one when trying to infer causal effects. There are many different concepts of and approaches to causality, some related and some coming from different view points, reflecting the complicated nature of inferring causal effects. Hume (1738) developed a set of necessary conditions for a relationship to be causal: namely that the cause must precede the effect in time, they must be universally associated, and spatio-temporally connected.

More mathematical approaches to causality, such as Granger causality, tried to formalise these conditions and create a practical way to determine that a relationship is causal. In Granger causality, some variable X_t (measured at time t) can be considered a cause for some effect Y_{t+1} (measured at time $t+1$) if Y_{t+1} can be better predicted with X_t than without X_t given knowledge of all other pertinent variables (Lechner, 2010). Our approach in the social contagion modelling of Chapter 3 is most similar to this, in that we rely on observed occurrences within the data to establish a causal effect. However, instead of considering the contribution to predictive ability, we established the possibility of a causal contagion effect by determining whether adding a contagion part to the transition probability of changing state improves the trade off between goodness-of-fit and model complexity measured by the information criteria.

A more practical approach favoured in epidemiology is given by the Bradford-Hill criteria, which are simply a set of nine criteria for the consideration of whether an effect may be causal (Hill, 1965). These include that the effect is strong, reproducible across multiple studies, specific in nature to e.g. a particular location or context, and coherent with experimental findings, as well as various other criteria. Similar to Granger causality, there are elements of Humean causality in that the cause proceeding the effect is yet another criteria. Unlike Granger causality, the Bradford-Hill criteria are not mathematically formalised. The more criteria the effect meets, the more justified we are in concluding it is causal. The social contagion modelling of Chapter 3 also includes at least one of these criteria, that of plausibility i.e. whether we have good reason to believe the effect is causal. This we established from experimental findings in the literature, as discussed in Section 3.1.

An alternative mathematical approach is that of potential outcomes, where the impact of different possible values of the causal variable X_t on the effect Y_{t+1} are considered counterfactually (i.e. instead of considering actual observed occurrences we consider different potential realities of what could have happened if X_t had achieved a different value) (Lechner, 2010). Alongside the use of directional structure, potential outcomes are core to the conceptualisation of causality favoured by Pearl (2009) and used in the methods of structural equations and Bayesian belief networks. We considered potential outcomes in the interventional inferences of Chapter 5 where we followed the approach of Pearl (2009) implicit in the use of Bayesian belief networks.

As mentioned before, there are also approaches that link causal modelling with predictive modelling. Peters et al. (2016) showed that the marginal distribution of some outcome of interest conditioned on a set of direct causal predictors of that outcome does not change no matter what interventions are made to variables (other than the outcome) within the system. Therefore invariances in conditional marginal distributions under different interventions within a system can help identify causal predictors of an outcome.

There are many other conceptualisations of causality established in philosophy, mathematics, and applied science. In our analyses we relied on the approaches most suitable to the needs of the data, methodology, and context of the problem. There is no one definitive approach to causal inference, so a practical choice of direction must always be made.

As a further measure to mitigate against making conclusions of causality that are too forceful, we have consistently relied on inferring probabilistic causality. In probabilistic causality we simply infer that taking some action has an effect only on the probability of some outcome, and not directly on the outcome such as in other defined forms of causality (Parascandola and Weed, 2001). This definition is not only more inclusive, but also not as strong as stating that "changing X by amount α makes Y change by an amount β".

Causal modelling is not the central point of this work, but is an important and necessary part of public health research as a whole and in more than one of these analyses we have shown how using more complex methods can improve the causal inferences performed. We cannot claim our causal inferences to be robust enough to give definite observations as there may indeed be confounding variables we were unable to develop assumptions against that would undermine the conclusions, but by using methods that build in heavier assumptions in more mechanistic and structural fashions we have been able to mitigate against confounding in a better way than generalised linear models and hypothesis testing.

6.4 Concordance with research aims

We believe the advantages that our analyses have provided have enabled us to meet the research aims outlined in Section 1.3. Our core aim was to utilise more modern and complex statistical methods that have seen use in fields such as machine learning and computer science in different analyses of socio-economic variables in public health where such methods have seen no substantial use before, in order to increase the robustness, flexibility, and utility of the findings and research performed above the limitations that have been found in standard methodology. Overall we believe this has been achieved.

In using and further developing a non-linear parametric model of social contagion, we managed to make causal inferences of the spread of mood and weight over adolescent friendship networks whilst successfully controlling for basic confounding in a way that generalised linear models cannot. In modelling social contagion as a form of complex contagion, and relying on maximum likelihood estimation and information-theoretic model selection, we achieved a more robust model comparison than that which could be done using linear contagion and statistical significance measures. This has enabled us to make robust contributions to the knowledge of what health aspects are and are not socially contagious.

In utilising Gaussian processes we were able to both increase the robustness of descriptive models of fertility patterns against being undermined by possibly wrongly assuming linear patterns, as well as open up social contagion analysis to the use of datasets that do not include social network data. The minimal assumptions made by Gaussian processes allow the data to dictate the shape of relationships the method finds, therefore overcoming the need to make unjustified a priori judgements of what these relationships should be like. The greater wealth of features built into a more complex method such as Gaussian processes, namely the use of covariance functions, can be utilised to overcome boundaries in approaching research questions using data that may not allow the use of more standard methods. This has enabled us to make much more flexible contributions to the literature on pregnancy in rural South Africa than studies that have come before.

In building Bayesian belief network models of household food security in rural South Africa, we have introduced directional structure as a key aspect of the model which is not so for the generalised linear regressions performed in past studies. This has enabled us to build models much more suited to causal modelling. On top of that, the indirect relationships modelled by the networks provide a much better reflection of the complexity inherent within the system. These models then provide us with a strong base from which to develop further models and simulate interventions in a much more effective fashion. This has enabled us to make contributions of much greater utility to the work against food insecurity within the Agincourt study area.

Overall, by looking to other areas of science that involve data analysis and have already adopted a greater range of methods, we have contributed to the library of public health methodology in a positive way. Alongside other researchers attempting to introduce much more suitable methods, we have established the importance and usefulness of these more complex methods. That stepping beyond linear models and hypothesis tests enables us to do better research and, equally importantly, do more research. Modern statistics and machine learning is home to a wealth of methods that public health researchers have often ignored in favour of standardised methodology that often cannot meet the needs of research in such a complex area. The methods we have used here are just some of those that must be taken on generally by public health researchers, and emphasises the importance of being more thoughtful about what the right methods for our research questions are.

Beyond the issue of methodology, we have been able to make good quality contributions to the pool of knowledge of three key areas of public health - social contagion amongst adolescents, adolescent pregnancy in rural South Africa, and household food security in rural South Africa. Though our focus was on the methodological improvements, we should not dismiss the end findings of each analysis (or the potential future findings of improved versions of these analyses). In the end, it is such findings that will inform public policy and help improve the health of many populations. We seek to improve the methods used in order to improve the conclusions reached, and therefore improve the actions taken by public health bodies. We believe the research performed here contributes significantly to this effort.

6.5 Future work

Our research here only covers select areas within the wider spectrum of public health research, and is certainly not the end point in using and developing more suitable methods for this research. Like all science, it is part of an iterative process that will lead to better research, more robust conclusions, and greater health implications for the general public. Probably the most key part of future work that must be done in this area is to take the methods presented here, and other such methods used in machine learning, computer science, and statistics, and apply them to further areas of public health research on socio-economic variables where they could lead to improvements over the methods that have been used before. Beyond that, there is also much we could still do in terms of further research and improvements in the individual analyses presented here.

We anticipate that future work in social contagion can further enhance the models used in Chapter 3 and section 4.4 in order to cope with a wider range of datasets and more realistically reflect the mechanisms underlying social contagion. We already presented movements in this direction in Section 3.5, where we presented various possible models of social contagion combining both higher and lower state friends to overcome the necessity of analysing these friend groups separately. The method used in Section 4.4 already presents an enhancement over that of Chapter 3 due to its ability to be applied to datasets with no social network data. However, the results from this method were a little ambiguous, so the development of a parametric method must be attempted. Possibilities in this direction were described in Section 4.4.5, including the use of similarity graphs to develop social contagion measures that could be used in a parametric model. Beyond these methodological developments, we must also take the methods used here (which already present improvements of those used in other studies) and apply them to other health behaviours such as smoking, suicide, and sex, as the proven existence of social effects could have great impact on the development of public health interventions.

Further research to extend our method used to model fertility patterns in Section 4.3 might include the following. First, to overcome the limitations mentioned, Gaussian process regression could be used for age rather than relying on a parametric model. A probabilistic classification technique such as Gaussian process classification could be used instead of regression techniques to overcome the issue of having to bin the observations together. In order to examine the apparent lack of a fertility decline in our results, and to make the research into the fertility decline more robust, non-linear modelling techniques such as Gaussian process regression could be applied to the fertility time series of the study area. In addition, other outcomes than fertility that have also been analysed using less innovative methods could be explored with this same technique. Finally, various work has shown the age-pattern of fertility to contain a secondary earlier age peak credited to premarital fertility (Garenne et al., 2000). It would also be of interest to see what happens when the parametric fertility-age model used here is replaced with a double peaked model such as that proposed by (Peristera and Kostaki, 2007).

We believe that the Bayesian belief network models of household food security we presented in Chapter 5 can form the basis of the iterative generation of a robust causal model of household food security in the Agincourt study area. As there is no established test of causal faithfulness beyond experimentation, the models can be used to inform possible interventional studies that could then be used to further develop the model and so on. In addition, it is possible that methods of testing causal faithfulness from observational data could be explored, but this comes with obvious limitations. An alternative to experimentation would be to use further expert elicitations. We did not have the resources to perform elicitations beyond the ones we did. With greater resources and more time, future work could perform more and larger elicitations, even involving a greater range of experts such as members of the Agincourt research unit. Another way to improve the causal applicability of the methods would be to develop them into dynamic Bayesian networks that include temporal relationships, which would also require further resources in terms of data and expert elicitations. Beyond the further development of these models, they can also form the basis of developing similar models of household food security for other similar populations.

6.6 The importance of collaboration

One final strength of this research that we wish to briefly discuss is the collaborative nature of it. Though this is the work of the author it was achieved via collaboration and discussion with scientists of many different backgrounds. Commonly though these scientists either came from a social science or mathematical background.

The development of the methods used in these analyses could only be achieved out of collaboration, and the development of further models will also require collaboration. Mathematicians bring a greater range of methodology such as those we have used here. As discussed in Section 1.2, textbooks in statistical methodology aimed at social scientists concern themselves only with the standard methodology of GLMs, hypothesis testing, and summary statistics. The work of a social scientist is context heavy such that they may have little time to go beyond these standard methodologies. However, social scientists are still essential in this work for framing the problem, developing the assumptions needed especially in causal modelling, and bridging the knowledge gap between experts and non-experts (Lowe et al., 2013). Given this collaboration is essential for the improvement of methodology in public health analysis.

