Estimation of synthetic neighbourhood boundaries for multilevel analysis of the contextual determinants of mental health and investigation of associated methodological issues

Mark Kelly

School of Medicine, Cardiff University

ProQuest LLC
789 East Eisenhower Parkway
P.O. Box 1346
Ann Arbor, MI 48106-1346

You will find, as a general rule, that the constitutions and the habits of a people follow the nature of the land where they live.

Hippocrates- 460-377BC

Summary

The investigation of associations between places, people and mental health is complicated and there are serious limitations in the current methodology. Using data from the Caerphilly Health and Social Needs Study (CHSNS), as well as the British Household Panel Survey, this thesis investigated some of these methodological issues.

Firstly, motivated by the skewed distribution of the Mental Health Inventory (MHI-5), methods for analysing the mental health score were examined. Five methods for deriving a cutpoint on the MHI-5 based on linking with the General Health Questionnaire, were investigated and cutpoints derived for each. These cutpoints and methods were compared and contrasted.

When investigating associations between place and health, hierarchical modelling is an extremely useful tool. Sparse levels of information are a potential problem when using this method. In the CHSNS, households represent a sparse level of context. A simulation study was conducted to explore the effect of sparse levels on the results of hierarchical analyses. It was found that, in general, the underestimation of fixed effect standard errors is smaller when a sparse level is included than when it is excluded.

Another methodological consideration for hierarchical modelling concerns the choice of geographical hierarchy to use. Administrative hierarchies have been criticised for being heterogeneous and arbitrary. An algorithm was developed to partition regions into areas that are homogenous with respect to a given set of variables, and was applied to Caerphilly county borough. The resulting sets of boundaries were compared with the 2001 census administrative boundaries. These new boundaries performed favourably in comparison with the administrative boundaries, indicating that administrative boundaries may not represent the most suitable hierarchy to employ in hierarchical analyses.

The thesis has led to a greater understanding of the effects of context on multilevel analysis and contributed to the area-effects on health literature.

Acknowledgements

This thesis would not have been possible without the help and support of a large number of people. I would like to acknowledge James Osborne for all his hard work in setting up the Condor computing service which helped strengthen my thesis considerably. Next I must thank Dr. Daniel Farewell; his willingness to help is rivalled only by his ability to do so. I must also pay tribute to my supervisors Professors David Fone and Frank Dunstan. Their expertise, advice, and enthusiasm throughout the course of my PhD were greatly appreciated. I would like to dedicate all of the punctuation, both; correct, and. incorrectly-placed, to' them. I should acknowledge my brothers, Donal and Robert at this point. Your contributions to the thesis were as important as they were fictional and for that I am truly grateful. Your contributions toward making me the person I am however, are great and I will continue to aspire to be more like (the good parts of) both of you. Penultimately I thank my girlfriend Olivia Roche who provides constant support, encouragement and affection and makes my life better just by being there. I couldn't have done it without you. To her I would like to dedicate all of the 119,782 vowels in this thesis. Last, but by no means least, I thank my parents Donal and Teresa Kelly. To adequately catalog their contributions to this thesis would require a hundred page appendix incorporating an economic assessment of the cost of raising a child through to financial independence (read: 27 years of age and counting), a section detailing the man-hours spent drumming times tables and countless other important concepts into my head and a glossary covering definitions of Renaissance-style patronage and Job-like patience. In the interests of brevity I will simply say this: Thank you, I owe it all to you.

Contents

Chapter 1

Introduction

1.1 Place and health

Epidemiologists and public health scientists have long been concerned with assessing the effect of various exposures on people's health. This is a long tradition stretching all the way back to Hippocrates (460-377 B.C) who first recognized the association of disease with place, water conditions, climate, eating habits and housing. The theories may have changed considerably from the four humours postulated by Hippocrates, but the spirit of questioning and investigation remain the same. It is fitting then that this thesis is concerned with one of the most basic exposures experienced by everyone: where one lives. Early efforts to investigate the relationship between area and health focussed on ascertaining why given diseases were more prevalent in certain areas. Often these investigations provided profound insights into the aetiology of the disease itself. An oft quoted example relates to the study performed by John Snow, where he mapped the cases of cholera in Soho, London during an epidemic, revealing that the cases centred around a public water pump (Snow, 1849). The water from this pump was found to be infected with cholera and the water borne nature of the disease was discovered.

Nowadays the link between place and health is more subtle, with far less chance of discovering a "smoking gun" cause such as the one described above. Nevertheless, health disparities between places abound to this day, with a recent book quoting a twenty-year age gap in life expectancy of men between areas separated by just twelve miles in Washington D.C. (Marmot, 2004). This is neither an isolated nor extreme example. Indeed, according to Subramanian et al (2003) *"The question, therefore, is not whether variations exist between various settings (they always do), but what is their source, that is, are the variations across settings compositional or contextual?"*. So to rephrase, are these differences due to the various attributes of the residents or

to some intrinsic property of the areas themselves? In this thesis this question and the methodology surrounding it will be investigated for the outcome of the common mental disorders of anxiety and depression.

There is already a large literature base concerned with the effect of area on health. In the past however, the focus was not on the areas themselves, but rather on the various physical properties of the areas which were treated as exposure variables (Macintyre et al., 1993). Researchers were concerned with identifying the aetiology of disease and areas were their unit of analysis. The attitude toward area effects was that they represented variability associated with some unmeasured individual-level variable. The approach was to measure the right exposure or individual risk factors in order to explain away the area effect (Pickett & Pearl, 2000). Macintyre et al (2002) discuss the reasons for this giving four separate explanations for why this was the case. Firstly, they claim that researchers were (and still are) determined to avoid the ecological fallacy, where relationships observed at an area (or group) level are found to be different from the equivalent individual-level relationship. They cite a seminal sociological paper which found that the individual and group relationships between foreign birth and illiteracy had opposite signs (Robinson, 1950). Macintyre et al's second reason is that statistical and computing advances allowed researchers unprecedented capability to analyse information on individuals. This, coupled with the availability of large datasets, meant that researchers were inclined to focus on individual-level analyses. Thirdly, Macintyre et al claim that *"from the 1950s onwards methodological, conceptual and political individualism was dominant in many industrialised countries"*. This manifested itself in an emphasis on the role of individual attributes on health (notably smoking, drinking, diet and exercise), thus drawing attention away from area-level effects. This concept is summed up by Margaret Thatcher's infamous quote *"there is no such thing as society, there are only individuals"*. The final reason concerns a lack of interest in using the geography of the study area to explicitly inform and conceptualise the hypotheses. More recently however, the focus has shifted away from treating the area as a nuisance factor that needs to be controlled for, into treating the area as a variable of interest in its own right. Indeed, the whole field has undergone something of a revival in recent years with the advent of new statistical tools to investigate such effects.

A book published in 2003, succinctly entitled "Neighbourhoods and Health" (Kawachi & Berkman, 2003), summarised the progress to date as well as indicating emerging issues in neighbourhood research. The authors categorise these issues under the following headings: Social Selection versus Social Causation; Contextual versus Compositional Effects; Psychosocial versus Material Explanations; Subjective versus Objective Assessments; Quantitative versus Qualitative Approaches and Neighbourhoods versus Communities.

The first of these, social selection versus social causation, refers to the debate be-

tween whether places affect their residents, or whether people shape where they live. An example might be "do people with poorer mental health choose to (or at least are complicit in the decision to) move to and stay in deprived areas, or do deprived areas have a detrimental effect on the psychological well-being of their residents?". Perhaps a more plausible causal pathway is that healthy people are better equipped to find the means to move away from deprived areas, thus decreasing the average health of an area. Essentially, this refers to the issue of causation. This issue could possibly be resolved using a longitudinal study design, which could tease out the causal pathway. While common sense might suggest that both processes are probably at work, it would still be an important step forward for the field if better evidence were available to address the issue of causation.

The contextual versus compositional debate, the second area of emerging interest cited, concerns a related question about the distinction between individual characteristics and area characteristics. As mentioned earlier, much of the previous literature has focused on individual-level characteristics, with the area-level being a nuisance factor which needs to be explained away. Macintyre and Ellaway (Macintyre et al., 2002) discuss this, saying

> *"Within both epidemiology and geography there has been a tendency to ascribe much within-country geographical variation to compositional differences, and until recently there has been apparent resistance to any role for contextual explanations. It has almost been an article of faith that differences between places are reducible to differences between the types of people living there".*

Clearly both compositional and contextual effects could possibly influence an individual's mental health. If for example the association between income and mental health is of interest, it would certainly be neccessary to include the individual-level income variable as an explanatory variable in the model. However, it would be prudent to also include a measure of area-level income (the aggregated income of all residents of an area) also, since the effect of earning minimum wage in an area rife with unemployment may be different to earning minimum wage in an affluent neighbourhood. In other words, it may not be just the magnitude of one's income that influences one's mental health, but the relative magnitude of one's income compared with one's neighbours. So, the question of whether the levels of mental health reported in an area depend on the type of people who live there (compositional effect), or whether they are due to area characteristics (contextual effect), may be an incorrectly framed question, since it may be the joint impact of both of these characteristics that is really of interest. Certainly the clear distinction between the two types of effect seen in the literature is not reflected in reality. Even what may be classified as a true contextual variable,

that is a variable that is not merely the aggregation of individual characteristics, (e.g. number of grocery shops in an area), could also be seen as dependent on the type of resident of that area (e.g. shops will open where there is demand for them). The problem occurs when variables considered to be compositional in nature, such as smoking, diet and exercise, are controlled for in an analysis. The idea is that if there remains a significant area-level effect after these individual characteristics have been controlled for, then this can be considered evidence for a contextual effect. The problem is, of course, that area characteristics may influence these variables, again discussed by Macintyre et al (2002). For instance living in an area with a high prevalence of smoking may make a person more likely to smoke. Living in an area where there are lots of fast food outlets may impact negatively on a person's diet. Similarly, living in an area where there are not many leisure facilities may result in a person exercising less. In this situation, the practice of controlling for confounding variables becomes a type of statistical overadjustment. The way to deal with this problem is far from obvious.

Another debate that continues in this field is the psychosocial versus material explanations debate (this is Kawachi and Berkman's (2003) third area of emerging interest). The psychosocial explanation theory seeks to explain poor health behaviours and outcomes in terms of the psychology of the residents. An example of this, given by the authors, is the contagion effect of high smoking prevalence on smoking initiation in adolescents. Here it is the complex interplay of social norms and peer pressure that results in poor health behaviour. The opposing theory, material explanations, seeks to explain such outcomes and behaviour in terms of more concrete and visible characteristics of an area. An example would be increased levels of obesity in a neighbourhood where there are no leisure facilities. Common sense would dictate that both mechanisms operate together, and indeed affect one another. Psychosocial characteristics may affect material ones and vice versa. An area with a high proportion of short term residents may result in residents not knowing each other (i.e. poor social networks). This in turn may decrease the demand or usage of community centres, resulting in their closure. Conversely, a lack of amenities in an area may result in fewer opportunities for social interaction, which could feasibly impact negatively on the mental health of its residents. What is needed is the development of new theories to describe the mechanisms by which these characteristics can explain health. Kawachi and Berkman (2003) call for researchers to eschew routinely collected data, and instead to gather information specific to the research question, through more bespoke primary data collection. They also call for more studies that measure mental health. This call has been echoed by a recent review of social capital (i.e. in the words of the authors "a way of describing social relationships within societies or groups of people") (De Silva et al., 2005). They contend that the evidence from the literature neither justifies specific social capital interventions in order to improve mental illness nor is it sufficient to inform how such

interventions might be designed

The fourth area of emerging interest highlighted surrounds the use of subjective versus objective assessments. In this context, subjective refers to information collected from residents themselves, for example, their ratings of their own area of residence. Objective corresponds to measurements of an area that are (relatively) indisputable and highly repeatable, e.g. number of shops in an area. The solution proposed by Kawachi and Berkman (2003) is that there is unique information to be gained from both approaches and that greater insight will be afforded to researchers willing to seek a happy medium between the two methods. A subset of this issue is the problem of same source bias. People with poor mental health may have a more pessimistic or negative outlook on life, and thus may rate their environment and local amenities more harshly than they perhaps deserve (Duncan & Raudenbush, 1999; O'Campo, 2003). This could bias the relationship between poor mental health and deprived neighbourhoods resulting in an observed positive relationship that exaggerates the true situation. The problem can be circumvented if data from different sources are combined, so that the people who provide the measurements of an area's characteristics are different to the ones whose mental health measurements are used in an analysis.