There are also further benefits to collaboration. It has been shown that paper authorships crossing a range of disciplines correlates with more citations and favourable responses from peer reviewers (Franceschet and Costantini, 2010). It has also been argued that mutually beneficial relationships can develop between different fields such as mathematics and biology where mathematical methodology can improve biological research and biological findings can spur the development of new mathematics (Cohen, 2004). Such a relationship could also exist between social science and statistics.

There are many barriers to collaboration (Fischer et al., 2011). It has been argued that due to these difficulties collaboration for its own sake is not worthwhile, but is extremely worthwhile when it can provide new possibilities (Sonnenwald, 2007). This research (alongside other similar research attempting to bring new methodology to public health) has shown how such collaboration can be fruitful due to the new possibilities in terms of increased robustness, flexibility, and usefulness they present.

CHAPTER 7

Summary

The modelling of socio-economic variables in public health is integral to the design of interventions related to social determinants of health in populations across the world. Such research has now become a central part of the aims of the World Health Organisation in order to address the health inequalities and inequities that exist in all countries. Due to the expense and ethical difficulties of experimenting on socio-economic variables, statistical inference from observational data is key to this research. However, these inferences are commonly done using a small library of conventional, but simplistic, methods such as hypothesis testing and generalised linear regression. Moving beyond these methods is essential to improving statistical inference in public health and addressing the criticisms many established studies have drawn. We addressed this issue in three key areas: social contagion in US adolescents, adolescent pregnancy in rural South Africa, and household food security in rural South Africa.

In Chapter 3 we used a non-linear parametric model to analyse the possibility of social contagion of mood and weight across the friendship networks of US adolescents using the Add Health dataset. We further generalised an already established model of social contagion where transition probabilities of possibly contagious behaviour states are stratified by the number of friends of the contagious state. By examining stratified transition probabilities, we were able to correctly distinguish contagion from other possible confounding phenomena. By modelling the contagion as a complex contagion, and using maximum likelihood and information-theoretic statistical techniques, we improved on similar studies that relied on linear contagion and statistical significance tests. By generalising from binary behaviour states to ranges of behaviour states we were able to capture greater subtleties of the contagion phenomena occurring. As such, we were able to find evidence of one-sided contagion of increasing weight, supporting the possibility of contagion of obesity. We were also able to help reconcile our previous findings of contagion of mood with the lack of contagion of depression found in previous studies using a simple Gaussian process regression model. Finally, we presented several possibilities of generalising the model further. Overall we were able to increase the robustness of causal inferences made about social contagion.

In Chapter 4 we used Gaussian processes to address two separate problems in modelling pregnancy in rural South Africa using the Agincourt HDSS dataset. The first involved building descriptive models of fertility-patterns over socio-economic variables such as age and socio-economic status, where established models exist for age but not for any other variable. By using Gaussian process regression combined with parametric regression over age, we were able to capture the pre-defined fertility age-pattern whilst also being able to infer fertility-patterns over other variables without making unjustified a priori assumptions of linearity. Gaussian process regression makes minimal assumptions, allowing the data to govern the shape of the relationship found and allowing for the possible findings of non-linearities. Though the fertility-pattern over socio-economic status was found to be constant, this does not diminish the importance of not assuming linearity when we have no reason to do so. We also analysed differences here between refugee and non-refugee subsets of the population, but found no substantial deviations of interest.

The second part of Chapter 4 involved taking advantage of the more unique features of Gaussian process classification, namely the covariance function, to perform an analysis of the possibility of social contagion of pregnancy risk behaviour on the Agincourt HDSS dataset that includes no social network data. This was achieved by defining a contagion variable derived from the covariance function fitted to a non-contagion model from the previous time step. In doing so, we were able to address a research question we could not have done otherwise without resorting to the much more difficult task of justifying a first principles parametric model of a non-social network contagion measure. We found no evidence for social contagion of pregnancy risk behaviour, though admittedly the results were somewhat ambiguous. As such, further research is needed, but we have still established the flexibility relying on more complex methods gives us.

In Chapter 5 we built three possible causal models of household food security in the Agincourt study area using the method of Bayesian belief networks. One learned from expert elicitation using the expert local knowledge of members of the Agincourt Community Advisory Group. The second learned from the data using a contraint-based structure learning algorithm. The third learned from a combination of the experts and the data by using the expert network as a prior for a hill climbing greedy search algorithm across the possible network structures. The inbuilt features of directional structure, modelling of indirect relationships, and conditional independence make Bayesian belief networks much more suitable as causal models we can make causal inferences from rather than any typical linear model or statistical test. Testing the causal faithfulness to the system was not possible, but various comparisons were performed including simulating a small set of established interventions on the networks. These implied that the data-expert combined structure performs the best. These networks can be used for more effective simulations of possible interventions, as well as informing the building of similar network models for other populations.

The social population level systems considered in public health have been long established as complex systems, therefore rendering simple dichotomous or linear models insufficient for many inference tasks on these systems. Though the methods we have used here include many limitations, not least the difficulties in establishing causal models from observational data, they achieve great improvements over previous studies in the key areas of robustness, flexibility, and utility. Not only have we contributed key findings to the contextual areas we considered, we have also brought methods in from areas such as machine learning and computer science to help grow the library of statistical methods in public health research. In doing so we hope to encourage more thoughtful analyses with more suitable and stronger methodology. We also hope to further encourage collaboration between social scientists and mathematical scientists, continuing the tradition already established by many, to ensure the quality of public health research in socio-economic variables to come.

Bibliography

Donald Acheson, D Barker, et al. Independent inquiry into inequalities in health: report, 1998.

Charu C Aggarwal. *Data mining: the textbook*. Springer, 2015.

Hirotugu Akaike. A new look at the statistical model identification. *IEEE Trans Autom Control*, 19(6):716–723, 1974.

Mir M Ali, Aliaksandr Amialchuk, Song Gao, and Frank Heiland. Adolescent weight gain and social networks: is there a contagion effect? *Appl Econ*, 44(23):2969–2983, 2012.

Douglas G Altman. Statistics in medical journals: some recent trends. *Statistics in Medicine*, 19(23):3275–3289, 2000.

AS Anater, M Chambard, M Barnes, L Morgan, K Woodward, S Beaulieu, et al. Current and prospective scope of hunger and food security in america: A review of current research. *Prepared for the president and congress; North Carolina: RTI International*, 2014.

Sinan Aral, Lev Muchnik, and Arun Sundararajan. Distinguishing influence-based contagion from homophily-driven diffusion in dynamic networks. *Proc Natl Acad Sci*, 106(51):21544–21549, 2009.

Peter Armitage, Geoffrey Berry, and John Nigel Scott Matthews. *Statistical methods in medical research*. John Wiley & Sons, 2008.

Samuelina Arthur, Martin Bangha, and Osman Sankoh. Review of contributions from HDSSs to research in sexual and reproductive health in low-and middle-income countries. *Trop Med Int Health*, 18(12):1463–1487, 2013.

Dawit Getnet Ayele. Determinants of fertility in ethiopia. *Afr Health Sci*, 15(2): 546–551, 2015.

Nicoletta Balbo and Nicola Barban. Does fertility behavior spread among friends? *Am Soc Rev*, 79(3):412–431, 2014.

Yaneer Bar-Yam. Improving the effectiveness of health care and public health: a multiscale complex systems analysis. *American Journal of Public Health*, 96(3): 459–466, 2006.

Elizabeth M Barbeau, Nancy Krieger, and Mah-Jabeen Soobader. Working class matters: socioeconomic disadvantage, race/ethnicity, gender, and smoking in nhis 2000. *American journal of public health*, 94(2):269–278, 2004.

MJ Barons, X Zhong, and JQ Smith. Dynamic bayesian networks for decision support and sugar food security. *CRiSM Rep. (submitted)*, 2014.

Mel Bartley and Ian Plewis. Accumulated labour market disadvantage and limiting long-term illness: data from the 1971–1991 office for national statistics' longitudinal study. *Int J Epidemiology*, 31(2):336–341, 2002.

Laura Bernardi. Channels of social influence on reproduction. *Population Research and Policy Review*, 22(5-6):427–555, 2003.

Laura Bernardi, Sylvia Keim, and Holger Von Der Lippe. Social influences on fertility: A comparative mixed methods study in eastern and western germany. *Journal of Mixed Methods Research*, 1(1):23–47, 2007.

Eszter A Bertha and Judit Balázs. Subthreshold depression in adolescence: a systematic review. *European Child & Adolescent Psychiatry*, 22(10):589–603, 2013.

J Martin Bland and Douglas G Altman. Multiple significance tests: the bonferroni method. *BMJ*, 310(6973):170, 1995.

Sara N Bleich, Marian P Jarlenski, Caryn N Bell, and Thomas A LaVeist. Health inequalities: trends, progress, and policy. *Ann Rev Pub Health*, 33:7–40, 2012.

Remco Ronaldus Bouckaert. *Bayesian belief networks: from construction to inference.* PhD thesis, Utrecht University, 1995.

SC Brailsford, Paul Robert Harper, B Patel, and M Pitt. An analysis of the academic literature on simulation and modelling in health care. *Journal of Simulation*, 3 (3):130–140, 2009.

Paula Braveman, Susan Egerter, and David R Williams. The social determinants of health: coming of age. *Ann Rev Pub Health*, 32:381–398, 2011.

Leo Breiman. Statistical modeling: The two cultures. *Statistical science*, 16(3): 199–231, 2001.

Glenn W Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

Christoph Bühler and Ewa Fratczak. Learning from others and receiving support: The impact of personal networks on fertility intentions in poland. *European Societies*, 9(3):359–382, 2007.

Rulof P Burger, Ronelle Burger, and Laura Rossouw. The fertility transition in south africa: A retrospective panel data analysis. *Development Southern Africa*, 29(5):738–755, 2012.

Mary A Burke and Frank Heiland. Social dynamics of obesity. *Economic Inquiry*, 45(3):571–591, 2007.

Victor N Bushamuka, Saskia de Pee, Aminuzzaman Talukder, Lynnda Kiess, Dora Panagides, Abu Taher, and Martin Bloem. Impact of a homestead gardening program on household food security and empowerment of women in bangladesh. *Food and Nutrition Bulletin*, 26(1):17–25, 2005.

Carol S Camlin, Michel Garenne, and Tom A Moultrie. Fertility trend and pattern in a rural area of south africa in the context of hiv/aids. *Afr J Reprod Health*, 8 (2):38–54, 2004.

Neil C Campbell, Elizabeth Murray, Janet Darbyshire, Jon Emery, Andrew Farmer, Frances Griffiths, Bruce Guthrie, Helen Lester, Phil Wilson, and Ann Louise Kinmonth. Designing and evaluating complex interventions to improve health care. *BMJ*, 334(7591):455, 2007.

Center for Disease Control. Ten great public health achievements–united states, 1900-1999. *MMWR*, 48(12):241, 1999.

Center for Disease Control. Percentile data files with lms values. http://www.cdc.gov/growthcharts/percentile_data_files.htm, 2009. Accessed: 6 February 2015.

Center for Disease Control. Ten great public health achievements–united states, 2001-2010. *MMWR*, 60(19):619, 2011a.

Center for Disease Control. Ten great public health achievements–worldwide, 2001–2010. *MMWR*, 60(24):814–818, 2011b.

Center for Disease Control. About bmi for children and teens. http://www.cdc.gov/healthyweight/assessing/bmi/childrens_bmi/about_childrens_bmi.html, 2014. Accessed: 18 August 2015.

Center for Disease Control. What is Public Health? https://www.cdcfoundation.org/what-public-health, 2017. Accessed: 14 September 2017.

Damon Centola. The spread of behavior in an online social network experiment. *Science*, 329(5996):1194–1197, 2010.

Damon Centola and Michael Macy. Complex contagions and the weakness of long ties. *Am J Sociol*, 113(3):702–734, 2007.

Mary W Chaffee and Margaret M McNeill. A model of nursing as a complex adaptive system. *Nursing Outlook*, 55(5):232–241, 2007.

Nicholas A Christakis and James H Fowler. The spread of obesity in a large social network over 32 years. *N Engl J Med*, 357(4):370–379, 2007.

Nicholas A Christakis and James H Fowler. The collective dynamics of smoking in a large social network. *N Engl J Med*, 358(21):2249–2258, 2008.

Nicholas A Christakis and James H Fowler. Social contagion theory: examining dynamic social networks and human behavior. *Stat Med*, 32(4):556–577, 2013.

Nicola J Christofides, Rachel K Jewkes, Kristin L Dunkle, Mzikazi Nduna, Nwabisa Jama Shai, and Claire Sterk. Early adolescent pregnancy increases risk of incident hiv infection in the eastern cape, south africa: a longitudinal study. *J Int AIDS Soc*, 17(1), 2014.

David Clayton and Michael Hills. *Statistical models in epidemiology.* OUP Oxford, 2013.

Joel E Cohen. Mathematics is biology's next microscope, only better; biology is mathematics' next physics, only better. *PLoS biology*, 2(12):e439, 2004.

Ethan Cohen-Cole and Jason M Fletcher. Detecting implausible social network effects in acne, height, and headaches: longitudinal analysis. *BMJ*, 337:a2533, 2008a.

Ethan Cohen-Cole and Jason M Fletcher. Is obesity contagious? social networks vs. environmental factors in the obesity epidemic. *Journal of health economics*, 27(5):1382–1387, 2008b.

Timothy J Cole and Pamela J Green. Smoothing reference centile curves: the lms method and penalized likelihood. *Stat Med*, 11(10):1305–1319, 1992.

Sarah Coll-Black, Daniel O Gilligan, John Hoddinott, Neha Kumar, and William Wiseman. Targeting food security interventions in ethiopia: The productive safety net. *Food and Agriculture in Ethiopia: Progress and Policy Challenges*, 74:280, 2012.

Mark A Collinson, Samuel J Clark, Annette M Gerritsen, Peter Byass, Kathleen Kahn, and Stephen Tollman. The dynamics of poverty and migration in a rural south african community, 2001-2005. *CSSS Working Paper Series*, (92):1–38, 2009.

Amy Colquhoun, Janis Geary, and Karen J Goodman. Challenges in conducting community-driven research created by differing ways of talking and thinking about science: a researcher's perspective. *International journal of circumpolar health*, 72(1):21232, 2013.

David Colquhoun. An investigation of the false discovery rate and the misinterpretation of p-values. *R. Soc. open sci.*, 1(3):140216, 2014.

Nancy R Cook. Use and misuse of the receiver operating characteristic curve in risk prediction. *Circulation*, 115(7):928–935, 2007.

John Coveney. A qualitative study exploring socio-economic differences in parental lay knowledge of food and health: implications for public health nutrition. *Public Health Nutrition*, 8(3):290–297, 2005.

Peter Craig, Paul Dieppe, Sally Macintyre, Susan Michie, Irwin Nazareth, and Mark Petticrew. Developing and evaluating complex interventions: the new medical research council guidance. *BMJ*, 337:a1655, 2008.

William H Crown. Potential application of machine learning in health outcomes research and some statistical cautions. *Value in Health*, 18(2):137–140, 2015.

Tegan Cruwys, S Alexander Haslam, Genevieve A Dingle, Catherine Haslam, and Jolanda Jetten. Depression and social identity: An integrative review. *Personality and Social Psychology Review*, 18(3):215–238, 2014.

Geoff Cumming. The new statistics: Why and how. *Psychological Science*, 25(1): 7–29, 2014.

David M Cutler and Adriana Lleras-Muney. Education and health: evaluating theories and evidence. Technical report, National Bureau of Economic Research, 2006.

Jayati Das-Munshi, David Goldberg, Paul E Bebbington, Dinesh K Bhugra, Traolach S Brugha, Michael E Dewey, Rachel Jenkins, Rob Stewart, and Martin Prince. Public health significance of mixed anxiety and depression: beyond current classification. *Br J Psychiatry*, 192(3):171–177, 2008.

Luis M de Campos, Juan M Fernández-Luna, and Juan F Huete. Bayesian networks and information retrieval: an introduction to the special issue. *Information Processing & Management*, 40(5):727–733, 2004.

Alfred Dean and Walter M Ensel. Modelling social support, life events, competence, and depression in the context of age and sex. *J Community Psychol*, 10(4):392–408, 1982.

K Didan. Mod13q1 modis/terra vegetation indices 16-day l3 global 250m sin grid v006. `https://doi.org/10.5067/modis/mod13q1.006`, 2015. Accessed: 31 January 2017.

Delali BK Dovie, Charlie M Shackleton, and ETF Witkowski. Conceptualizing the human use of wild edible herbs for conservation in south african communal areas. *Journal of Environmental Management*, 84(2):146–156, 2007.

Valerie J Ehlers. Adolescent mothers utilization of contraceptive services in south africa. *Int Nursing Rev*, 50(4):229–241, 2003.

Beverley Ellis. Complexity in practice: understanding primary care as a complex adaptive system. *Journal of Innovation in Health Informatics*, 18(2):135–140, 2010.

David M Erceg-Hurn and Vikki M Mirosevich. Modern robust statistical methods: an easy way to maximize the accuracy and power of your research. *American Psychologist*, 63(7):591, 2008.

Fabrice Etilé. Social norms, ideal body weight and food attitudes. *Health Economics*, 16(9):945–966, 2007.

Robert W Eyre. Social networks and health. Master's thesis, University of Warwick, United Kingdom, 2014.

Mieke Faber, Chantell Witten, and Scott Drimie. Community-based agricultural interventions in the context of food and nutrition security in south africa. *South African Journal of Clinical Nutrition*, 24(1):21–30, 2011.

Alessio Farcomeni and Laura Ventura. An overview of robust methods in medical research. *Statistical Methods in Medical Research*, 21(2):111–133, 2012.

Arnout RH Fischer, Hilde Tobi, and Amber Ronteltap. When natural met social: a review of collaboration between the natural and social sciences. *Interdisciplinary Science Reviews*, 36(4):341–358, 2011.

Food and Agriculture Organization of the United Nations. Food security: Policy brief. 2006. URL `http://www.fao.org/3/a-a0800e.pdf`.

Edward Fottrell, Peter Byass, and Yemane Berhane. Demonstrating the robustness of population surveillance data: implications of error rates on demographic and mortality estimates. *BMC Med Res Methodol*, 8(1):1, 2008.

The Health Foundation. Complex adaptive systems. evidence scan. 2010.

James H Fowler and Nicholas A Christakis. Dynamic spread of happiness in a large social network: longitudinal analysis over 20 years in the framingham heart study. *BMJ*, 337:a2338, 2008.

Massimo Franceschet and Antonio Costantini. The effect of scholar collaboration on impact and quality of academic papers. *Journal of informetrics*, 4(4):540–553, 2010.

Sarah W Fraser and Trisha Greenhalgh. Complexity science: Coping with complexity: Educating for capability. *BMJ*, 323(7316):799, 2001.

Nir Friedman, Michal Linial, Iftach Nachman, and Dana Pe'er. Using bayesian networks to analyze expression data. *J Comp Bio*, 7(3-4):601–620, 2000.

Cheryl D Fryar, Margaret D Carroll, Cynthia L Ogden, et al. Prevalence of obesity among children and adolescents: United states, trends 1963–1965 through 2009–2010. *National Center for Health Statistics*, 1960, 2012.

Santosh Gaihre, Janet Kyle, Sean Semple, Jo Smith, Madhu Subedi, and Debbi Marais. Type and extent of trans-disciplinary co-operation to improve food security, health and household environment in low and middle income countries: systematic review. *BMC Public Health*, 16(1):1093, 2016.

Dilrukshi Hashini Galhena, Russell Freed, and Karim M Maredia. Home gardens: a promising approach to enhance household food security and wellbeing. *Agriculture & Food Security*, 2(1):8, 2013.

Michel Garenne and Veronique Joseph. The timing of the fertility transition in sub-saharan africa. *World Development*, 30(10):1835–1843, 2002.

Michel Garenne and Julien Zwang. Premarital fertility in namibia: trends, factors and consequences. *J Biosoc Sci*, 38(2):145–167, 2006.

Michel Garenne, Stephen Tollman, and Kathleen Kahn. Premarital fertility in rural south africa: a challenge to existing population policy. *Stud Fam Plann*, 31(1): 47–54, 2000.

Michel L Garenne, Stephen M Tollman, Mark A Collinson, and Kathleen Kahn. Fertility trends and net reproduction in agincourt, rural south africa, 19922004. *Scand J Public Health*, 35(69 suppl):68–76, 2007.

Andrew Gelman and David Weakliem. Of beauty, sex and power: Too little attention has been paid to the statistical challenges in estimating small effects. *American Scientist*, 97(4):310–316, 2009.

Andrew Gelman, John B Carlin, Hal S Stern, David B Dunson, Aki Vehtari, and Donald B Rubin. *Bayesian data analysis*. CRC press Boca Raton, FL, 2014.

Clara Goldberg-Freeman, Nancy Kass, Andrea Gielen, Patricia Tracey, Barbara Bates-Hopkins, and Mark Farfel. Faculty beliefs, perceptions, and level of community involvement in their research: a survey at one urban academic institution. *Journal of Empirical Research on Human Research Ethics*, 5(4):65–76, 2010.