The fifth emerging area of interest the authors cite is the conflict between quantitative and qualitative research. It seems slightly out of place to categorise this debate as emerging, since it would appear that this conflict has been around as long as the two approaches themselves have. The stance the authors advocate is the same as for all the other issues, namely that the approaches offer different yet mutually complementary information. Qualitative studies can tease out relationships that even the most detailed and complicated quantitative studies could never hope to replicate. Conversely, qualitative studies have been criticised for being too subjective and non-generalisable, and as such will usually require additional support from a quantitative approach to implement policy change. Quantitative analysis proponents on the other hand claim that their approach is objective, generalisable and most importantly can provide information regarding the confidence a researcher can place in a given estimate. The detractors of quantitative analysis suggest that in many cases statistical methodology is misused and is open to abuse. The supposed objectivity of the quantitative approach can be questioned also, since there is often no consensus from the statistical community over the best way to model a given situation. Moreover, qualitative researchers point out that without their research, quantitative researchers would not be able to generate hypotheses of interest. It seems obvious then that the best way to proceed is to utilise the symbiotic nature of both approaches in order to best harness their different, but complementary strengths.

The last emerging issue for area based health research is the conflict between neighbourhoods and communities. This concerns the choice of context to adopt when per-

forming area based research. The authors urge the collection of information from various contexts apart from the traditional neighbourhood context. They list workplaces, schools and virtual communities as contexts worthy of further investigation and indeed hedge their bets slightly by saying *"some years hence it may indeed turn out to be the case that neighbourhoods explain rather less of the variations in health than do other contexts"*. The final advice the authors have for prospective researchers in this field is to simultaneously model multiple levels of context. They point out that neighbourhood characteristics may be merely reflections of larger macroeconomic forces and more widescale political decisions. The methodology exists now to handle, and indeed unravel, the complex interplay between the small scale influences and the more global ones, even if datasets rich enough to encapsulate such information are few and far between.

This thesis will directly address two of the six areas highlighted by this book, namely the question of context versus composition, as well as the issue of neighbourhoods versus communities. The first of these will be examined using hierarchical (or multilevel) modelling which will be introduced in chapter 5. The second issue concerns the choice of what to use to define area of residence or neighbourhood. There is no practical or usable definition of the term neighbourhood. Galster (2001) summarizes the situation neatly by saying:

> *"Urban social scientists have treated 'neighbourhood' in much the same way as courts of law have treated pornography: as a term that is hard to define precisely, but everyone knows it when they see it. "*

He then goes on to define it as *"...the bundle of spatially based attributes associated with clusters of residences, sometimes in conjunction with other land uses"*. For the purposes of statistical modelling however, this definition is far from workable. Part of this thesis will be concerned with deriving a method to delineate areas.

1.2 Mental Health

Independently of the aforementioned revolution in modelling area effects, the public health importance of mental health is also being increasingly recognised. The mental health status of an individual is complicated and difficult to quantify accurately. Indeed, what might constitute good mental health is difficult to determine. The medical literature tends to focus on identifying poor mental health, as opposed to explicitly defining good mental health. The World Health Organisation (WHO) (World Health Organisation, 1992) defines a mental health "disorder" as a term used to:

"imply the existence of a clinically recognizable set of symptoms or behaviour associated in most cases with distress and with interference with personal functions. Social deviance or conflict alone, without personal dysfunction, should not be included in mental disorder as defined here."

Along the same lines, but more expansive, is the definition given by the American Psychiatric Association Diagnostic and Statistical Manual of Mental Disorders (DSM IV, 2000). They say that a mental disorder is:

"conceptualized as a clinically significant behavioural or psychological syndrome or pattern that occurs in an individual and that is associated with present distress (e.g., a painful symptom) or disability (i.e., impairment in one or more important areas of functioning) or with significantly increased risk of suffering death, pain, disability, or an important loss of freedom. In addition, this syndrome or pattern must not be merely an expectable and culturally sanctioned response to a particular event, for example, the death of a loved one. Whatever its original cause, it must currently be considered a manifestation of a behavioural, psychological, or biological dysfunction in the individual. Neither deviant behaviour (e.g., political, religious, or sexual) nor conflicts that are primarily between the individual and society are mental disorders unless the deviance or conflict is a symptom of a dysfunction in the individual as described above."

These disorders represent a significant public health issue, and are leading causes of morbidity and disability. A WHO publication examined the global burden of all diseases, using Disability-Adjusted Life Years (DALYs) and Years lived with Disability (YLDs) (Murray & Lopez, 1996). Of the top ten causes of YLDs for the world, five belong to the neuro-psychiatric category (unipolar major depression, alcohol use, bipolar disorder, schizophrenia and obsessive-compulsive disorder) and account for almost 22%. All neuro-psychiatric conditions are estimated to account for just under 30% per cent of all YLDs, making them the single most important contributors. The authors also state that *"neuro-psychiatric conditions account for 10.5% of the global burden of disease and injury and that uni-polar major depression is the fourth most important cause of DALYs"*. It seems then, that an absence of such disorders could be deemed as good mental health. Such a definition, however, is unworkable if one wishes to examine large numbers of people. Instead a continuous measure, capable of assessing the whole spectrum of mental health, is needed. Quantifying something as complicated as mental health would be a considerable undertaking, even if it were static through time. However, a person's mental health today may not be the same as their mental health tomorrow. This is a problem not easily overcome, and indeed no solution is addressed in this thesis either.

Instead, this thesis focuses on the minor anxiety and depressive disorders, termed the common mental disorders (CMD) (Goldberg & Huxley, 1992) in the general population. The prevalence of CMD in Wales has been estimated at 31.0% of the population (Weich et al., 2001). The burden of common mental disorders is considerable. In economic terms alone they constitute a large problem with one-third of days lost from work due to ill health being attributable to common mental disorders (Jenkins, 1985). Nearly twenty years ago they were estimated to cost up to £6 billion in the UK (Croft-Jefferys & Wilkinson, 1989). Moreover, in terms of health care service provision, they place a sizeable burden on the system, with common mental disorders accounting for one-fifth of general practice consultations (Williams et al., 1986). Aside from the obvious economic cost however, the common mental disorders have serious consequences for those that suffer from them. Previously, lists of the most important public health issues considered just the mortality effects of disease, but more recently, both mortality and disability effects of disease were examined simultaneously. Since depressive disorders have a high prevalence, a high impact on functioning, and an early age of onset, this change has pushed mental disorders into the WHO's top 10 of public health priorities (Üstün et al., 2004).

This thesis combines both of these areas of interest (area effects on health and mental health) by investigating the contextual determinants of mental health. The message from the literature surrounding area effects on mental health is quite discordant, with no clear consensus emerging about whether they even exist, let alone what magnitude of influence they could be expected to exert. In the past five years various studies have come to quite different conclusions regarding the existence of area effects on mental health. Some studies have found no evidence of an area effect (Reijneveld & Schene, 1998; Reijneveld et al., 2000; Drukker & van Os, 2003; Ross, 2000; Propper et al., 2005; Weich et al., 2003a), while others have found significant area effects (Fone & Dunstan, 2006; McCulloch, 2001; Skapinakis et al., 2005; Weich et al., 2003b; Wainwright & Surtees, 2003; Fone et al., 2007a,b,c). One review claimed that the evidence for area effects on mental health is "reasonably sound" (Ellen et al., 2001). The authors investigated five health outcomes, of which one was mental health (the others being health-related behaviours, low birth weight and infant mortality, adult physical health, overall mortality). The review was restricted to multilevel analyses incorporating both individual and area-level information. For the outcome of mental health, they assert that associations have been found to exist between the quality of social networks and social cohesion and various non-psychotic psychiatric disorders, right across the life cycle. They also note that there is evidence to support the theory that neighbourhood violence has negative psychological effects for children. No evidence for the same effect has been identified among adults; however there is evidence for the relationship between living in a high poverty area and poor mental health. The authors conclude

by saying:

> *"...the relationship between neighbourhoods and mental health is underexplored, but the current evidence does suggest a provisional story. In brief, we find the strongest evidence for independent neighbourhood effects on overall mortality as well as on health outcomes that can be expected to develop and be discernible fairly quickly, such as health-related behaviours and mental health."*

A systematic review of neighbourhood socioeconomic context and health outcomes published in 2000 found that *"the evidence for modest neighbourhood effects on health is fairly consistent"* (Pickett & Pearl, 2000). The health outcomes reviewed were mortality, morbidity and health behaviours. Only one paper had mental health as the outcome of interest (Reijneveld & Schene, 1998) and found no evidence for an area effect of socioeconomic deprivation after individual socioeconomic status had been included. Two studies investigating the effect of area deprivation on mental health in Wales found that area-level deprivation explains some of the geographical variability in mental health (Skapinakis et al., 2005; Fone et al., 2006b).

The waters are further muddied by the fact that there is no standard methodology adopted by researchers in the field, with different area sizes being used as proxies for neighbourhoods, different sample sizes, different ways of measuring mental health and different compositional factors controlled for. If there is a general consensus however, it is that the effect on mental health of neighbourhoods is dwarfed by the effect of individual characteristics. This is to be expected, but enough studies have uncovered significant area effects after adjusting for compositional variables to indicate that the mental health of individuals is not solely determined by their individual characteristics.

The suggestions for the direction of future research are more consistent. One recurring theme is that there is a need for better theories regarding the causal pathway between neighbourhood characteristics and mental health (Drukker & van Os, 2003; Weich et al., 2003b; Blakely & Woodward, 2000; Diez Roux, 2001). Once these theories are formulated they can be examined and tested to see if they are consistent with the data. This would elucidate the mechanisms by which area characteristics such as deprivation and unemployment levels may adversely affect mental health. This, in turn, would inform researchers about the salient information necessary to conduct such studies as well as providing policy makers with a complete map of the causal pathway, indicating the most efficacious point of intervention.

Another area requiring further research involves investigating the statistical methodology of hierarchical modelling. Hierarchical modelling allows for the correct analysis of hierarchically structured data. An implicit assumption in hierarchical modelling is that the data at each level of the analysis represent a random sample from the popula-

tion at that level. Moreover, the sample size for each unit at each level should be large enough so that the variation can be partitioned between each contextual level. Recent literature has suggested that the household level is an important context to include in multilevel analyses of mental health. Advocates of the explicit modelling of this level are Weich et al, who have published three papers in the past four years calling attention to the importance of including household as an important level to include in analyses of this type (Chandola et al., 2003, 2005; Propper et al., 2005; Weich et al., 2003b, 2006, 2005). There is a potential problem, however, with obtaining sufficient responses from each household unit so that the household contextual effect can be separated from the other levels. A related problem involves the effect on the results of a multilevel analysis of either ignoring or including household as a level under low response conditions.

There is also benefit to be gained from utilising a Bayesian approach. With a Bayesian approach, information from other studies can be incorporated into the analysis. This is appropriate if a given parameter is particularly well understood or studied. The knowledge about this parameter can be explicitly modelled in a Bayesian framework, a distinctly different approach to classical analysis. In classical analysis, accumulated knowledge from previous studies is ignored (except perhaps to inform the choice of model to fit) and the estimates produced depend solely on the current data. Proponents of Bayesian analysis would say that this is akin to reinventing the wheel with each study. Opponents of Bayesian analysis would accuse the selection of prior information of being too subjective and non-scientific. Both approaches, used responsibly, have their uses in different settings. Aside from this, recent advances in the implementation of Monte Carlo Markov Chain (MCMC) methods, have led to the development of software that can tackle complex Bayesian analyses. The flexibility of this method, and the ease of its implementation, make it a very attractive solution. Bayesian models incorporating minimal prior information will be used in this thesis to examine the spatial distribution of mental health in Caerphilly.

1.3 Motivation for thesis

In 2001 the Caerphilly Health and Social Needs Study (CHSNS) survey (Fone, 2005) was carried out (described in chapter 2). Information on a random sample of the residents of Caerphilly county borough was collected providing a large dataset. This dataset was used to investigate various hypotheses and a number of publications ensued (Fone & Dunstan, 2006; Fone et al., 2006b, 2007b,a,c). The CHSNS was also used as the basis for an MD thesis (Fone, 2005). During the analysis of the CHSNS dataset a number of methodological issues became apparent. The inconclusive literature on area effects on health provided further motivation for the thesis. These issues will be

described in the next section.

1.4 Overall Objectives

The objectives of this thesis are as follows:

1. To investigate properties surrounding the distribution of the mental health score used in the study, as well as evaluating a cutpoint to identify cases of common mental disorder.

2. To investigate the spatial variation of common mental disorders in Caerphilly county borough in a Bayesian framework.

3. To investigate the robustness of multilevel modelling techniques to sparse levels of data.

4. To develop an algorithm that can partition an area into internally homogenous areas, using data from the Caerphilly Health and Social Needs Study as an example.

5. To compare (quantitatively) the operationalisations of area for both administrative and synthetic boundaries

The first objective is to examine the properties of the measure of mental health employed in the study in order to investigate the most appropriate way to model it. Various approaches including Normal modelling, transforming the scale, ordinal modelling and dichotomising the scale will be considered. Attention will also be given to the problem of identifying a cutpoint on the scale to identify common mental disorders.

The second objective involves using Bayesian modelling in order to examine the spatial variation of mental health in Caerphilly county borough. The Bayesian modelling framework is very flexible and it will be used to implement the Besag, York and Mollie model (introduced in chapter 4). This allows spatial dependence to be incorporated into the model. Bayesian smoothing will be used to provide a clearer picture of the spatial distribution of mental health status in Caerphilly county borough.

The third objective concerns hierarchical modelling itself. Undoubtedly a useful tool, there remain concerns about how it is used in practical situations. Motivated by the problem of low response households, this objective concerns assessing the impact of including or excluding a sparse level of data on the results of a multilevel analysis.

The fourth objective is to construct an algorithm which, using information about the composition and geography of Caerphilly, can create new area boundaries which

partition the borough into internally homogenous, distinct, contiguous regions. Henceforth, these new areas will be referred to as "synthetic" areas. The algorithm will be developed using the CHSNS dataset, but will also be generalisable so that potentially, it could be applied to any area and any outcome.

The fifth objective is to compare the operationalisations of area for both administrative and synthetic boundaries. The data will be analysed firstly using the administrative boundaries, and then with the synthetic boundaries. The models fitted will be exactly the same except for the hierarchies employed.

1.5 Structure of thesis

Chapter 2 will focus on the data itself. The dataset was collected as part of the Caerphilly Health and Social Needs Study in 2001 (Fone, 2005). A brief summary of the background of this study will be presented. The dataset will be described in terms of the response rate, variables collected and scores calculated. Finally a critique of the study will be undertaken, indicating both the strengths and weaknesses of the dataset.

Chapter 3 will be concerned with the measure of mental health used in the study. Clearly mental health is a highly complicated variable to measure and model. There is no universally accepted measure to quantify the quality of a person's mental health. The measure used in the CHSNS was the Short Form 36 (SF-36). This instrument can be used to assess the general health status of a person. It includes a mental health specific scale (the Mental Health Inventory (MHI-5)) which focuses on measuring mental health. The MHI-5 is a useful and suitable tool for quantifying the mental health of a population. Unfortunately, as shall be seen, the resulting variable is negatively skewed which causes a problem for statistical analysis. Other studies which use this mental health score ignore this problem (Wainwright & Surtees, 2003; Drukker & van Os, 2003; Skapinakis et al., 2005) and even the developers of the scale recommend using a z-transformation (subtracting the mean and dividing by the standard deviation) and treating it as being Normally distributed (Ware et al., 2000a). Various approaches to dealing with this issue will be presented in chapter 3. These include transforming the scale, dichotomising the scale and utilising ordinal regression.

Bayesian analysis will be introduced in chapter 4. This chapter will then use Bayesian analysis to investigate the issue of spatial dependence of mental health status in Caerphilly county borough. Areas that are close to each other geographically are likely to be similar in other respects, such as access to green areas, availability of services, and infrastructure. This is likely to induce spatial correlation. A model developed by Besag, York and Mollié (1991) includes terms to separate such spatial variation from unstructured variation. Their model uses information about which areas are adjacent, and can therefore distinguish between patterns of variation that are

spatially linked and ones that are not. The adjacency information is also used to "borrow strength" from neighbouring areas to provide more reliable estimates. This is particularly useful if an area has a low number of respondents. In such a situation, information from the surrounding areas can be incorporated into the crude estimates, with the effect of drawing extreme values toward the mean.

In Chapter 5 the main statistical method of the thesis, hierarchical modelling (also called multi-level modelling), will be introduced. Features, parameters and diagnostic tests associated with standard hierarchical models will be described. The benefits of hierarchical modelling will also be explained. All of this will then be illustrated using an example from the CHSNS dataset.

Recent studies advocate (either implicitly or explicitly) the inclusion of the household-level in multilevel studies of mental health (Chandola et al., 2003, 2005; Propper et al., 2005; Weich et al., 2003b, 2006, 2005). However, very few studies have information regarding household-level. Failure to account for this extra level of clustering can lead to variation being attributed to higher or lower levels, thus inflating their significance (Moerbeek, 2004). The CHSNS dataset does include such information, and so the effect of including the household-level can be assessed. There are, however, other issues of a methodological nature surrounding the use of the household-level. The most frequent number of responses from a household is just one (over 90% of the households in the dataset are single response households). This limits the power available to attribute variation to the household-level since in over 90% of the households, the household effect cannot be separated from the individual effect. This problem is not unique to the Caerphilly dataset. The papers by Weich et al (2003a; 2003b; 2005) are based on datasets where at least 37% of the households must be single response (under the extremely conservative assumption that no household submitted more than two responses). Simulated datasets will be used in chapter 6 to investigate how much of a problem this might prove to be.

An individual's health is clearly related to their own individual characteristics, e.g. age, gender, employment status and social class. There is a belief that an individual's health (in particular mental health) is also associated with some exposure related to where they live. This poses the question: how do you define where people live? Is it their postcode, their enumeration district, their electoral ward? What constitutes a neighbourhood? The majority of published studies of mental health and context have used administratively defined areas to act as proxies for neighbourhoods. There are a number of reasons to use administrative areas in studies of this type. Firstly, the use of administrative boundaries is very straightforward. They are already delineated and no effort needs to be made to define them. Also, the use of administrative boundaries allows for comparisons between studies to be made more easily. Finally, there is often no choice but to use them, as datasets are routinely aggregated up from individual-level to

administrative area-level for confidentiality purposes. There are problems with this approach however. There is no reason to believe that administrative areas are adequate proxies for neighbourhoods, other than the fact that people who are geographically close are grouped together. There is no guarantee that people from widely different socio-economic backgrounds will not end up in the same grouping. This becomes an issue when aggregate statistics are used as summary measures for areas. In chapter 7 neighbourhoods will be defined using the concept of spatially distributed attributes (e.g. social class) so that "synthetic" boundaries can be created which attempt to group similar people together. There is still no guarantee that the residents of these areas would identify with their new boundaries. However, the composition of these new areas will be more homogenous than the administrative areas. It is hoped that these homogenous "synthetic" boundaries will be able to elucidate the link between area of residence and mental health better than the administrative boundaries. The operationalisations of neighbourhood produced by both administrative and synthetic boundaries will be compared. Both types of boundary will be compared using summary statistics relating to internal homogeneity in order to compare the two types of boundaries. Moreover, hierarchical models will be fitted using both sets of boundaries in order to assess the relationship between the internal homogeneity of the boundaries used and the results of fitting a hierarchical model.

Chapter 8 incorporates all of the results from the previous chapters and uses them to investigate whether individual mental health is associated with area of residence. This investigation will be compared with a previous analysis of the data.

Chapter 9 will summarise the results of the thesis in relation to the objectives set out in this chapter, outline the practical implications that this research might have in a research setting and indicate some areas that would benefit from further research.

Chapter 2

The Caerphilly Health and Social Needs Study

2.1 Geography of Caerphilly county borough

Located in South Wales (figure 2.1 reproduced from Ordnance Survey map data by permission of the Ordnance Survey © Crown copyright 2001) Caerphilly county borough has a 2001 census population of 169,519. The borough contains about fifty towns and villages, the largest being Caerphilly town itself (population $\sim$ 28,000). Other significant settlements include Bargoed, Blackwood, Newbridge, Risca and Ystrad Mynach (figure 2.2). In the 1950's it was a thriving industrial borough with 29 operational coal pits providing employment for 24,000 people. The last pit was closed in 1990. The decline of the coal industry had a dramatic effect on the area. Today, a generation on, the people who live there still suffer from raised rates of unemployment and poverty. Indeed two census wards in the Upper Rhymney Valley in the north of the borough are in the most deprived 5% of wards in England and Wales (Moriah and Twyn Carno) (Glennerster et al., 1999). The borough itself encompasses a large variety of socio-economic backgrounds. The southern parts of the borough are generally the most affluent, housing many commuters who work in Cardiff. The northern parts of the borough are generally more deprived than the southern parts as they were more dependent on the mining industry and thus more affected by its collapse. Since devolution in Wales, government policy has focussed on identifying areas for improvement. An example of this is the "Communities First" programme which targeted the 100 most deprived wards in Wales (of the 865 wards in the 1998 boundary revision) (National Assembly for Wales, 2001). These were identified by ranking the Welsh Index of Multiple Deprivation (National Assembly for Wales, 2000). Thirteen of Caerphilly county borough's 36 wards were included in these 100 most deprived (New Tredegar, Tir Phil, Darran Valley, Aberbargoed, Bargoed, Hengoed, Gilfach, Twyn Carno, Pontlottyn,

Figure 2.1: Map of the UK showing Wales, Gwent and Caerphilly county borough

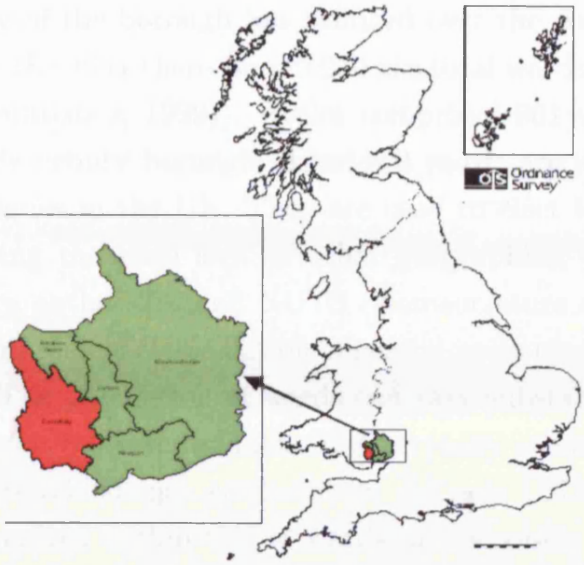


Figure 2.2: Map of Caerphilly county borough showing geographical features

Moriah, Abertysswg, Argoed and Aber Valley). In 1998 Caerphilly county borough council and the former Gwent Health Authority collaborated to develop a strategic programme to improve health in the borough. One part of this was the Caerphilly Health and Social Needs Study (Fone et al., 2002; Fone, 2005).