Chris Goldspink. Modelling social systems as complex: Towards a social simulation meta-model. *Journal of Artificial Societies and Social Simulation*, 3(2):1–23, 2000.

Irving John Good. Rational decisions. *J Roy Stat Soc B*, 14(1):107–114, 1952.

Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

Steven N Goodman. Toward evidence-based medical statistics. 1: The p value fallacy. *Annals of Internal Medicine*, 130(12):995–1004, 1999.

Sander Greenland, Stephen J Senn, Kenneth J Rothman, John B Carlin, Charles Poole, Steven N Goodman, and Douglas G Altman. Statistical tests, p values, confidence intervals, and power: a guide to misinterpretations. *European Journal of Epidemiology*, 31(4):337–350, 2016.

K. Harris et al. The national longitudinal study of adolescent to adult health: Research design. `http://www.cpc.unc.edu/projects/addhealth/design`, 2008. Accessed: 27 July 2015.

Penelope Hawe, Alan Shiell, and Therese Riley. Complex interventions: how out of control can a randomised controlled trial be? *BMJ*, 328(7455):1561, 2004.

Matthew J Hayat, Amanda Powell, Tessa Johnson, and Betsy L Cadwell. Statistical methods used in the public health literature and implications for training of public health professionals. *PloS one*, 12(6):e0179032, 2017.

Helen Keller International. Homestead food production model contributes to improved household food security, nutrition and female empowerment–experience from scaling-up programs in asia (bangladesh, cambodia, nepal and philippines). *Nutrition Bulletin*, 8(1), 2010.

Alison L. Hill, David G. Rand, Martin A. Nowak, and Nicholas A. Christakis. Emotions as infectious diseases in a large social network: the SISa model. *Proc R Soc B*, 277(1701):3827–3835, 2010a.

Alison L Hill, David G Rand, Martin A Nowak, and Nicholas A Christakis. Infectious disease modeling of social contagion in networks. *PLoS Comput Biol*, 6(11): e1000968, 2010b.

Austin Bradford Hill. The environment and disease: Association or causation? *Proceedings of the Royal Society of Medicine*, 58(5):295–300, 1965.

E. M. Hill, F. E. Griffiths, and T. House. Spreading of healthy mood in adolescent social networks. *Proc R Soc B*, 282(1813), 2015. ISSN 0962-8452. doi: 10.1098/rspb.2015.1180.

Christer Hogstedt. *The Swedish public health policy and the National Institute of Public Health*, volume 64. Taylor & Francis, 2004.

Søren Højsgaard. Graphical independence networks with the grain package for r. *Journal of Statistical Software*, 46(10):1–26, 2012.

B Holmes. The influence of food security and other social and environmental factors on diet in the national low income diet and nutrition survey. *The Proceedings of the Nutrition Society*, 67(OCE):E88, 2008.

Lucas R Hope and Kevin B Korb. A bayesian metric for evaluating machine learning algorithms. In *Australian Conference on Artificial Intelligence*, pages 991–997. Springer, 2004.

Nicholas J Horton and Suzanne S Switzer. Statistical methods in the journal. *New England Journal of Medicine*, 353(18):1977–1979, 2005.

David W Hosmer and Stanley Lemeshow. *Applied logistic regression*. 2001.

Brian Houle, Samuel J Clark, F Xavier Gómez-Olivé, Kathleen Kahn, and Stephen M Tollman. The unfolding counter-transition in rural south africa: mortality and cause of death, 1994–2009. *PLoS One*, 9(6):e100420, 2014.

David Hume. *A Treatise of Human Nature*. Oxford University Press, 1738.

Lori M Hunter, Wayne Twine, and Laura Patterson. "locusts are now our beef": Adult mortality and household dietary use of local environmental resources in rural south africa. *Scandinavian Journal of Public Health*, 35(69_suppl):165–174, 2007.

John PA Ioannidis. Why most published research findings are false. *PLoS Medicine*, 2(8):e124, 2005.

Barbara A Israel, Amy J Schulz, Edith A Parker, and Adam B Becker. Review of community-based research: assessing partnership approaches to improve public health. *Annual review of public health*, 19(1):173–202, 1998.

Rachel Jewkes, Caesar Vundule, Fidelia Maforah, and Esme Jordaan. Relationship dynamics and teenage pregnancy in south africa. *Social Science & Medicine*, 52 (5):733–744, 2001.

Xia Jiang and Gregory F Cooper. A bayesian spatio-temporal method for disease outbreak detection. *Journal of the American Medical Informatics Association*, 17 (4):462–471, 2010.

F Johnson, L Cooke, H Croker, and Jane Wardle. Changing perceptions of weight in great britain: comparison of two population surveys. *BMJ*, 337:a494, 2008.

Aimable Robert Jonckheere. A distribution-free k-sample test against ordered alternatives. *Biometrika*, 41(1/2):133–145, 1954.

Kathleen Kahn, Stephen M Tollman, Mark A Collinson, Samuel J Clark, Rhian Twine, Benjamin D Clark, Mildred Shabangu, Francesc Xavier Gómez-Olivé, Obed Mokoena, and Michel L Garenne. Research into health, population and social transitions in rural south africa: Data and methods of the agincourt health and demographic surveillance system. *Scand J Public Health*, 35(69 suppl):8–20, 2007.

Kathleen Kahn, Mark A Collinson, F Xavier Gómez-Olivé, Obed Mokoena, Rhian Twine, Paul Mee, Sulaimon A Afolabi, Benjamin D Clark, Chodziwadziwa W Kabudula, Audrey Khosa, SImon Khoza, Mildred G Shabangu, Bernard Silaule, Jeffrey B Tibane, Ryan G Wagner, Michel L Garenne, Samuel J Clark, and Stephen M Tollman. Profile: Agincourt health and socio-demographic surveillance system. *Int J Epidemiol*, 41:988–1001, 2012.

Sarah Kaschula. Using people to cope with the hunger: social networks and food transfers amongst hiv/aids afflicted households in kwazulu-natal, south africa. *AIDS and Behavior*, 15(7):1490, 2011.

Carol E Kaufman, Thea Wet, and Jonathan Stadler. Adolescent pregnancy and parenthood in south africa. *Studies in family planning*, 32(2):147–160, 2001.

Sylvia Keim, Andreas Klärner, and Laura Bernardi. Qualifying social influence on fertility intentions: composition, structure and meaning of fertility-relevant social networks in western germany. *Current Sociology*, 57(6):888–907, 2009.

Shane Khan and Vinod Mishra. Youth reproductive and sexual health. *DHS Comparative Reports 19, USAID*, 2008.

Dudley Kirk and Bernard Pillet. Fertility levels, trends, and differentials in subsaharan africa in the 1980s and 1990s. *Stud Fam Plann*, 29(1):1–22, 1998.

Tracy M Kirkland, Robert J Kemp, Lori M Hunter, and Wayne M Twine. Toward improved understanding of food security: a methodological examination based in rural south africa. *Food, Culture & Society*, 16(1):65–84, 2013.

Betty R Kirkwood and Jonathan AC Sterne. *Essential medical statistics.* John Wiley & Sons, 2010.

Daniel N Klein, Catherine R Glenn, Derek B Kosty, John R Seeley, Paul Rohde, and Peter M Lewinsohn. Predictors of first lifetime onset of major depressive disorder in young adulthood. *J Abnormal Psych*, 122(1):1, 2013.

Daphne Koller and Nir Friedman. *Probabilistic graphical models: principles and techniques*. MIT press, 2009.

Igor Kononenko. Machine learning for medical diagnosis: history, state of the art and perspective. *Artificial Intelligence in Medicine*, 23(1):89–109, 2001.

Kevin B Korb and Ann E Nicholson. *Bayesian artificial intelligence*. Florida: Chapman & Hall/CRC, 2003.

Christos L Koumenides and Nigel R Shadbolt. Combining link and content-based information in a bayesian inference model for entity search. In *Proceedings of the 1st Joint International Workshop on Entity-Oriented and Semantic Search*, page 3. ACM, 2012.

Adam DI Kramer, Jamie E Guillory, and Jeffrey T Hancock. Experimental evidence of massive-scale emotional contagion through social networks. *Proc Natl Acad Sci*, 111(24):8788–8790, 2014.

Øystein Kravdal. Education and fertility in sub-saharan africa: Individual and community effects. *Demography*, 39(2):233–250, 2002.

Demetre Labadarios, Zandile June-Rose Mchiza, Nelia Patricia Steyn, Gerda Gericke, Eleni Maria Winifred Maunder, Yul Derek Davids, and Whadi-ah Parker. Food security in south africa: a review of national surveys. *Bulletin of the World Health Organization*, 89(12):891–899, 2011.

Steffen L Lauritzen and David J Spiegelhalter. Local computations with probabilities on graphical structures and their application to expert systems. *J Roy Stat Soc B*, pages 157–224, 1988.

Michael Lechner. The relation of different concepts of causality used in time series and microeconometrics. *Econometric Reviews*, 30(1):109–127, 2010.

Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521 (7553):436–444, 2015.

Jeff Leek. Don't use deep learning your data isn't that big. `https://simplystatistics.org/2017/05/31/deeplearning-vs-leekasso/`, 2017. Accessed: 29 September 2017.

Jeffrey T Leek and Roger D Peng. Statistics: P values are just the tip of the iceberg. *Nature*, 520(7549):612, 2015.

Stefanie Lemke, Hester H Vorster, NS Jansen van Rensburg, and Joachim Ziche. Empowered women, social networks and the contribution of qualitative research: broadening our understanding of underlying causes for food and nutrition insecurity. *Public Health Nutrition*, 6(8):759–764, 2003.

Paul S Levy and Kristine Stolte. Statistical methods in public health and epidemiology: a look at the recent past and projections for the next decade. *Statistical Methods in Medical Research*, 9(1):41–55, 2000.

GH Leyna, KS Mnyika, EJ Mmbaga, A Hussain, E Klouman, C Holm-Hansen, and KI Klepp. Food insufficiency in rural kilimanjaro, tanzania. *East African Medical Journal*, 84(4):163, 2007.

Zachary C Lipton. The mythos of model interpretability. *arXiv preprint arXiv:1606.03490*, 2016.

Daniel Lois and Oliver Arránz Becker. Is fertility contagious? using panel data to disentangle mechanisms of social network influences on fertility decisions. *Advances in Life Course Research*, 21:123–134, 2014.

Yin Lou, Rich Caruana, and Johannes Gehrke. Intelligible models for classification and regression. In *Proceedings of the 18th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 150–158. ACM, 2012.

Tom Love and Chris Burton. General practice as a complex system: a novel analysis of consultation data. *Family Practice*, 22(3):347–352, 2005.

Philip Lowe, Jeremy Phillipson, and Katy Wilkinson. Why social scientists should engage with natural scientists. *Contemporary Social Science*, 8(3):207–222, 2013.

Russell Lyons. The spread of evidence-poor medicine via flawed social-network analysis. *Stat Politics Policy*, 2(1), 2011.

Majola L Mabuza, Gerald F Ortmann, Edilegnaw Wale, and Munyaradzi J Mutenje. The effect of major income sources on rural household food (in)security: Evidence from swaziland and implications for policy. *Ecology of Food and Nutrition*, 55(2): 209–230, 2016.

Michael Marmot. Social determinants of health inequalities. *The Lancet*, 365(9464): 1099–1104, 2005.

GFH McLeod, LJ Horwood, and DM Fergusson. Adolescent depression, adult mental health and psychosocial outcomes at 30 and 35 years. *Psychological Medicine*, 46(7):1401, 2016.

Nicola Mead, Helen Lester, Carolyn Chew-Graham, Linda Gask, and Peter Bower. Effects of befriending on depressive symptoms and distress: systematic review and meta-analysis. *Br J Psychiatry*, 196(2):96–101, 2010.

Nolwazi Mkhwanazi. Understanding teenage pregnancy in a post-apartheid south african township. *Culture, Health, & Sexuality*, 12(4):347–358, 2010.

Mark R Montgomery and John B Casterline. Social learning, social influence, and new models of fertility. *Population and Development Review*, 22:151–175, 1996.

Tom A Moultrie and Ian M Timaeus. Trends in south african fertility between 1970 and 1998. *Cape Town: South African Medical Research Council*, 2002.

Tom A Moultrie and Ian M Timæus. The south african fertility decline: Evidence from two censuses and a demographic and health survey. *Popul Stud*, 57(3): 265–283, 2003.

Marcus Müllner, Hugh Matthews, and Douglas G Altman. Reporting on statistical methods to adjust for confounding: a cross-sectional survey. *Annals of Internal Medicine*, 136(2):122–126, 2002.

Barbara Hazard Munro. *Statistical methods for health care research*, volume 1. Lippincott Williams & Wilkins, 2005.

Kevin P Murphy. *Machine learning: a probabilistic perspective.* MIT press, 2012.

A Must and SE Anderson. Body mass index in children and adolescents: considerations for population-based applications. *Int J Obes*, 30:590–594, 2006.

Aviva Must, Jennifer Spadano, Eugenie H Coakley, Alison E Field, Graham Colditz, and William H Dietz. The disease burden associated with overweight and obesity. *Jama*, 282(16):1523–1529, 1999.

Florence K MKaibi, Nelia P Steyn, Sophie Ochola, and Lisanne Du Plessis. Effects of agricultural biodiversity and seasonal rain on dietary adequacy and household food security in rural areas of kenya. *BMC Public Health*, 15(1):422, 2015.

Raphael J Nawrotzki, Kristin Robson, Margaret J Gutilla, Lori M Hunter, Wayne Twine, and Petra Norlund. Exploring the impact of the 2008 global food crisis on food security among vulnerable households in rural south africa. *Food security*, 6 (2):283–297, 2014.

Carin Nilses, Gunilla Lindmark, Stephen Munjanja, and Lennarth Nyström. Trends in fertility patterns of women in rural zimbabwe. *Health Care Women Int*, 18(4): 369–382, 1997.

Hiroshi Nitta, Shin Yamazaki, Takashi Omori, and Tosiya Sato. An introduction to epidemiologic and statistical methods useful in environmental epidemiology. *Journal of Epidemiology*, 20(3):177–184, 2010.

Hans Noel and Brendan Nyhan. The unfriending problem: The consequences of homophily in friendship retention for causal estimates of social influence. *Social Networks*, 33(3):211–218, 2011.

Vincent Nsabuwera, Bethany Hedt-Gauthier, Mohammed Khogali, Mary Edginton, Sven G Hinderaker, Marie Paul Nisingizwe, Jean de Dieu Tihabyona, Benoit Sikubwabo, Samuel Sembagare, Antoinette Habinshuti, and Peter Drobac. Making progress towards food security: evidence from an intervention in three rural districts of rwanda. *Public Health Nutrition*, 19(07):1296–1304, 2016.

Martin Palamuleni, Ishmael Kalule-Sabiti, and Monde Makiwane. Fertility and childbearing in south africa. In *Families and households in post-apartheid South Africa: Socio-demographic perspectives*, pages 113–134. 2007.

ME Palamuleni and AS Adebowale. Patterns of premarital childbearing among unmarried female youths in sub-saharan africa: Evidence from demographic health survey. *Scientific Research and Essays*, 9(10):421–430, 2014.

Saadhna Panday, M Makiwane, C Ranchod, and T Letsoala. Teenage pregnancy in south africa: with a specific focus on school-going learners. *Human Sciences Research Council*, 2009.

Mark Parascandola and Douglas L Weed. Causation in epidemiology. *Journal of Epidemiology & Community Health*, 55(12):905–912, 2001.

Judea Pearl. Causal inference in statistics: An overview. *Statistics Surveys*, 3: 96–146, 2009.

Judea Pearl. Graphical models for probabilistic and causal reasoning. In *Computing Handbook, Third Edition: Computer Science and Software Engineering, Volume I.*, pages 1–29. 2013.

Jennifer Peat and Belinda Barton. *Medical statistics: A guide to data analysis and critical appraisal*. John Wiley & Sons, 2008.

Alberto Pellegrini Filho. Public policy and the social determinants of health: the challenge of the production and use of scientific evidence. *Cadernos de Saude Publica*, 27:s135–s140, 2011.

Laura M Pereira, C Nicholas Cuneo, and Wayne C Twine. Food and cash: understanding the role of the retail sector in rural food security in south africa. *Food security*, 6(3):339–357, 2014.

Paraskevi Peristera and Anastasia Kostaki. Modeling fertility in modern populations. *Demographic Research*, 16(6):141–194, 2007.

Jonas Peters, Peter Bühlmann, and Nicolai Meinshausen. Causal inference by using invariant prediction: identification and confidence intervals. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 78(5):947–1012, 2016.

A Petrie and C Sabin. *Medical Statistics at a Glance (Third Edition)*. Oxford, Blackwell, 2009.

Philayrath Phongsavan, Tien Chey, Adrian Bauman, Robert Brooks, and Derrick Silove. Social capital, socio-economic status and psychological distress among australian adults. *Social Science & Medicine*, 63(10):2546–2561, 2006.

Anna Pilgrim, Mary Barker, Alan Jackson, Georgia Ntani, Sarah Crozier, Hazel Inskip, Keith Godfrey, Cyrus Cooper, and Sian Robinson. Does living in a food insecure household impact on the diets and body composition of young children? findings from the southampton women's survey. *Journal of Epidemiology & Community Health*, 2011. doi: doi:10.1136/jech.2010.125476.

Sebastian Pink, Thomas Leopold, and Henriette Engelhardt. Fertility and social interaction at the workplace: Does childbearing spread among colleagues? *Advances in Life Course Research*, 21:113–122, 2014.

Paul E Plsek and Trisha Greenhalgh. Complexity science: The challenge of complexity in health care. *BMJ*, 323(7313):625, 2001.

Roberto Poli. A note on the difference between complicated and complex social systems. *Cadmus*, 2(1):142, 2013.

Craig Evan Pollack, Sekai Chideya, Catherine Cubbin, Brie Williams, Mercedes Dekker, and Paula Braveman. Should health studies measure wealth?: A systematic review. *Am J Preventive Medicine*, 33(3):250–264, 2007.

Rina Pradhan, Karen Wynter, and Jane Fisher. Factors associated with pregnancy among adolescents in low-income and lower middle-income countries: a systematic review. *J Epidemiol Community Health*, 69(9):918–924, 2015.

Paul M Pronyk, Trudy Harpham, Joanna Busza, Godfrey Phetla, Linda A Morison, James R Hargreaves, Julia C Kim, Charlotte H Watts, and John D Porter. Can social capital be intentionally generated? a randomized trial from rural south africa. *Social Science & Medicine*, 67(10):1559–1570, 2008.

Lenore Sawyer Radloff. The ces-d scale a self-report depression scale for research in the general population. *Appl Psych Meas*, 1(3):385–401, 1977.

Astha Ramaiya, Ligia Kiss, Paula Baraitser, Godfrey Mbaruku, and Zoe Hildon. A systematic review of risk factors for neonatal mortality in adolescent mothers in sub saharan africa. *BMC research notes*, 7(1):750, 2014.

Gabriela Ramos. Its not just the economy: society is a complex system too. `http://oecdinsights.org/2016/09/30/its-not-just-the-economy-society-is-a-complex-system-too/`, 2016. Accessed: 28 September 2017.

Carl Edward Rasmussen and Christopher KI Williams. *Gaussian processes for machine learning.* MIT press Cambridge, 2006.

Alan S Rigby, Gillian K Armstrong, Michael J Campbell, and Nick Summerton. A survey of statistics in three uk general practice journal. *BMC Medical Research Methodology*, 4(1):28, 2004.

Donald Rose and Karen E Charlton. Prevalence of household food poverty in south africa: results from a large, nationally representative survey. *Public Health Nutrition*, 5(3):383–389, 2002.

Donald Diego Rose. Interventions to reduce household food insecurity: a synthesis of current concepts and approaches for latin america. *Revista de Nutrição*, 21: 159s–173s, 2008.

Stacey Rosen, Birgit Meade, Keith Fuglie, and Nicholas Rada. International food security assessment, 2014-2024. *Economic Research*, 2014:2024, 2016.

Sandra Yu Rueger, Christine Kerres Malecki, and Michelle Kilpatrick Demaray. Relationship between multiple sources of perceived social support and psychological and academic adjustment in early adolescence: comparisons across gender. *J Youth Adolesc*, 39(1):47–61, 2010.

Harry Rutter, Natalie Savona, Ketevan Glonti, Jo Bibby, Steven Cummins, Diane T Finegood, Felix Greaves, Laura Harper, Penelope Hawe, Laurence Moore, et al. The need for a complex systems model of evidence for public health. *Lancet*, 2017.

Paul M Sacher, Maria Kolotourou, Paul M Chadwick, Tim J Cole, Margaret S Lawson, Alan Lucas, and Atul Singhal. Randomized controlled trial of the mend program: a family-based community intervention for childhood obesity. *Obes*, 18 (S1):S62–S68, 2010.

R Keith Sawyer. *Social emergence: Societies as complex systems*. Cambridge University Press, 2005.

Milo Schield. Correlation, determination and causality in introductory statistics. *American Statistical Association, Section on Statistical Education*, 1995.

Gideon Schwarz et al. Estimating the dimension of a model. *The Annals of Statistics*, 6(2):461–464, 1978.

Heather M Sedibe, Kathleen Kahn, Kerstin Edin, Tabitha Gitau, Anneli Ivarsson, and Shane A Norris. Qualitative study exploring healthy eating practices and physical activity among adolescent girls in rural south africa. *BMC Pediatrics*, 14 (1):211, 2014.