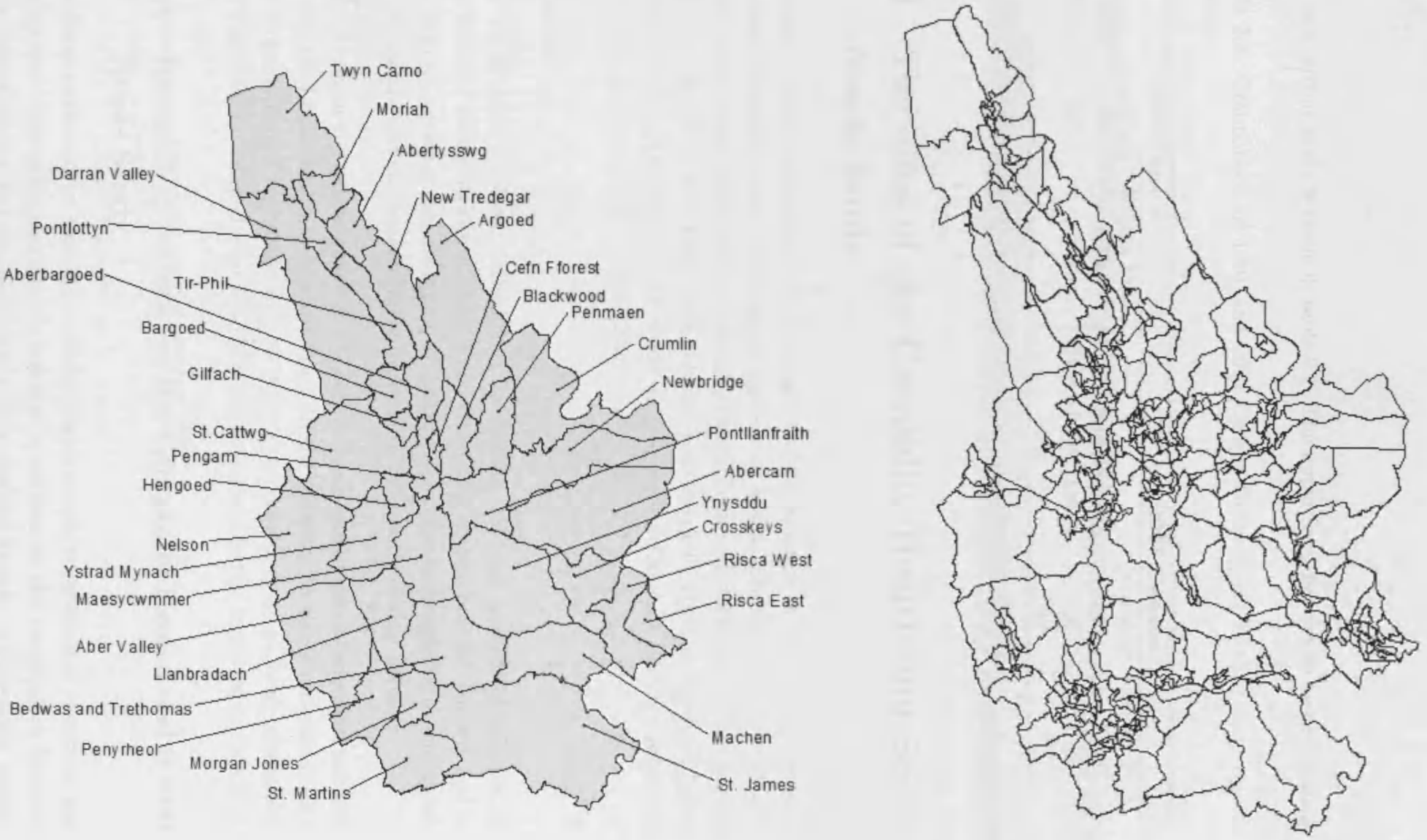
2.1.1 1991 Census geography

The census geography of the borough has changed over the past decade, in line with the rest of the UK. In the 1991 there were 9,930 electoral wards in England and Wales (Office for National Statistics, 1999)). Wales comprised 901 wards and of these, 36 belonged to Caerphilly county borough. Electoral wards are a key building block of administrative boundaries in the UK. They are used to elect local government councillors, as well as being the base unit of other geographies, such as parliamentary constituencies, unitary authorities and NUTS (Nomenclature of Units for Territorial Statistics) areas (a hierarchical classification of spatial units used for statistical production across the EU). The population of wards can vary substantially, but the average for England and Wales is 5,500 individuals (in Caerphilly county borough it is 4,700). These wards were subdivided into enumeration districts (EDs) in the 1991 census. There were 325 EDs nested within the 36 wards in Caerphilly county borough. The average population of an ED in the study area is about 520 individuals. Boundaries for both electoral wards and enumeration districts are shown in figure 2.3.

2.1.2 2001 Census geography

The 2001 Census introduced new administrative boundaries, in particular output areas. Output areas were generated for the 2001 census (in England and Wales) by combining adjacent postcodes (University of Southampton, 2000). They were designed to have similar population size (on average 300 residents). They are also designed to take into account "measures of population size, mutual proximity and social homogeneity" (Office for National Statistics, 2006b; Vickers & Rees, 2007). Homogeneity is based on the nature of tenure of household and type of dwelling. Wherever possible, OAs do not straddle urban/rural boundaries and their boundaries frequently take account of obvious geographical features, such as main roads. The minimum number of households permitted in an OA is 40 (with 100 resident individuals); however, the recommended size is 125 households. There are 9,769 OAs in Wales and 559 in Caerphilly county borough. Super Output Areas (SOAs) are built up from OAs and were designed to replace electoral wards for the purposes of presenting statistical information. As mentioned above the population of electoral wards can vary substantially. This makes them unsuitable for nationwide comparisons as well as causing confidentiality problems, when data from smaller wards cannot be released (Office for National Statistics, 2006a). The SOAs are defined by population size and so avoid this problem. There are three SOA layers. The Lower Super Output Layer (LSOA)has a minimum population of 1,000 and mean of 1,500. Typically they consist of 4 to 6 merged OAs. The middle layer is slightly bigger containing at least 5,000 individuals with an average of 7,200 individuals. These are built up from lower layer SOAs. A comparison between the old

Figure 2.3: Map of Caerphilly county borough, showing Ward and Enumeration District Boundaries



and new output areas is made in table 2.1. The upper layer has yet to be determined.

Table 2.1: Comparison of 1991 and 2001 Census Geographies in Caerphilly county borough

1991 Census			2001 Census		
	Number	**Mean Pop.**		**Number**	**Mean Pop.**
Wards	36	4,700	SOAs-Middle Layer	24	7,000
EDs	325	520	SOAs-Lower Layer	110	1,500
			OAs	559	300

2.2 The aims of the Caerphilly Health and Social Needs Study

In order to gather information on the health and social needs of the borough population and to inform a new public health agenda on health inequalities, Gwent Health Authority and Caerphilly county borough council made commitments to partnership working in the autumn of 1998. This initiated the Caerphilly Health and Social Needs Study (CHSNS). The aims of the study as proposed in October 1998 are presented here in the form reported by Fone (2005).

Aims:

- To establish a robust methodology for sharing and joint analysis of information between Gwent Health Authority and Caerphilly county borough council, and to achieve a greater understanding of the relations between health status and social, economic and environmental deprivation in Caerphilly county borough;
- To inform the development of the health needs assessment information required by the Local Health Group and Local Health Alliance for developing the Health Improvement Programme and to inform the development of local community regeneration strategies for health improvement and better targeting of resources.

2.2.1 Research questions for the Caerphilly Health and Social Needs Study

In addition to the service aims stated above, the academic component of the study was developed to investigate two specific research questions on the associations between mental health status and compositional and contextual factors in Caerphilly county

borough, and gain a greater understanding of the relations between mental health, people and places (Fone, 2005). The specific research questions to be addressed were listed as follows.

1. Is individual mental health status associated with factors that measure socioeconomic deprivation, urban/rural status and social capital at contextual level (place) after adjusting for the composition (people) of these places?

2. Do any associations between mental health and contextual factors vary (a) between population groups, and (b) with the size of geographical areas?

2.3 Caerphilly Health and Social Needs Survey dataset

In order to address these research questions a population-based survey was carried out in 2001. This thesis will make use of the data resulting from this survey. It was a population-based questionnaire survey of residents of Caerphilly county borough, aged 18 and over. The sampling frame was a stratified (by electoral ward) random sample of the 132,613 adult residents as identified by the GP administrative age-sex register. A commercial company called Beaufort Research, Cardiff carried out the survey. A sample size calculation was performed (Fone, 2005) giving a target of 12,600 responses. Assuming a 60% response rate, this translated to a total sample size of 22,290. At the start of the study 22,236 questionnaires were posted out to residents of the borough. Of these 2,267 were reported to have moved away, 84 had died or were too ill to complete the survey and 98 were living in nursing homes and were excluded. This reduces the number of people sampled to 19,787. From these, 12,408 completed questionnaires were received giving a response rate of 62.7%. Figure 2.4 below summarizes the survey response. Excluding the over 75s, those with incomplete mental health information, and 316 individuals who had defaced the identifying barcode on the questionnaire and so had missing geographical information, the final sample size was reduced to 10,653.

The Caerphilly Health and Social Needs Survey produced a number of publications, disseminating the results of the study (Fone et al., 2006b, 2007c,b,a; Fone, 2005).

Mental health was assessed using the MHI-5 of the SF-36 Version 2 (Ware et al., 2000a), which will be examined in detail in chapter 3. A large amount of information was collected from each respondent including age, gender, social class, employment status, housing tenure and gross household income. This information was augmented with area-level (or contextual) information from other data sources (Fone et al., 2002). The main sources of information were the Department of Work and Pensions (DWP), the Office for National Statistics (ONS), and the Paycheck dataset (a commercially available dataset used to estimate gross household income at area levels) (CACI, 1999).

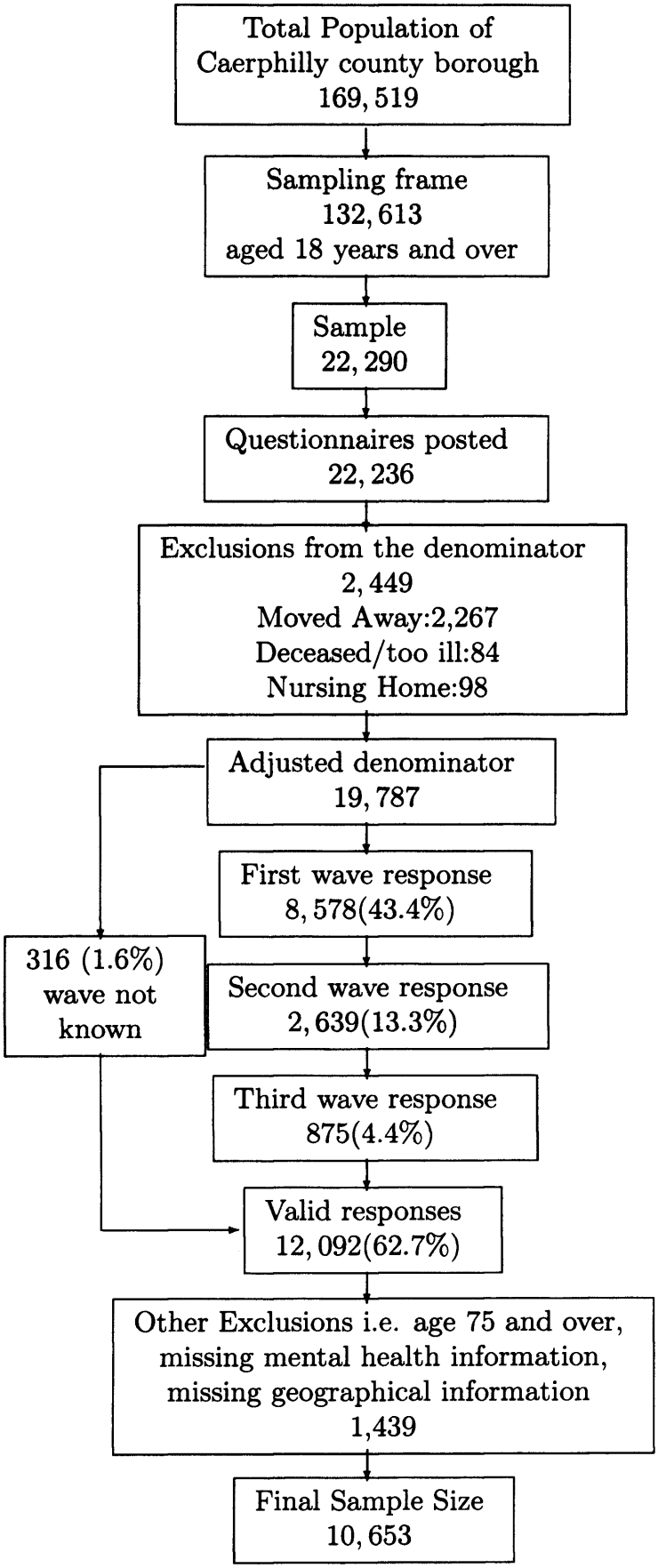


Figure 2.4: Flowchart of the Caerphilly Health and Social Needs Study

The DWP provided information on both means-tested and non-means-tested benefits (Fone et al., 2007c). The ONS census data from 2001 was used to calculate an index of deprivation called the Townsend Index (described in the next section). The Council Tax and Benefits division of the Caerphilly county borough council supplied the 2002 (collected in February) council tax register, containing information on the council tax valuation band of each property in the borough (Fone, 2005; Fone et al., 2006b). Caerphilly county borough also provided information regarding the proportion of unemployed in the borough.

The next section will focus attention on the variables of interest for this thesis. These will be categorised according to whether they are measured at individual, household or area level. The outcome measure is the mental health score which will be described in detail in chapter 3.

2.3.1 Description of Variables

Individual-level variables

The gender of the respondents is tabulated in table 2.2. More women than men responded, with over 55% of the dataset being female.

Table 2.2: Survey response by gender and five year age group

Age category	Female	%	Male	%	Total	(%)
18-24	503	(4.7)	368	(3.5)	871	(8.2)
25-29	458	(4.3)	284	(2.7)	742	(7.0)
30-34	574	(5.4)	395	(3.7)	969	(9.1)
35-39	618	(5.8)	447	(4.2)	1,065	(10.0)
40-44	587	(5.5)	447	(4.2)	1,034	(9.7)
45-49	612	(5.7)	456	(4.3)	1,068	(10.0)
50-54	627	(5.9)	602	(5.7)	1,229	(11.5)
55-59	580	(5.4)	476	(4.5)	1,056	(9.9)
60-64	480	(4.5)	464	(4.4)	944	(8.9)
65-69	455	(4.3)	482	(4.5)	937	(8.8)
70-74	389	(3.7)	349	(3.3)	738	(6.9)
Total	5,883	(55.2)	4,770	(44.8)	10,653	(100.0)

An important variable is social class. This was based on the 1991 Standard Occupational Classification (ONS, 1991). It was derived from the following survey questions:

- Which best describes your situation? Employed (full time or part time)/ Unemployed and seeking work/ Looking after home or children full time/ Retired from paid work/ Long term carer/ Permanently unable to work due to illness or disability/ On a government training scheme

Table 2.3: Social Class Frequencies

Social Class	Frequency	(%)
I&II (Professional &Intermediate)	2,407	(22.6)
IIINM (Skilled Non-manual)	2,103	(19.7)
IIIM (Skilled Manual)	2,171	(20.4)
IV&V (Semi-skilled & Unskilled)	2,647	(24.8)
Other	635	(6.0)
Missing	690	(6.5)
Total	10,653	(100.0)

- In your present or most recent job, are (were) you: A manager/ A foreman or supervisor/ Self employed (with employees)/ Self employed (without employees)/ I have never been in paid employment
- What is your job title (if you are not in work state what your previous title was)? (Please answer this question even if you are not working now)
- Industry sector/ field of employment
- Main things done in job

The last three questions were open response questions. The final social class variable used in the study was split into six categories (chosen to divide those respondents in social classes I to V into approximately equal numbers). The divisions were class I & II, III non-manual, III manual, IV & V, Missing and Other (this category contained the armed forces, full time education, youth training scheme, housewife or carer at home, not working due to disability and unemployed-never worked). In all models social class I & II will be the reference category. Table 2.3 shows the numbers of respondents in each category.