Alan Shiell, Penelope Hawe, and Lisa Gold. Complex interventions or complex systems? implications for health economic evaluation. *BMJ*, 336(7656):1281, 2008.

Galit Shmueli. To explain or to predict? *Statistical Science*, 25(3):289–310, 2010.

Jim Q Smith. *Bayesian decision analysis: principles and practice*. Cambridge University Press, 2010.

Katherine E Smith and Kerry E Joyce. Capturing complex realities: understanding efforts to achieve evidence-based policy and practice in public health. *Evidence & Policy*, 8(1):57–78, 2012.

Edward Snelson, Zoubin Ghahramani, and Carl E Rasmussen. Warped gaussian processes. In *Advances in neural information processing systems*, pages 337–344, 2004.

Diane H Sonnenwald. Scientific collaboration. *Annual review of information science and technology*, 41(1):643–681, 2007.

South African History Online. The Homelands. `http://www.sahistory.org.za/article/homelands`, 2017. Accessed: 13 July 2017.

South African Social Security Agency. Child Support Grant. `http://www.sassa.gov.za/index.php/social-grants/child-support-grant`, 2017a. Accessed: 27 October 2017.

South African Social Security Agency. Grants for Older Persons. `http://www.sassa.gov.za/index.php/social-grants/grants-for-older-persons`, 2017b. Accessed: 27 October 2017.

Sophie Staniszewska and Simon Denegri. Patient and public involvement in research: future challenges. *Evidence-Based Nursing*, 16(3):69–69, 2013.

John D Sterman. Learning from evidence in a complex world. *Am J Pub Health*, 96 (3):505–514, 2006.

Ruth Stewart, Carina van Rooyen, Kelly Dickson, Mabolaeng Majoro, and Thea de Wet. What is the impact of microfinance on poor people?: a systematic review of evidence from sub-saharan africa. *EPPI-Centre, Social Science Research Unit, Institute of Education, University of London*, 2010.

Valerie Tarasuk, Andy Mitchell, and Naomi Dachner. *Household food insecurity in Canada, 2012*. Research to Identify Policy Options to Reduce Food Insecurity, 2014.

AC Thomas. The social contagion hypothesis: comment on social contagion theory: examining dynamic social networks and human behavior. *Stat Med*, 32(4):581–590, 2013.

Ian M Timæus and Tom A Moultrie. Teenage childbearing and educational attainment in south africa. *Studies in Family Planning*, 46(2):143–160, 2015.

AJ Tomiyama, JM Hunger, J Nguyen-Cuu, and C Wells. Misclassification of cardiometabolic health when using body mass index categories in nhanes 2005–2012. *International Journal of Obesity*, 40(5):883–886, 2016.

Stefan Topolski. Understanding health from a complex systems perspective. *Journal of Evaluation in Clinical Practice*, 15(4):749–754, 2009.

Ioannis Tsamardinos, Constantin F Aliferis, and Alexander Statnikov. Time and sample efficient discovery of markov blankets and direct causal relations. In *Proceedings of the 9th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pages 673–678. ACM, 2003.

Ioannis Tsamardinos, Laura E Brown, and Constantin F Aliferis. The max-min hill-climbing bayesian network structure learning algorithm. *Machine Learning*, 65(1):31–78, 2006.

Compton J Tucker. Red and photographic infrared linear combinations for monitoring vegetation. *Remote Sensing of Environment*, 8(2):127–150, 1979.

Rhian Twine, Mark A Collinson, Tara J Polzer, and Kathleen Kahn. Evaluating access to a child-oriented poverty alleviation intervention in rural south africa. *Scand J Pub Health*, 35(69_suppl):118–127, 2007.

Wayne Twine and Lori Mae Hunter. Adult mortality and household food security in rural south africa: Does aids represent a unique mortality shock? *Development Southern Africa*, 28(4):431–444, 2011.

Koji Ueno. The effects of friendship networks on adolescent depressive symptoms. *Soc Sci Res*, 34(3):484–510, 2005.

United Nations. Sustainable Development Goals. `http://www.un.org/sustainabledevelopment/sustainable-development-goals/`, 2015a. Accessed: 14 September 2017.

United Nations. Millennium Development Goals Report. `http://www.un.org/millenniumgoals/`, 2015b. Accessed: 14 September 2017.

Thomas W Valente. Social network thresholds in the diffusion of innovations. *Soc Net*, 18(1):69–89, 1996.

Carina Van Rooyen, Ruth Stewart, and Thea De Wet. The impact of microfinance in sub-saharan africa: a systematic review of the evidence. *World Development*, 40(11):2249–2262, 2012.

Tyler J VanderWeele, Elizabeth L Ogburn, and Eric J Tchetgen Tchetgen. Why and when" flawed" social network analyses still yield valid tests of no contagion. *Statistics, Politics, and Policy*, 3(1), 2012.

Alfredo Vellido, José David Martín-Guerrero, and Paulo JG Lisboa. Making machine learning models interpretable. In *ESANN*, volume 12, pages 163–172, 2012.

Paul Voosen. How ai detectives are cracking open the black box of deep learning. `http://www.sciencemag.org/news/2017/07/how-ai-detectives-are-cracking-open-black-box-deep-learning`, 2017. Accessed: 29 September 2017.

Larry Wasserman. *All of statistics: a concise course in statistical inference*. Springer Science & Business Media, 2013.

Yair Weiss, Givat Ram Campus, and Judea Pearl. Belief propagation–perspectives. *Communications of the ACM*, 53(1), 2010.

WHO Commission on Social Determinants of Health. *Closing the gap in a generation: health equity through action on the social determinants of health: Commission on Social Determinants of Health final report*. World Health Organization, 2008.

Mick Wilkinson. Clinical and practical importance versus statistical significance: limitations of conventional statistical inference. *International Journal of Therapy and Rehabilitation*, 21(10):488–494, 2014.

Jill Williams, Latifat Ibisomi, Benn Sartorius, Kathleen Kahn, Mark Collinson, Stephen Tollman, and Michel Garenne. Convergence in fertility of south africans and mozambicans in rural south africa, 1993–2009. *Global Health Action*, 6, 2013.

Tim Wilson and Tim Holt. Complexity science: complexity and clinical care. *BMJ*, 323(7314):685, 2001.

Kate Wood and Rachel Jewkes. Blood blockages and scolding nurses: barriers to adolescent contraceptive use in south africa. *Reproductive Health Matters*, 14(27): 109–118, 2006.

World Health Organisation. Depression - Fact Sheet No. 369. `http://www.who.int/mediacentre/factsheets/fs369/en/`, 2014. Accessed: 27 July 2015.

Xiao-Hua Zhou, Donna K McClish, and Nancy A Obuchowski. *Statistical methods in diagnostic medicine*, volume 569. John Wiley & Sons, 2009.

APPENDIX A

Section 3.3 goodness of fit test results

Here we present the details and results of the goodness of fit tests for the results presented on weight contagion in Section 3.3.

We follow Hill et al. (2015) by adapting the Hosmer-Lemeshow test (HL) test, which analyses the distribution of residual errors, to the kind of regression we have performed (Hosmer and Lemeshow, 2001). The residual error function for our model stratified by number of friends is defined as

$$\mathcal{E} = \left(\sum_{k=0}^{10} \left(N_k - M_k \left(\theta \right) \right)^2 \right)^{1/2} \tag{A.1}$$

where N_k is the number of individuals with k higher/lower state friends who increased/decreased in state in the data, and $M_k\left(\theta\right)$ is the modelled number of such individuals given parameters θ. By definition, $\mathcal{E} \geqslant 0$ and will tend to zero for models that perfectly capture the data.

As an analytical distribution of $\mathcal{E}$ is not available, we use a parametric bootstrap method to simulate the fitted models. This involved applying the models to the initial friendship network from wave 1 of the data $N = 10^4$ times. The proportions of individuals with k higher/lower state friends who increased/decreased in state were extracted and used to calculate residuals.

	Figure A.1	Figure A.2
a	0.792	0.940
b	0.752	0.751
c	0.854	0.927
d	0.861	0.980
e	0.603	0.892
f	0.590	0.605
g	0.595	0.834
h	0.586	0.699
i	0.577	0.565
j	0.595	0.725
k	0.565	0.548
l	0.589	0.653
m	0.576	0.565
n	0.595	0.606
o	0.567	0.554
p	0.569	0.694

Table A.1: p-values from the goodness of fit test results of the weight contagion analysis. Figure and subfigure labels refer to those in Figures A.1 and A.2.

Figures A.1 and A.2 show the observed and simulated residual values for increasing/decreasing state stratified by higher/lower state friends for general weight change and clinically significant weight change for the four different models (with p-values given in Table A.1). Model 1 has both increasing and decreasing state dependent on friend states. Model 2 has neither dependent on friend states. Model 3 has increasing state alone dependent on friend states. Model 4 has decreasing state alone dependent on friend states. It should be noted that $\mathcal{E}$ has no asymptotic properties that imply it can be used for model selection in a manor similar to AIC, so no threshold of p-value should be considered. Simply, a larger p-value shows the model to form a better fit to the data. Overall, these results support our conclusions about which models are preferred in all cases.