Employment status was recorded as one of the categories listed in table 2.4. Note the high proportion of people who classify themselves as "Permanently unable to work due to illness or disability". This compares with a UK average of 5.5% (based on the age group 16-74) (Office for National Statistics, 2003). In all models the employed category will be the reference category. Official Caerphilly county borough council records showed the unemployment rate to be 2.9% (95% CI: 2.4% to 3.6%), which is not significantly different to the reported rate of 2.7% (Fone, 2005).

The level of education of respondents was also investigated in the questionnaire. The multiple choice question "What is your highest educational qualification?" was used for this purpose. A summary of the survey responses is provided in table 2.5.

Table 2.4: Employment Status Frequencies

Employment Status	Frequency	(%)
Employed (full or part time)	5,507	(51.7)
Unemployed and seeking work	286	(2.7)
Full time student/school/government training scheme	190	(1.8)
Looking after home or children fulltime/Long term carer	804	(7.5)
Permanently unable to work due to illness or disability	1,274	(12.0)
Retired from paid work	2,111	(19.8)
Missing	481	(4.5)
Total	10,653	(100.0)

There is a large percentage of individuals with no educational achievement in the dataset.

Table 2.5: Educational Achievement of Respondents

Educational Level	Frequency	%
Degree/Professional/NVQ Levels 4 or 5	1,385	(13.0)
HNC/HND	421	(4.0)
A Level/Advanced GNVQ/NVQ Level 3	946	(8.9)
School certificate/City & Guilds	947	(8.9)
O Level/GCSE A*- C/GNVQ/NVQ Level 2	1,785	(16.8)
O Level D-E/GCSE D-G/GNVQ/NVQ Level 1	481	(4.5)
No educational qualifications	3,708	(34.8)
Missing	980	(9.2)
Total	10,653	(100.0)

Household-level variables

As will be described in chapter 6, the household-level has been identified as being an important context to include in mental health studies (Chandola et al., 2003, 2005; Propper et al., 2005; Weich et al., 2003b, 2005). The household variables used in this thesis will now be described.

In the interests of minimising non-response the question "What is your total current gross weekly or yearly household income?", was framed as a multiple response question. This avoids the problem of individuals having to divulge too much specific information about their income. The three possible choices are listed in table 2.6. Those with a gross income lower than £215 per week after housing costs (i.e. less than 60% of the UK median gross income) are classified as living in "poverty" according to the

UK definition (Office for National Statistics, 2004). It is worth noting that under this classification at least 45% of the respondents are living in poverty. This income information was aggregated to both ward and ED level and compared to area-level income data from Paycheck (CACI, 1999) which provided evidence for the validity of the income information (Fone, 2005). In all models the £95-£215 per week/£5,000-£11,250 per year category is the reference category.

Table 2.6: Self-reported gross household income frequencies

Self-reported Income	**Frequency**	**(%)**
Less than £95 per week/Less than £5,000 per year	960	(9.0)
£95-£215 per week/£5,000-£11,250 per year	3,810	(35.8)
More than £215 per week/More than £11,250 per year	5,158	(48.4)
Missing	725	(6.8)
Total	10,653	(100.0)

Council tax valuation band (CTVB) information on each residential property in the borough was obtained from the council tax register and matched to the sampling frame (Fone, 2005; Fone et al., 2006b). There are eight different CTVBs, which are labelled A to H. These provide a measure of residential property value. Since the numbers of houses in the higher bands are relatively small, this variable was dichotomised for many of the analyses. The first two categories (A and B) were combined and compared to the last six (C-H). Table 2.7 summarises the council tax band information for individuals (not properties) in the CHSNS dataset. In all models category C-H is the reference category.

Housing tenure was recorded as one of four categories: "I own it or live with the

Table 2.7: Council Tax Valuation Bands

CTVB	**Property Value**	**Frequency**	**(%)**
A	< £30,000	2,326	(24.3)
B	£30,001-£39,000	3,988	(37.4)
C	£39,001-£51,000	1,677	(15.7)
D	£51,001-£66,000	862	(8.4)
E	£66,001-£90,000	487	(4.6)
F	£90,001-£120,000	193	(1.8)
G	£120,001-£240,000	41	(0.4)
H	> £240,001	2	(0.0)
Missing		1,077	(10.1)
Total		10,653	(100.0)

Table 2.8: Housing Tenure

Tenure	Frequency	(%)
Owner Occupied	8,562	(80.4)
Not Owner Occupied	1,943	(18.2)
Missing	148	(1.4)
Total	10,653	(100.0)

person who owns it", "Rented from local council or housing association/trust", "Rented from a private landlord" and "Other (e.g. live rent free or home comes with job". For the analysis however, this was simplified by merging the last three categories, as in table 2.8. In all models the owner occupied category is the reference category.

Area-level variables

To investigate possible associations between area of residence and mental health, contextual measures need to be recorded. These variables can then be included at the correct level of analysis and their influence assessed. In order to investigate the contextual impact of deprivation, two area-level deprivation measures were included in the analysis. The Townsend social and material deprivation score (Townsend et al., 1988) was calculated from Census small area statistics using unemployment, car ownership, owner occupation and overcrowding information. It was calculated for both ward and enumeration district levels. The 2001 census was used to calculate the ward-level score, but since the enumeration district level was not used in the 2001 census, the 1991 census was used for the ED level score. Table 2.9 shows the Townsend scores for wards in Caerphilly county borough split into quintiles. The four constituent Townsend variables are each standardised to have zero mean and standard deviation of 1. These four variables are then summed, producing a score with a mean of 0, and a standard deviation of 4 (Townsend et al., 1988). Higher scores indicate more deprivation. The most deprived ward is Aberbargoed, with a Townsend index of 6.75, (followed closely by Twyn Carno with 6.58), while the least deprived is St. Martins with a score of -2.99.

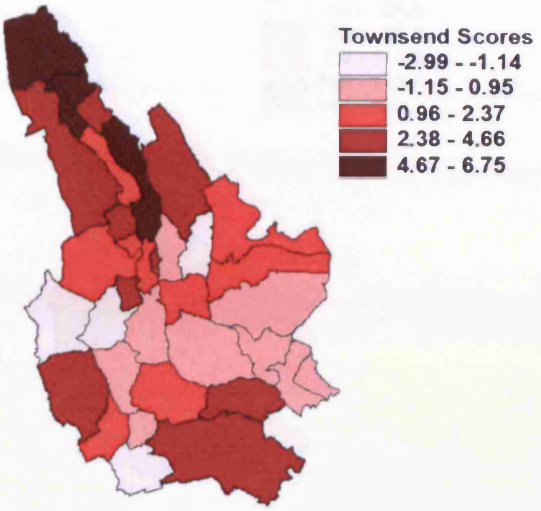
The Department of Work and Pensions (DWP) provided information on benefits data for the adult working age population of Caerphilly county borough. This comprised the number of claimants of long term Incapacity Benefit and Severe Disablement Allowance residing in the borough in August 2001. Individuals aged between 16 and state pension age, who had been on statutory sick pay for 28 weeks or more, and who had made sufficient National Insurance contributions were eligible for Incapacity Benefit. Severe Disablement Allowance was paid to individuals who had never been able to

Table 2.9: Townsend Scores by Ward

Ward Name	Townsend Score	Ward Name	Townsend Score
Aber Valley	4.26	Moriah	5.61
Aberbargoed	6.75	Nelson	−1.14
Abercarn	0.37	New Tredegar	5.65
Abertysswg	3.6	Newbridge	1.14
Argoed	4.46	Pengam	1.54
Bargoed	3.89	Penmaen	−1.47
Bedwas and Trethomas	1.37	Penyrheol	1.5
Blackwood	0.32	Pontllanfraith	1.98
Cefn Fforest	2.92	Pontlottyn	5.56
Crosskeys	−0.12	Risca East	0.68
Crumlin	1.28	Risca West	0.27
Darran Valley	4.6	St. Cattwg	2.32
Gilfach	1.26	St. James	4.66
Hengoed	3.65	St. Martins	−2.99
Llanbradach	0.66	Tir-Phil	2.37
Machen	2.94	Twyn Carno	6.58
Maesycwmmer	−0.26	Ynysddu	0.76
Morgan Jones	0.95	Ystrad Mynach	−2.18

work, or who did not meet the eligibility criteria for Incapacity Benefit. These benefits, taken together, provide an estimate of the total working age population classified as incapable of work. In August 2001, 17,493 (15%) of the 116,990 working age residents of Caerphilly were claiming these benefits. This information was supplied for persons anonymised in five-year age ranges and was aggregated to ward level. No gender information was supplied. Indirect age-standardised ward ratios were calculated and are referred to as the Incapacity Claimant Ratio (ICR). These were multiplied by 100, so that 100 indicates the overall borough average incapacity for work. This information is summarised in table 2.10 and illustrated in figure 2.6. Again, there is considerable variation throughout the borough, ranging from areas with less than half the expected number of claimants (St. Martins), to areas with 60% more claimants than average (Aberbargoed). There is a clear gradient of increasing proportions of claimants from the southern part of the borough to the northern part. This matches the pattern of deprivation seen in figure 2.5.

Figure 2.5: Townsend Scores, split into quintiles, by Ward



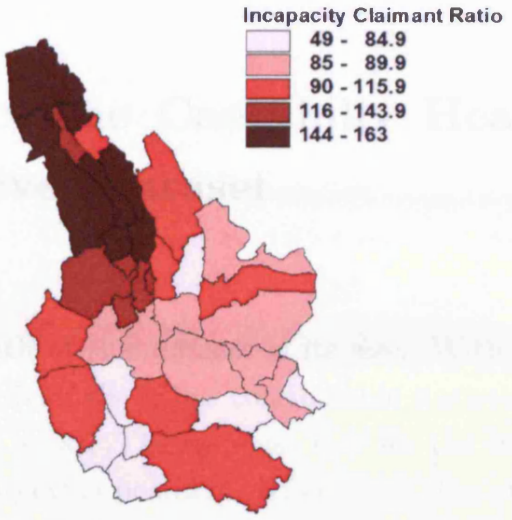
2.4 Analysis of Survey Response

The response to the Caerphilly Health and Social Needs Survey was extensively investigated for evidence of differential response (Fone, 2005). The age of respondents was cross-validated with information on the sampling frame to determine whether the

Table 2.10: Incapacity claimant ratios by ward

Ward Name	ICR	Ward Name	ICR
Aber Valley	109.54	Moriah	144.88
Aberbargoed	162.94	Nelson	94.04
Abercarn	89.46	New Tredegar	147.48
Abertysswg	115.25	Newbridge	91.90
Argoed	124.29	Pengam	117.07
Bargoed	147.89	Penmaen	68.98
Bedwas and Trethomas	96.32	Penyrheol	79.21
Blackwood	90.29	Pontllanfraith	86.84
Cefn Fforest	116.57	Pontlottyn	143.25
Crosskeys	85.94	Risca East	73.08
Crumlin	89.69	Risca West	88.18
Darran Valley	148.77	St. Cattwg	120.44
Gilfach	126.03	St. James	102.50
Hengoed	143.50	St. Martins	49.32
Llanbradach	82.97	Tir-Phil	147.21
Machen	80.20	Twyn Carno	152.28
Maesycwmmer	89.70	Ynysddu	86.11
Morgan Jones	73.95	Ystrad Mynach	86.47

Figure 2.6: Incapacity Claimant Ratio by Ward



sample was representative of the borough. Work was done to investigate the response rate by council tax band and deprivation also (Fone, 2005).

2.4.1 Exclusions from the denominator

As illustrated in the flowchart in figure 2.4, 2,449 (11%) of the original sample were excluded from the denominator for various reasons. Of these, 1,570 were male and 879 female. At every age group except for the over 80s there was a greater percentage of males excluded than females.