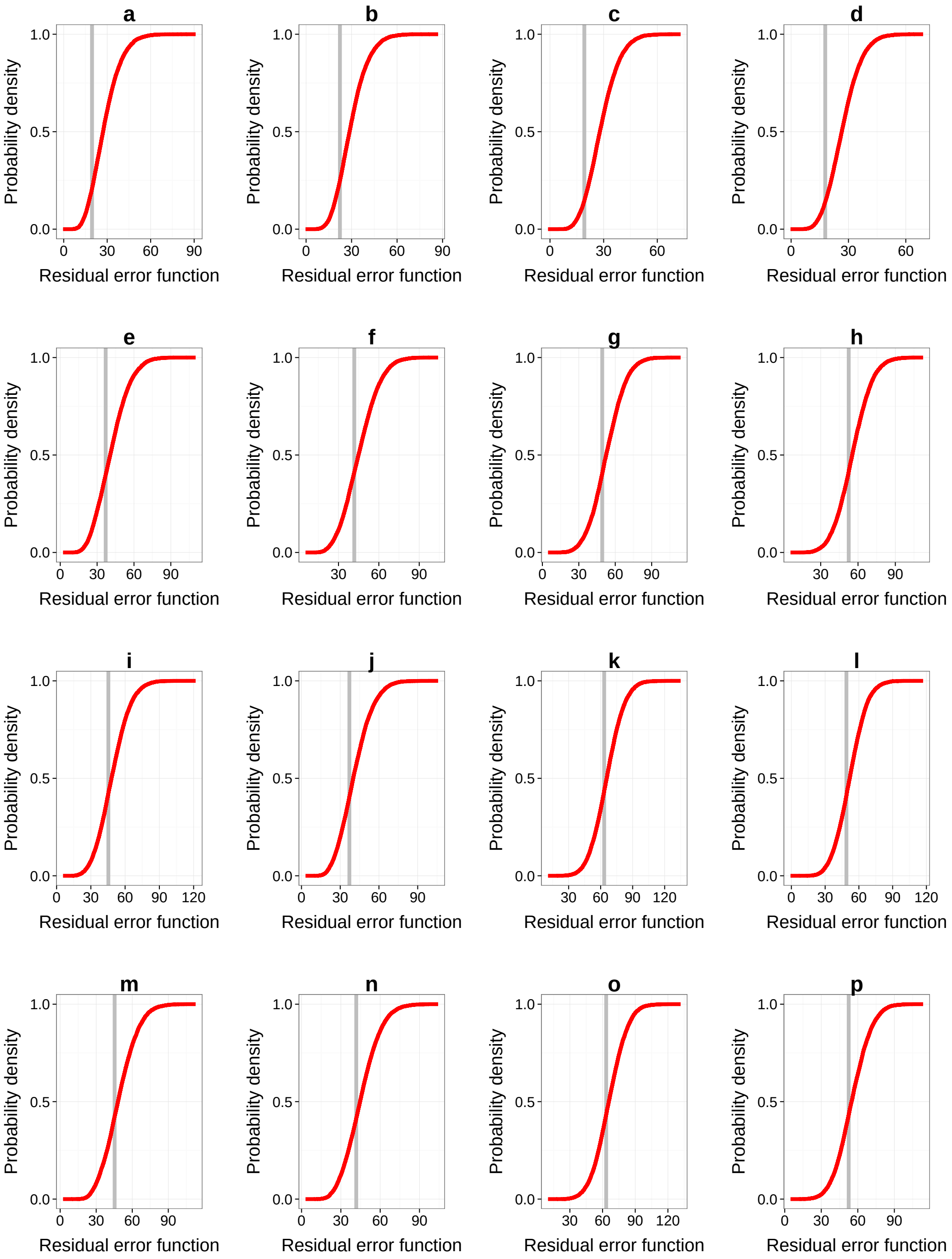


Figure A.1: Residual cdfs (in red) with observed residuals (in grey) for the fitted models for weight change. The first row (a-d) are for Model 1. The second row (e-h) are for Model 3. The third row (i-l) are for Model 4. The fourth row (m-p) are for Model 2. In each row, the left two plots are for the models dependent on higher weight friends, and the right two for those dependent on lower weight friends. In each of these pairs, the left plot shows the increasing state residuals and the right shows the decreasing state residuals.

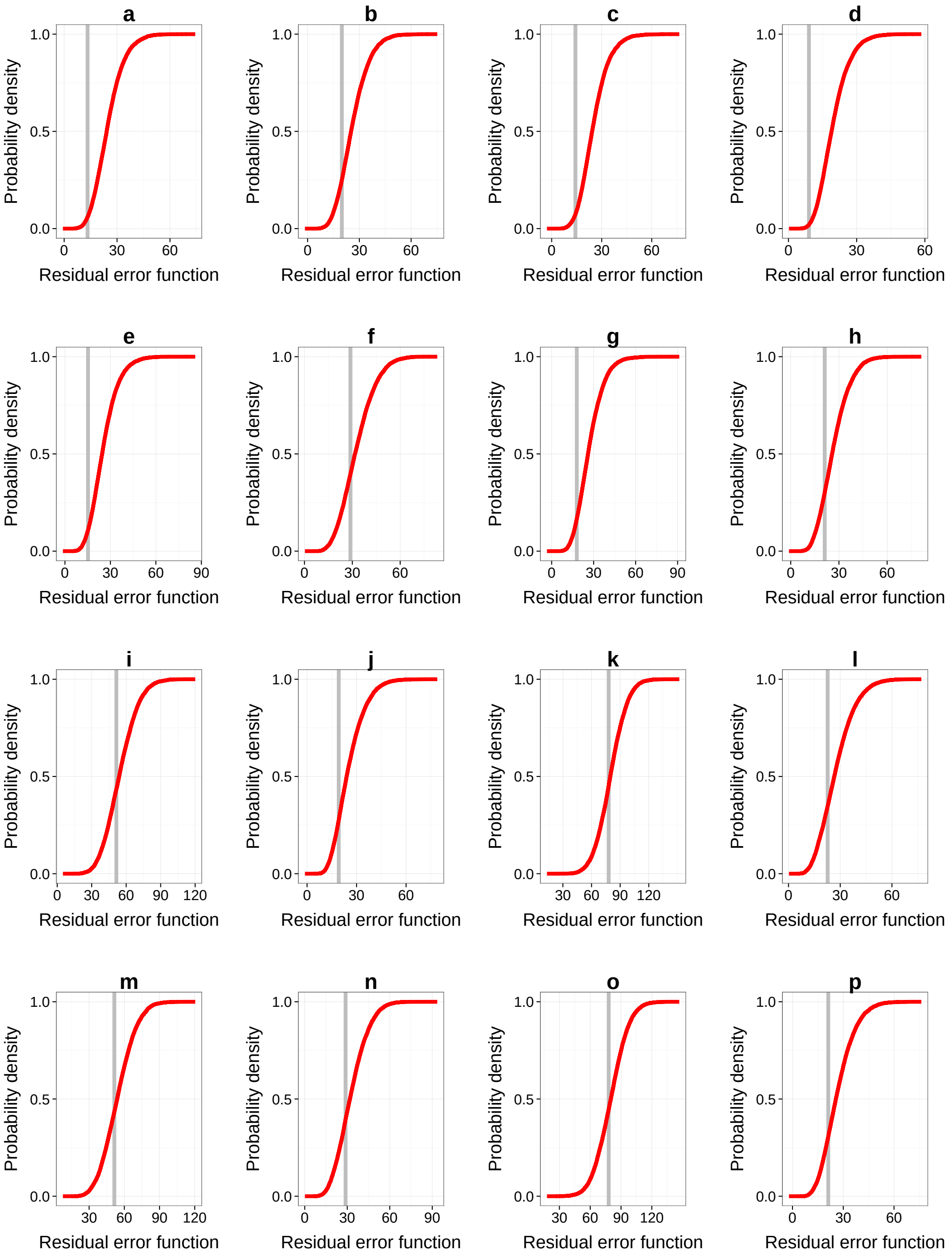


Figure A.2: Residual cdfs (in red) with observed residuals (in grey) for the fitted models for clinically significant weight change. The first row (a-d) are for Model 1. The second row (e-h) are for Model 3. The third row (i-l) are for Model 4. The fourth row (m-p) are for Model 2. In each row, the left two plots are for the models dependent on higher weight friends, and the right two for those dependent on lower weight friends. In each of these pairs, the left plot shows the increasing state residuals and the right shows the decreasing state residuals.

APPENDIX B

Section 4.3 cross validation and statistical test results

Here we present the results of the cross validation and statistical tests for the results presented in Section 4.3. The main results are shown in Table B.1, whilst more detailed results for particular models are shown in Tables B.2 to B.7.

Covariates	Age quantiles	SES quantiles	Average Briers score	KS p-value
s	50	10	0.04147	see table B.4
	100	20	0.04147	see table B.4
	125	25	0.04147	see table B.4
	150	30	0.04146	see table B.4
t	50	-	0.04620	4.219×10^{-15}
	100	-	0.04253	4.540×10^{-14}
	125	-	0.04193	0.0001
	150	-	0.04167	1.366×10^{-14}
e	50	-	0.04487	see table B.6
	100	-	0.04448	see table B.6
	125	-	0.04449	see table B.6
s, t	50	10	0.04155	0.04526
	100	20	0.04155	0.0008
	125	25	0.04155	0.0001
	150	30	0.04155	0.00001
s, e	25	10	0.04442	see table B.2
	50	10	0.04439	see table B.2
t, e	50	-	0.04248	0.0004
	100	-	0.04216	0.0053
	125	-	0.04217	0.0017
s, t, e	25	10	0.04221	0.3928
	50	10	0.04219	0.0672

Table B.1: Average Briers scores from cross validation and Kolmogorov-Smirnov test p-values of the different possible model and resolution combinations. The different possible models are formed from combinations of the different possible covariates of socio-economic status (SES) s, education in years e, and time in years t. The different possible resolutions are the the number of quantiles the data was binned to for age and SES. Briers scores for the s, e, and $\{s, e\}$ covariate combination models are averaged over years. Non-time-averaged Briers scores, as well as p-values, for these models can be found in Tables B.2 to B.7.

Resolution	Year					
	2001	2003	2005	2007	2009	2011
25x10	0.9854	0.9999	0.9960	0.9997	0.9998	0.9962
50x10	0.7786	0.9805	0.9097	0.9497	0.9533	0.9180

Table B.2: p-values for individual years and resolutions for the $\{s, e\}$ covariate combination model.

Resolution	Year					
	2001	2003	2005	2007	2009	2011
25x10	0.05790	0.03826	0.04382	0.04141	0.04267	0.04244
50x10	0.05784	0.03827	0.04382	0.04136	0.04268	0.04237

Table B.3: Briers scores for individual years and resolutions for the $\{s, e\}$ covariate combination model.

Resolution	Year					
	2001	2003	2005	2007	2009	2011
50x10	0.8057	0.9466	0.8226	0.9092	0.9234	0.8662
100x20	0.3353	0.6500	0.3976	0.5790	0.6000	0.4667
125x25	0.2358	0.4824	0.2269	0.3548	0.4472	0.3141
150x30	0.1703	0.3563	0.1697	0.3315	0.3546	0.2129

Table B.4: p-values for individual years and resolutions for the s covariate combination model.

Resolution	Year					
	2001	2003	2005	2007	2009	2011
50x10	0.04004	0.03813	0.04379	0.04161	0.04279	0.04249
100x20	0.04002	0.03813	0.04378	0.04160	0.04278	0.04248
125x25	0.04002	0.03813	0.04378	0.04159	0.04278	0.04248
150x30	0.04002	0.03813	0.04379	0.04159	0.04279	0.04248

Table B.5: Briers scores for individual years and resolutions for the s covariate combination model.

Resolution	Year					
	2001	2003	2005	2007	2009	2011
50	0.8318	0.9531	0.9185	0.9438	0.9796	0.9376
100	0.4922	0.6741	0.5458	0.6145	0.7812	0.5885
125	0.4312	0.5330	0.4490	0.4897	0.6633	0.4440

Table B.6: p-values for individual years and resolutions for the e covariate combination model.

Resolution	Year					
	2001	2003	2005	2007	2009	2011
50	0.05823	0.03822	0.04407	0.04159	0.04311	0.04398
100	0.05814	0.03822	0.04381	0.04156	0.04273	0.04244
125	0.05818	0.03821	0.04381	0.04154	0.04273	0.04245

Table B.7: Briers scores for individual years and resolutions for the e covariate combination model.