2.4.2 Response Rate

The response rate for women was better than for men with 6,887 females responses (constituting a response rate of 66.8% for females). There were 5,519 male response, representing a response rate of 58.3%. This was also broken down by five year age groups and it was found that women had higher response rates than men up to (and including) the age group 60-64, but after that men had higher response rates.

It was also found that there was an under-representation of those under 45 years old and an over representation of those over 45 years old. This was reflected in the fact that the mean age of responders was 50.5 while the mean age of non-responders was 43.8.

The relationship between response rate and age group (split by gender) was not found to vary substantially between the three waves of data collection, with males reporting response rates of 39.9%, 13.2% and 4.5% for the first, second, and third waves. Females had response rates of 46.5%, 13.5% and 4.4% for the first, second, and third

waves. The differences between waves were not found to vary substantially between the deprivation of the ward, as measured by the Townsend Index of Deprivation (Townsend et al., 1988).

2.5 Critique of the Caerphilly Health and Social Needs Survey dataset

2.5.1 Strengths

The most striking strength of this dataset is its size. With complete mental health information on 10,653 individuals, it has considerable statistical power to detect even small differences between areas. The response rate for the study was consistent with (if not slightly better than) other postal questionnaires. The study achieved a response rate of 62.7%. This is typical of the response rates from postal questionnaire surveys, and is in fact better than the response rate of 57.7% achieved by the Welsh Health Survey in Caerphilly county borough (The National Assembly for Wales, 1999).

Secondly, Caerphilly county borough is a meaningful study area, in that it is a well defined contiguous region. Conclusions drawn from the study have a practical and useful interpretation. This is in contrast with other studies which sample people at random from the entire UK, with small numbers of people from each area sampled. Cummins et al (2005) discuss this issue, concluding:

> "*Our approach has shown that 'true' area data on social and material context is complex and difficult to collect at national level and produced caveats in terms of data comparability, generation and interpretation.*"

Thirdly, the multi-agency nature of the Caerphilly Health and Social Needs study facilitated data linkage between diverse datasets. Of particular importance was the linkage of the Caerphilly questionnaire data with data from the Office for National Statistics. This provided census based deprivation indices. Information from the Department of Work and Pensions (DWP) provided information on incapacity claimants in the borough. Income data was provided by Paycheck (a commercial available dataset) (CACI, 1999). Add to this the fact that the Caerphilly Health and Social Needs questionnaire recorded a large amount of information about each respondent and the resultant dataset has a richness and depth unusual for a study of this type. In particular the level to which individuals can be geographically pinpointed is quite good, with information down to postcode and household for each individual, which facilitated the linkage of the questionnaire data with other datasets.

A further benefit of the dataset being an amalgamation of information from different sources is that it helps avoid same source bias. This is useful for the DWP benefits

information. Area based variables derived simply by aggregating individual-level survey responses are not objective measures of an area (e.g. percentage unemployment). It could be the case that people with poor mental health, tend to exaggerate (or understate) their incapacity to work. This could result in a stronger (or weaker, respectively) association between being incapable of work due to disability and mental health being reported. This same source bias problem is avoided because the incapacity benefits data was obtained from a different dataset.

The quality of the dataset is reflected in the number of original publications it has produced. These include an analysis investigating the utility of local authority data for investigating health inequalities (Fone et al., 2002), the association between council tax valuation bands and socio-economic status and a number of health outcomes (Fone et al., 2006b), the association between neighbourhood cohesion and individual mental health status (Fone et al., 2006c), the effect of perceived and geographical access to accident and emergency departments (Fone et al., 2006a), the effect of community and individual level deprivation and social cohesion on mental health (Fone et al., 2007b), an investigation of the relationship between the Mental Health Needs Index (MINI) (Glover et al., 1998), the common mental disorders (Fone et al., 2007a), and a multilevel analysis of economic inactivity (Fone et al., 2007c) as well as providing the data for an MD thesis (Fone, 2005).

2.5.2 Weaknesses

Firstly, weaknesses of the study which are not addressed in this thesis are presented. The biggest weakness of the study is that it is cross-sectional in nature. One of the suggestions for future research recommended by Kawachi and Berkman (2003) is to disentangle the effects of social selection versus social causation. Both social selection and social causation are theories that attempt to explain why certain attributes differ between areas. Basically, this can be summarised with the question "do poor people choose to live in poor areas or do poor areas contribute to making their residents poorer?". It might seem reasonable to suggest that both mechanisms operate simultaneously in most areas, or to quote Macintyre (1993) "People create places and places create people". To study such mechanisms (and quantify their relative importance) a longitudinal study is needed. Such a study would be able to make causal inferences, and provide insight into which of these two mechanisms is most potent. The answers to such questions are central to obtaining a deeper understanding of the relationship between mental health and place. Unfortunately, since these data are cross-sectional in nature, there is no possibility of any such kind of inference being made. At best, such data can provide evidence for an association between any two variables, say deprivation and mental health. It can not provide any information as to which came first however.

Another potential limitation of this dataset is that despite there being a wealth of area-level geographical information there is no information on the geographical position of individuals within a postcode. For the purposes of using GIS techniques people are mapped as residing at the centroid of their postcode. It would have been preferable had it been possible to assign individuals from the same postcode into different synthetic boundaries. This is not possible in the CHSNS dataset.

The dataset provides a number of contextual variables for the purposes of multilevel modelling. Most of these variables however are not "true" contextual variables. What is meant by this is that they are calculated from aggregated individual variables, such as percentage unemployment. A "true" contextual variable is one that is only measurable at the higher levels. Examples would be the number of shops in an area, or the number of burnt out cars, or the percentage of available surfaces covered in graffiti. It has been suggested that more use should be made of such "true" contextual variables (Macintyre et al., 1993). While it is a pity that more of these contextual variables are not available, the same authors note that aggregated contextual variables contain different and important information about area characteristics.

There is an issue with regard to differential response rates from the survey. This is a problem endemic to postal questionnaire surveys, with low response rates known to be associated with younger age groups, male gender, individuals from lower social classes and less educated people (Etter & Perneger, 1997). Young males (aged under 45 years) in this study were the worst responders with a response rate of only 58%. However, similar studies which used the SF-36 achieved comparable rates from men: 51% (Avery et al., 1998), 64% (Department of Public Health Medicine, 1993), 72% (Bowling et al., 1999) and 58% (The National Assembly for Wales, 1999). The surveys with response rates of 64% and 72% were conducted in much more affluent areas than Caerphilly county borough (Oxford and Hertfordshire respectively). This would be a far greater problem if the goal of the study was to make population inferences about Caerphilly county borough (e.g. if the objective was to make population inference regarding the mental health of residents of Caerphilly county borough in order to ascertain the level of service provision were required for the area). Serious bias could be introduced by differential response rates if this were the case. However, since the work is focussed on identifying relationships between various exposure variables and mental health, the non-representative nature of the survey is not an impediment. Similarly, if the relationships between the explanatory variables and mental health differed between responders and non-responders this would be a potential problem. Depending on the nature of these differences they could bias the associations between mental health and the explanatory variables in either direction. It is not possible to investigate this in the CHSNS dataset as no mental health information is available on the non-respondents.

Next, weaknesses of the study which are addressed in this thesis are presented.

The outcome of interest for this thesis is mental health. This was measured using a component of the SF-36 questionnaire. This component is known as the Mental Health Inventory (MHI-5) and consists of just five questions. While this is a valid and reliable score (as shall be reviewed in chapter 3) it unfortunately does not have a clinically validated or meaningful cutpoint, indicating whether people are "cases" of CMD or not. Other mental health scales such as the General Health Questionnaire(GHQ) (Goldberg & Williams, 1988) do have such a cutpoint. Various approaches to dealing with this issue will be presented in chapter 3.

Another potential problem with the CHSNS dataset is the household level. None of the published analyses from the CHSNS used households as a level in a multilevel analysis. This has been highlighted as an important context to model. However, since the dataset contains a high proportion of single response households (84.8% of the responses are from individuals belonging to single response households), the household effect is practically inseparable from the individual level. The effect of including such a level is under-investigated as will be discussed in chapter 6. Here a simulation study investigates the possible impact of such identifiability problems.

Finally, a criticism of the study is that the only area boundaries available to locate respondents geographically are administrative census boundaries. These boundaries are not created for the purpose of modelling mental health, and so may not be suited to the task. From a statistical point of view boundaries which group similar people together are preferable to those which group heterogenous people together, since aggregate measures of heterogenous groups are rather less meaningful than those of homogenous groups. A separate issue is that the effect of changing the hierarchy on the results of a multilevel analysis is not known. Since administrative boundaries change regularly (different census geographies were used in the 1991 and 2001 censuses) this is potentially a large problem. This will addressed in chapter 7.

2.6 Conclusion

This chapter described and critiqued the CHSNS dataset. Many of the criticisms of the CHSNS dataset are common to many datasets which investigate area-effects on health. For instance, confidentiality considerations preclude recording individual's geographic location at any level lower than postcode. Similarly, questionnaire non-response for males and young people are widespread problems for postal questionnaire surveys. It was concluded that the Caerphilly Health and Social Needs Survey dataset is an excellent resource for studying area-level effects on mental health. The original publications from the study suffer from a number of methodological limitations however, and this thesis will address those limitations as well as augmenting the wider multilevel modelling methodology literature.

Chapter 3

Measurement of mental health status

In the first chapter of this thesis, background was provided on the public health importance of studying mental health in general, as well as the more specific problem of the common mental disorders, on which this thesis focusses. Before this can begin however, it is important to investigate the method by which common mental disorders will be measured. This chapter will address the second main objective outlined in the introduction, to investigate properties surrounding the distribution of the mental health score used in the study, as well as evaluating a cutpoint to identify cases of common mental disorder.

This will be done in the following sections.

1. The mental health scale used in the CHSNS, the mental health inventory (MHI-5) of the SF-36 version 2, will be described.

2. The concepts of validity and reliability will be introduced. Standard techniques to assess these concepts will be described.

3. Literature surrounding previous attempts to demonstrate the validity and reliability of the MHI-5 will be described and summarised.

4. Various approaches to modelling the mental health score produced by the MHI-5 will be described and illustrated. Emphasis will be placed on deriving a cutpoint to define a case of common mental disorder.

5. The results and conclusions of the chapter will be summarised.

3.1 Description of the SF-36

Understanding the epidemiology of population mental health relies heavily upon questionnaire measures to quantify the state of a person's mental health. The gold standard, of course, would be a clinically administered interview; however in most cases this is either too time-consuming, expensive or impractical. Even if it were feasible, a large number of clinicians would be required in order to assess sufficiently many subjects, which introduces concerns over inter-rater reliability. A properly designed self-administered questionnaire can overcome some of these concerns.

There are many questionnaires that measure mental health status, but only the General Health Questionnaire (GHQ-12) (Goldberg & Williams, 1988), the Mental Health Inventory (MHI-5) (which is embedded in the SF-36) (Ware et al., 2000a) and the mental health component score (MCS) (Ware et al., 2000a) of the SF-36 will be considered in this thesis. The mental health measure in the CHSNS dataset was the MHI-5, since this is included in the SF-36 scale. On the construction of the SF-36 the authors have said (Ware & Gandek, 1998):

> *"Much remains to be discovered about population health in terms of functional health and well-being, the relative burden of disease, and the relative benefits of alternative treatments. One reason for this has been the lack of practical measurement tools appropriate for widespread use across diverse populations. The SF-36 was constructed to provide a basis for such comparisons. The SF-36 was constructed to satisfy minimum psychometric standards necessary for group comparisons. The eight health concepts measured in the SF-36 were selected from dozens included in the Medical Outcomes Study (MOS) and represent the most frequently measured concepts in widely-used health surveys that have been shown to be affected by disease and treatment."*

It is designed then, as a tool for general population surveys and can be employed via self-administration, telephone administration, or administration during a personal interview. It has been used in Denmark, France, Germany, Italy, the Netherlands, Norway, Spain, Sweden, United Kingdom, United States, Singapore, China and Australia (Keller et al., 1998; Li et al., 2003; Thumboo et al., 2001; McCallum, 1995) and work has been done to validate the scale in each of them. The validation of the SF-36 will be investigated in section 3.2.

The SF-36 is divided into eight subscales. These are Physical Functioning, Role-Physical, Bodily Pain, General Health, Vitality, Social Functioning, Role-Emotional and Mental Health. The MHI-5 included in the SF-36 version 2 (the version of the SF-36 in the CHSNS dataset) consists of five questions summarised in table 3.1.

Table 3.1: The MHI-5 included in the SF-36 version 2

	How much of the time during the past four weeks	**Responses**	**Score**
1.	have you been a very nervous person?	all of the time	1
2.	have you felt so down in the dumps	most of the time	2
	that nothing could cheer you up?	some of the time	3
3.	have you felt downhearted and low?	a little of the time	4
		none of the time	5
4.	have you felt calm and cheerful?	all of the time	5
5.	have you been a happy person?	most of the time	4
		some of the time	3
		a little of the time	2
		none of the time	1

The five possible Likert scale responses which are "All of the time", "Most of the time", "Some of the time", "A little of the time", and "None of the time". These responses are coded as numbers from 1 to 5. Two of the questions are reverse coded (the fourth and fifth questions). The sum of the responses is calculated. This sum is transformed to a scale ranging from 0-100 by subtracting the lowest possible score from it and dividing the result by the range of possible scores as in equation 3.1.

$$\text{Transformed Scale} = \left(\frac{\text{Actual raw score} - 5}{20} \right) * 100 \tag{3.1}$$

Data imputation for the mental health scale was performed according to the SF-36 manual guidelines (Ware et al., 2000a). For the mental health scale, this entailed imputing values for individuals who failed to provide answers for one or two of the five items on the mental health scale. Imputed values were calculated by taking the average score for the items the individual answered.

Another way of using the SF-36 is to calculate two summary scores; a Physical Component Summary (PCS) score and a Mental Component Summary (MCS) score. These summary scores sum the eight component scales with varying weights (derived using principal components analysis) in order to produce fine grained scales measuring physical and mental health. The are calculated as follows. All eight component scales are calculated (in the same fashion as the mental health score described above). Then each scale is standardised by subtracting the population mean score and dividing by the scale standard deviation (this procedure is called a z-transformation), both of which are estimated from previous studies. The developers of the SF-36 provide US population norms for this purpose (Ware et al., 2000a), but UK norms have been

published (Jenkinson et al., 1997). These z-transformed scales are then summed with different weights, or factor loadings, again derived from population studies. There are a set of factor loadings for both the Physical Component Summary score and the Mental Component Summary score. Both resulting scores are then transformed by multiplying by 10 and adding 50. Since the PCS and MCS require information from all eight component scales, missing values represent a larger problem for them than for each of the component scales. The MCS was not used in the CHSNS as 13.4% of the respondents had missing information precluding the calculation of their MCS scores (Fone, 2005). It is presented here because it will be used in section 3.4.3.

It should be noted that the terms "mental health scale of the SF-36" and "MHI-5" will be used interchangeably in this thesis.

3.2 Validity and Reliability

Since the main outcome measure of the study is mental health it is important to give some consideration to the method of measuring it. The development of a scale to measure mental health is a complicated procedure and it must pass many tests before it can be deemed a success. Chief among these concerns are reliability and validity. Reliability and validity are not the same, but they are highly related. Reliability refers to how consistently a scale measures some construct. It incorporates notions of internal consistency as well as temporal consistency. Validity on the other hand, refers to whether the scale is actually measuring what it is supposed to measure. The two are related, since a scale which is not reliable is unlikely to display high validity. Indeed, the reliability of a scale provides an upper bound for the validity of the scale (Streiner & Norman, 2003). Both properties will be described in more detail now.

3.2.1 Validity

Verifying the validity of a measure is perhaps a more complicated and subjective procedure than verifying the reliability of a scale. There are many different types of validity and ideally all should be addressed. Streiner separates validity into three different types: content, criterion and construct (Streiner & Norman, 2003).

Content validity refers to whether the scale comprehensively covers whatever construct it is attempting to quantify. For instance, a scale which purported to measure bodily pain, but only asked questions regarding back pain, would not demonstrate content validity. This is essentially the same as face validity, so called because it looks at whether "*on the face of it, the instrument appears to be assessing the desired qualities*" (Streiner & Norman, 2003).

Criterion validity refers to the performance of the scale when compared with an

already existing scale which measures the same or similar construct. Ideally, this pre-existing scale is a gold standard for the construct. This makes intuitive sense, in that if a scale is known to measure a given construct well, then a new scale should perform similarly. Of course, if the new scale performs better than the "gold standard", it may not perform similarly at all; however for most practical situations, such leaps in performance would not be expected.

Construct validity is the most general type of validity and as such can (and should) be demonstrated in a number of different ways. If a scale is designed to measure a certain construct (say mental health) then it should be possible to present some hypotheses that would demonstrate whether or not the scale really does measure mental health. Successfully testing such hypotheses demonstrates construct validity. For instance, the effect of being permanently unable to work due to disability has been shown to be associated with poor mental health in the CHSNS dataset (Fone, 2005), therefore, a new mental health scale would be expected to score individuals who are on incapacity benefit as having worse mental health than the general population. Such a hypothesis is testable, and could form part of the validation process for a new mental health scale. Construct validity depends on a clear and explicit model of how and why the scale measures whatever it purports to measure. Streiner lists four different ways to demonstrate construct validity (Streiner & Norman, 2003).

1. Firstly there is extreme groups construct validity. This essentially involves setting up a case control study. To continue the mental health scale analogy, this would involve administering the scale to individuals with poor mental health as well as to individuals with good mental health. For instance, psychiatric patients could be compared with non-psychiatric patients. If the scale in development actually does measure mental health it would be expected that there would be a difference in scores between these two groups. This is sometimes referred to as discriminant validity.

2. Secondly, there is convergent validity. Convergent validity appears to be exactly the same as criterion validity, in that it refers to comparing the scale with other measures of the same construct.

3. Divergent validity ensures that the scale is unrelated to constructs with which it should not be related. So, for instance, if a scale attempts to measure how intelligent an individual is, but does so by asking them to translate a paragraph of Spanish, then individuals from Spain or South America may score highly on this test regardless of intelligence, while highly intelligent people with no knowledge of Spanish will score poorly. Here the scale would be highly related to Hispanic ancestry which would not be expected. Confusingly, Streiner (2003) refers to

divergent validity as discriminant validity, however here it will be referred to as divergent validity.

4. The fourth method of investigating construct validity is called the multitrait-multimethod matrix. This simply involves combining the previous three methods into one study, by applying multiple measures (including the measure in development) to multiple different types of groups displaying traits both related and unrelated to the construct in question. This design allows the calculation of a correlation matrix, with the individual measure's reliabilities on the diagonal, and correlations between different measures off the diagonal. These correlations comprise three types: homotrait-heteromethod (the same construct assessed by different measures), heterotrait-homomethod (different constructs assessed by the same method) and heterotrait-heteromethod (different constructs assessed by different measures). If the new measure is good, the highest correlations should be the reliability of the measure itself, the lowest correlation should be for the heterotrait-heteromethod, and the remaining correlations should be somewhere in between.

More recent trends have moved away from this trichotomy of content, criterion and construct validity. Instead, construct validity has been identified as the most important category and its remit widened so that it permits any testable hypothesis resulting from the theoretical model of the construct. Moreover, the focus has shifted from assessing the test itself, to identifying the strength and reliability of the inferences that can be made from the test. So for instance, if a scale is designed to measure propensity for violence then a validation study could include assessing how well the scale predicts which individuals will go on to commit violent crimes. Essentially, any hypothesis that makes statements about individuals who score high or low on a scale (i.e. "high-scorers on this scale are more likely to go to university" or "low scorers on scale A can be expected to score low on scale B") can be thought of a validation test.

3.2.2 Reliability

When a measurement is made, there will almost always be an error associated with that measurement. This error can be a result of rounding, a limitation of the precision of the instrument (or observer) or due to the variability of what is being measured. Hopefully this measurement error is small enough to be unimportant. For example, when human height is measured it is usually reported with a certain degree of rounding. For most situations the height measurement needs only to be accurate to within an inch and indeed most methods of measuring height are only accurate to within a half inch or so. It is generally understood therefore that a height measurement of six feet

refers to a height less than half an inch from six feet. The error of a half inch is small in relation to the possible range of people's heights, and so it is deemed unimportant. The reliability of a scale provides an estimate of how large this measurement error is likely to be. Two types of reliability that are frequently investigated in the validation literature surrounding the SF-36 will be presented: temporal and internal. Temporal reliability is assessed by test-retest, while internal reliability by Cronbach's Alpha.

Test-retest

Test-retest is a method for measuring how consistently an instrument performs when different observers observe the same individual, or when the same observer observes the same individual more than once. The former investigates the extent of observer bias and precision while the latter investigates how stable the measure is likely to be over time. Both types of reliability can be assessed using test-restest; however since the mental health score in the CHSNS dataset was a self-administered questionnaire only temporal stability is of importance. Temporal stability is an important property for a measure to display. A measure may have poor temporal stability if its measurement error is large or if the construct it is measuring varies rapidly over time. For an example of the latter problem, consider a measure of height that uses the length of shadow cast by an individual. This measure would have poor consistency if measurements were taken at different times of the day, year, or even at different latitudes. This kind of situation is controlled for by administering the measure twice in quick succession to negate the chance of real differences being introduced. In situations where scores would not be expected to vary much over time (such as adult height) then the period between successive tests is not important. In situations where scores may well vary over time (such as mental health) it is important that the period between successive tests is short enough that actual changes are unlikely, but far enough apart that the respondent does not remember the test.

In essence the method involves examining the differences between baseline and follow-up readings of a measure (Bland & Altman, 1986). The difference between scores is plotted against the average of those scores. The average difference, $\overline{d}$ is calculated, along with the standard deviation of differences, s. If the differences are distributed normally 95% of these differences would be expected to fall within the interval $[\overline{d} - 1.96s, \overline{d} + 1.96s]$. This reference interval gives an indication of how large measurement error is likely to be. Bias can be assessed by constructing a confidence interval around the mean difference. Even if the mean difference is significantly different from zero, the authors urge the use of expert knowledge to determine whether the magnitude of the bias is sufficient to represent an important difference for a given field.

Cronbach's alpha

Next there is internal consistency. This is relevant to multi-item scales and measures how interrelated those items are. The classic test for internal reliability of multiple item instruments is Cronbach's Alpha. It works under the assumption that measurement errors for the underlying construct (or latent variable, in this case mental health) are random, uncorrelated between the items and uncorrelated with the latent variable. Other assumptions are that each item is affected equally by the latent variable, and that each item has equal measurement error (Dukes, 1999). It is derived as follows. Say there are k items on a given instrument to measure some latent variable L, with n subjects answering each of them. A variance-covariance matrix can be calculated where the term at position (i, j) denotes the covariance between the i^{th} item and the j^{th} item, as in equation 3.2.

$$\text{Covariance Matrix} = \begin{pmatrix} \sigma_{11}^2 & \sigma_{12} & \dots & \sigma_{1k} \\ \sigma_{21} & \sigma_{22}^2 & \dots & \sigma_{2k} \\ \vdots & \vdots & \dots & \vdots \\ \sigma_{k1} & \sigma_{k2} & \dots & \sigma_{kk}^2 \end{pmatrix} \tag{3.2}$$

The diagonal elements are simply the within item variances. Then a comparison can be made between the diagonal elements of the matrix (termed the unique variation, σ_i^2), and the sum of all of the elements (the total variation $\sum_{j=1}^{k}\sum_{i=1}^{k}\sigma_{ij} = \text{var}(Y)$). This ratio of the unique variation to the total variation, is subtracted from one, and scaled to reduce the dependency on k, the number of items. This gives Cronbach's alpha as in equation 3.3.

$$\alpha = \frac{k}{k-1}\left(1 - \left[\sum_{i=1}^{k}\frac{\sigma_{ii}^2}{\text{var}\,Y}\right]\right) \tag{3.3}$$

If the inter-item correlations are large (indicating that the items are all measuring the same underlying variable) the ratio of unique variation to total variation will be small. This will result in a large Cronbach's alpha. Similarly, small inter-item correlations (indicating a lack of internal consistency) will result in a small Cronbach's alpha. Since Cronbach's alpha is defined as the proportion of the total variation attributable to the latent variable (Dukes, 1999) (i.e. excluding all measurement error) Cronbach's alpha can only assume values between zero and one. An instrument with no internal reliability would score zero (here the questions would need to be completely unrelated), while an instrument with maximum internal reliability (i.e. if the answer to any of the questions was sufficient to exactly predict the answers to all the other questions), would score one.

Guidelines for interpreting values of Cronbach's alpha vary, with one author suggesting that less than 0.6 is unacceptable, greater than 0.8 is very good and over 0.9 implies that shortening the questionnaire should be considered (De Vellis, 1991). The correct guideline to use depends on the research question of interest however, with another author (Bland & Altman, 1997) suggesting that values between 0.7 and 0.8 would be sufficient for research tools, while 0.9 would be the minimum needed for a clinical application.