APPENDIX C

Section 4.4 covariance function comparison results

Here we present the Bayesian information criterion values used in Section 4.4 to compare the different possible covariance functions. Tables C.1 to C.5 show these for the different analyses performed.

Year	Covariance function			
	Squared exponential	Matern 3/2	Matern 5/2	Rational quadratic
2001	334.3	333.0	333.5	331.8
2002	1168	1165	1166	1165
2003	1194	1193	1194	1193
2004	1206	1205	1205	1205
2005	1390	1389	1390	1388
2006	1228	1227	1228	1226
2007	1305	1304	1304	1304
2008	1575	1574	1574	1573
2009	1660	1659	1659	1658
2010	1405	1404	1404	1404
Average	1247	1245	1246	1245

Table C.1: Bayesian information criterion values for Model 1 (no contagion) of both the global and local contagion analyses for each of the different possible covariance functions. The rational quadratic covariance function is generally preferred.

Year	Covariance function			
	Squared exponential	Matern 3/2	Matern 5/2	Rational quadratic
2002	1174	1172	1172	1173
2003	1202	1202	1202	1202
2004	1213	1214	1214	1213
2005	1397	1396	1396	1395
2006	1236	1235	1235	1234
2007	1312	1310	1310	1312
2008	1582	1581	1582	1581
2009	1666	1665	1666	1666
2010	1411	1411	1411	1411
Average	1355	1354	1354	1354

Table C.2: Bayesian information criterion values for Model 2 (contagion) of the global contagion analysis with an annual contagion period for each of the different possible covariance functions. The rational quadratic covariance function is generally preferred.

Year	Covariance function			
	Squared exponential	Matern 3/2	Matern 5/2	Rational quadratic
2003	1202	1202.	1202	1202
2004	1214	1213	1213	1213
2005	1397	1397	1398	1396
2006	1236	1235	1235	1234
2007	1311	1311	1311	1311
2008	1581	1581	1580	1581
2009	1667	1666	1666	1665
Average	1373	1372	1372	1372

Table C.3: Bayesian information criterion values for Model 2 (contagion) of the global analysis with a biennual contagion period for each of the different possible covariance functions. The rational quadratic covariance function is generally preferred.

Year	Covariance function			
	Squared exponential	Matern 3/2	Matern 5/2	Rational quadratic
2002	1175	1173	1174	1173
2003	1202	1202	1202	1201
2004	1214	1214	1214	1213
2005	1399	1398	1398	1396
2006	1235	1233	1234	1233
2007	1313	1312	1312	1311
2008	1582	1581	1581	1580
2009	1668	1666	1667	1664
2010	1414	1412	1412	1412
Average	1356	1355	1355	1354

Table C.4: Bayesian information criterion values for Model 2 (contagion) of the local contagion analysis with an annual contagion period for each of the different possible covariance functions. The rational quadratic covariance function is generally preferred.

Year	Covariance function			
	Squared exponential	Matern 3/2	Matern 5/2	Rational quadratic
2003	1202	1202	1202	1202
2004	1214	1214	1214	1213
2005	1399	1398	1398	1396
2006	1236	1234	1235	1234
2007	1311	1310	1311	1308
2008	1581	1580	1581	1579
2009	1669	1667	1668	1665
Average	1373	1372	1373	1371

Table C.5: Bayesian information criterion values for Model 2 (contagion) of the local contagion analysis with a biennual contagion period for each of the different possible covariance functions. The rational quadratic covariance function is generally preferred.

APPENDIX D

Section 4.4 biennual contagion period results

Here we present the results from Section 4.4 when contagion of pregnancy risk behaviour is assumed to happen over the span of two years rather than just one. As mentioned in Section 4.4.4 using a biennual contagion period generally results in reduced performance, so in the main analysis we focused on the results from using an annual period. The results here show that despite the worsening in performance a biennual contagion period produces very similar results to those shown in Section 4.4.4.

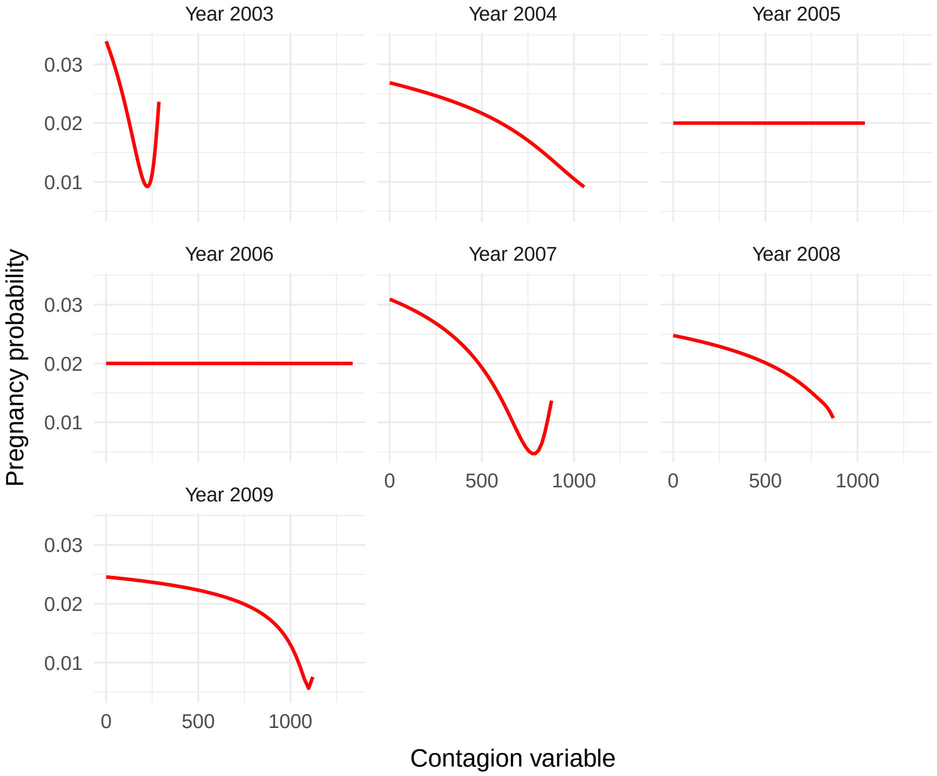


Figure D.1: Fitted marginal pregnancy probabilities over contagion variable value for each year of the global contagion analysis (with the contagion variable ranging up to the maximum value within the data). Results are similar to those with an annual contagion period (Figure 4.8).

Year	No contagion S_B	Contagion S_B	Contagion lengthscale
2003	0.0891	0.0891	4.465
2004	0.0761	0.0761	24.47
2005	0.0797	0.0797	772.9
2006	0.0874	0.0874	924.3
2007	0.0879	0.0878	5.397
2008	0.0742	0.0742	4.088
2009	0.0736	0.0739	2.457
Average	0.0811	0.0812	5.397

Table D.1: Comparison between Model 1 (no contagion) and Model 2 (contagion) from the global contagion analysis for each year shown by the Briers scores (S_B) from predicting the pregnancies for the next timestep. Contagion lengthscale values from the Gaussian process classification for Model 2 are also included. The performance of both models is generally worse than when using an annual contagion period (Table 4.4). Otherwise the results are similar. Note the average lengthscale is the median, whilst for Briers scores they are means.

Year	No contagion S_B	Contagion S_B	Contagion lengthscale
2003	0.0891	0.0891	145.7
2004	0.0761	0.0761	5196
2005	0.0797	0.0797	2327
2006	0.0874	0.0870	8.312
2007	0.0879	0.0880	1.342
2008	0.0742	0.0742	5.538
2009	0.0736	0.0739	0.8944
Average	0.0811	0.0811	8.312

Table D.2: Comparison between Model 1 (no contagion) and Model 2 (contagion) from the local contagion analysis for each year shown by the Briers scores (S_B) from predicting the pregnancies for the next timestep. Contagion lengthscale values from the Gaussian process classification for Model 2 are also included. The performance of both models is generally worse than when using an annual contagion period (Table 4.5). Otherwise the results are similar. Note the average lengthscale is the median, whilst for Briers scores they are means.

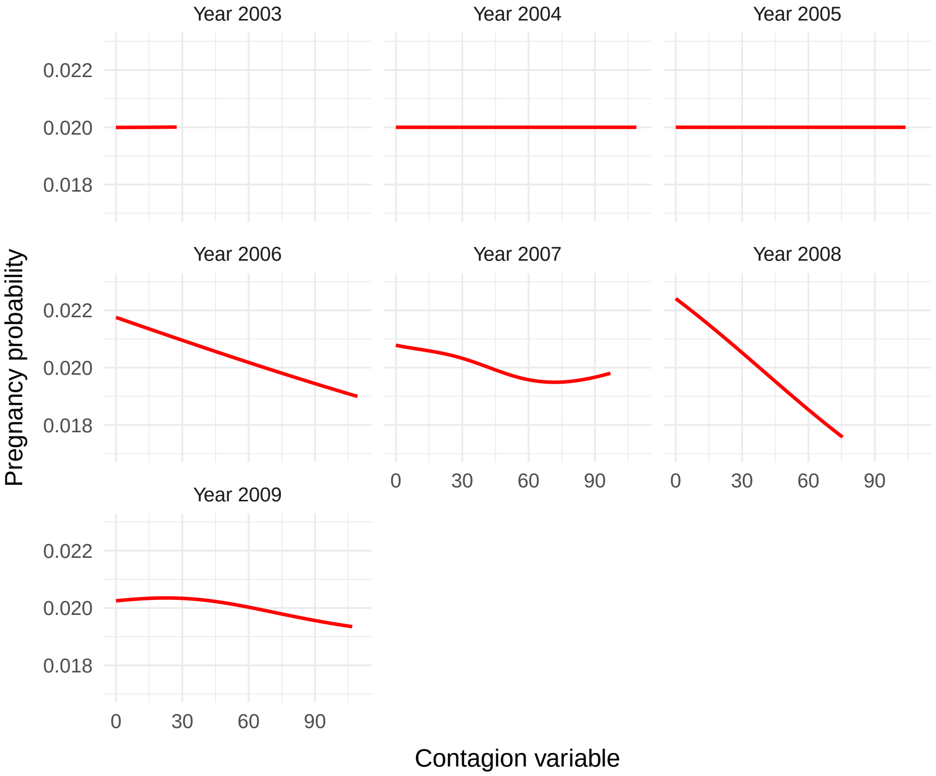


Figure D.2: Fitted marginal pregnancy probabilities over contagion variable value for each year of the local contagion analysis (with the contagion variable ranging up to the maximum value within the data). Results are similar to those with an annual contagion period (Figure 4.9).