Cronbach's alpha, while widely used, should be interpreted carefully. Schmitt (1996) warns that the alpha coefficient measures internal consistency and not internal homogeneity. The distinction between the two is described by Schmitt as follows

> "*Internal consistency refers to the interrelatedness of a set of items, whereas homogeneity refers to the unidimensionality of the set of items. Internal consistency is certainly necessary for homogeneity, but it is not sufficient*"

In other words a questionnaire's items can be highly interrelated, without being highly homogenous. This can happen if the questionnaire is measuring two or more different factors or constructs, which are related but not the same. In this case alpha could be high, without the questionnaire being unidimensional or homogenous. For example, a questionnaire designed to assess customer perceptions of a group of products may comprise questions relating to how cheap the products are, as well asking whether those products represent good value for money. There may be a correlation between the price of the products, and how likely they are to be considered good value (with cheaper products being more likely to represent better value) but the two are nevertheless separate issues. Thus the items of the questionnaire may have high reliability, while the questionnaire is not unidimensional. Schmitt recommends reporting alpha coefficients with the associated item correlations and corrected item correlations. This feature of Cronbach's alpha may not represent a problem for this study, since it can be argued that the common mental disorders, are not unidimensional, comprising both anxiety and depression.

Another important feature of Cronbach's alpha is that it is positively correlated with the length of the questionnaire (i.e. number of items on the questionnaire). Cortina (1993) reported that if a scale has more than 20 items, then an alpha coefficient of greater than 0.7 is achievable, even if the inter-item correlations are small. This is not an issue for the MHI-5 however, since it is such a short scale.

3.3 Validity and reliability of the MHI-5 and the SF-36

Literature searches were performed in the Web of Knowledge and Medline search engines, seeking papers which assessed the validity and reliability of the SF-36. Papers which included a keyword from each of the following groups: (SF- 36, MHI-5, Short form), (assess*, validity, validating, psychometric) ,(mental, psychol*, depress*, anxi*). Other papers investigating the validity of the SF-36 were obtained from the SF-36 manuals (Ware et al., 2000a,b).

This resulted in 15 papers which discuss the validation of the SF-36. These will be examined under the following headings: Validity, Reliability, Elderly Populations, Version 1 vs Version 2. The older respondents section will investigate how appropriate the scale is for use in older populations, while the final section will describe the differences (and consequent improvements) of the SF-36 version 2 over version 1.

3.3.1 Validity

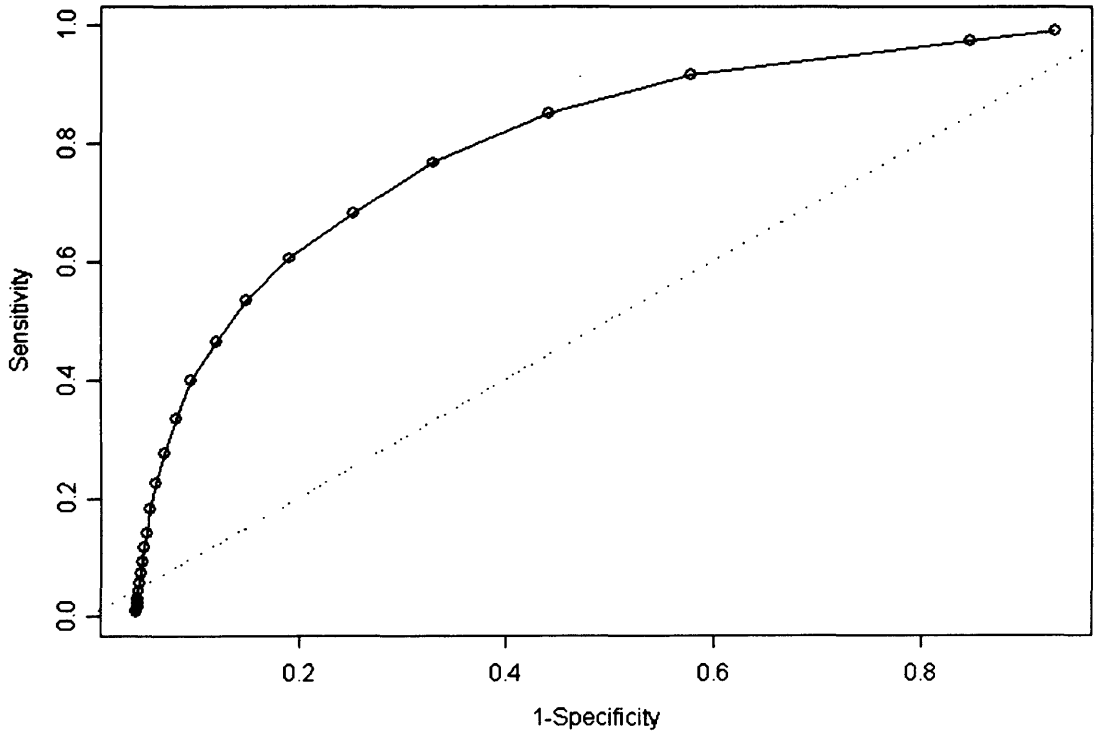
These papers are split into two groups, those that compare the SF-36 with another scale, and those that do not.

Comparison with other scales

Two papers compared the SF-36 with a Diagnostic Interview Schedule (DIS), while another compared it with the GHQ-12. The first of these compared four scales with DIS, namely the mental health component of the SF-36, the MHI-18, the General Health Questionnaire 30 (GHQ-30), and the Somatic Symptom Inventory (SSI-28) (Berwick et al., 1991). There is a consensus that interview administration produces more reliable responses than self administration, since the interviewer can ensure that the respondent understands each question and provide further clarification if required. An interviewer can also screen out contradictory answers. There are a number of drawbacks to interview administration however. One worry is that interviewers (if not trained properly) can lead subjects toward certain answers. This may even happen through no fault of the interviewer, since the act of observing may change the outcome (e.g. an interviewer asking respondents very personal or embarrassing questions may be answered less truthfully than if they received an anonymous self-administered questionnaire). Inter-rater reliability must also be assessed and monitored if an interview-administered questionnaire is to be employed. The DIS, used in this paper, is an interview-administered questionnaire, while the others are self-administered. The subjects were 5,291 people aged 20-64, selected from a health register in Eastern Massachusetts. Using the scores from the DIS these people were classified into one of five categories: those with a current

DIS disorder, the GHQ cluster (consisting of DIS disorders such as major depression, dysthymic disorder, panic disorder), affective disorders (e.g. bipolar disorder), anxiety disorders (e.g. generalised anxiety disorder, obsessive compulsive disorder) and major depression. Each method was assessed for each outcome. Their method of comparing the various measures was to use the area under the Receiver Operating Characteristic (ROC) curve. Basically the ROC curve is plotted with sensitivity on the y-axis and one minus specificity on the x-axis. A perfect ROC curve would describe a right angle, travelling straight up the y-axis at the origin and horizontal at $y = 1$. A totally uninformative measure would describe a diagonal at 45 degrees. An example of a ROC curve is given in figure 3.1. The area under the curve can be used as a way to quantify

Figure 3.1: An example of a ROC curve



measures that lie in between these two extremes. Using this measure the MHI-5 performs as well as the MHI-18 and the GHQ-30 in the detection of any DIS disorder, the "GHQ cluster", anxiety disorders and major depression (it performs worse for the full range of affective disorders). It is consistently as good as the GHQ-30, and is worse than the MHI-18 only for detecting the affective disorders group. The authors note the MHI-5 has "excellent detection capabilities". By comparing the MHI-5 with these other measures they are demonstrating convergent validity.

The second paper involving the DIS compared three mental health scales with it,

namely the GHQ-30, the Mental Health Inventory (MHI-5) and the Somatic Symptom Inventory (SSI) (Weinstein et al., 1989). The DIS was used to separate the sample into different clinical groupings. There were four such groupings: presence of any current DIS disorder (denoted Any DIS Disorder), presence of DIS diagnoses specifically related to the face content of the GHQ-30 (the GHQ cluster), the presence of any DIS affective disorder (Affective Disorders) and the presence of any anxiety disorders (Anxiety Disorders). For each of these four groups receiver operating characteristic curves were plotted for each of the three health scales. The MHI-5 outperformed the other two scales for all but one of the four clinical groupings (as measured by the area under the curve (AUC)). For the anxiety grouping the MHI-5 was not the best (and in fact was the worst) at detecting cases, whereas the SF-36 had an AUC of 0.745, while the GHQ-30 and the SSI had AUCs of 0.753 and 0.760 respectively. The authors conclude by saying "our results suggest that the MHI may be an especially promising screening instrument in a primary care setting". The main criticism that could be levelled at this study surrounds the small sample size of 364. Considering that information on four different scales including a DIS were included, this sample size is large. The clinical groupings chosen make sense in that they refer to psychiatric conditions. This paper provides convincing evidence that the MHI-5 has good convergent validity.

A paper by McCabe et al (1996) compared the GHQ-12 with the MHI-5. Both surveys were sent to 3,000 people randomly selected from two GP registers in Doncaster. One of the registers covered a more socially and economically advantaged area than the other. The construct validity of the scale, was assessed in a fairly ad hoc, qualitative fashion. The results of the study showed that for both the MHI-5 and the GHQ-12, women, the unemployed, those that left full time education earlier, carers of long-term mentally or physically disabled adult dependents, people who reported not having anyone to confide in about personal or emotional problems reported worse mental health, while homeowners and those living in the more socially and economically advantaged areas all reported better mental health. These relationships are currently known and undisputed in the literature, and were taken by the authors as evidence of the convergent validity of the mental health scale (they refer to this as construct validity however). The MHI-5 also correlated highly with the GHQ-12 providing further evidence for convergent validity. The one exception to this was the relationship with age, with the GHQ-12 being positively related with age and the MHI-5 being unrelated to age ($\chi^2 = 24.94$, p-value < 0.01 for the GHQ-12, $\chi^2 = 4.42$, p-value $= 0.45$ for the MHI-5). The authors postulate that the GHQ-12 may have items that are related to both age and mental health, citing item 1 about the respondents "ability to concentrate" as a possible example. While perhaps less rigorous than the previous two papers, this paper provides evidence that the MHI-5 is equally related to mental health as the GHQ-12.

Comparison across different subject groups

A number of papers did not compare the MHI-5 with another mental health scale, instead choosing to assess the performance of the MHI-5 by comparing scores for individuals from different populations.

The first of these used 1317 (75.5% response rate) patients from the north east of Scotland (Grampian) who presented to their GP with one of the following ailments: low back pain, menorrhagia, suspected peptic ulcer, varicose veins (Garratt et al., 1993). Within this group there were referred and non-referred patients, resulting in eight distinct subgroups of patients. The GP was asked to categorise their patient's symptom severity as one of the following: none, mild, moderate, severe. These were compared with 542 (response rate 60.2%) people selected randomly from the electoral register for Aberdeen. The mental health scale identified statistically significant differences (p-value < 0.01) between all but one of the eight groups and the general population. The only group there was not a significant difference for was the non-referred patients suffering from varicose veins. This is not surprising giving that the symptoms from varicose veins can be quite mild, and also that this particular subgroup had by far the smallest sample size of 58. The mental health scale also distinguished between the doctor-assessed ratings for patient's symptoms. Whether this demonstrates convergent validity is questionable. No hypotheses are presented to explain why the mental health of people with suspected peptic ulcers or menorrhagia sufferers would be expected to be different to that in the general population. The fact that the scale does show significant differences could merely be a result of the large sample size. The authors argue that by displaying that the mental health scale is correlated with general health, they have demonstrated convergent validity. This argument, however, seems much more applicable to the physical functioning, role-physical, bodily pain, general health and vitality scales, than it is to the social functioning, role-emotional or mental health scales. Had there been a psychiatric ailment included in the patient groups, then the utility of the MHI-5 could have been much more successfully demonstrated.

Another paper which used different patient groups to demonstrate the convergent validity of the MHI-5 comes from a paper from 1993 (McHorney et al., 1993). This study was carried out in the cities of Boston, Chicago and Los Angeles. The final sample size was 1,014. These participants were divided into four patient groups in this study, defined by the symptoms they exhibit: minor chronic medical conditions, serious chronic medical conditions, psychiatric conditions only and both serious medical and psychiatric conditions. The National Institute of Mental Health's DIS was used to identify depression and measure its severity; however the MHI-5 was not directly compared with this DIS. The MHI-5 displayed divergent validity by distinguishing between the serious medical and the minor medical poorly, compared with the other

seven SF-36 scales. Convergent validity was demonstrated by showing that the mental health scale distinguished between the psychiatric and minor medical groups better than any other scale and was the third best scale at distinguishing between the both serious medical and psychiatric group and the minor medical group. This paper provides good evidence that the MHI-5 is indeed measuring an underlying mental health construct. Without comparing to another mental health score, however, it is difficult to determine how well it does so.

Another paper by Ware et al (1995) examined the convergent validity of the SF-36. The specific focus of their work was to compare the individual scales of the SF-36 with the physical component summary (PCS) and mental component summary (MCS). They examined aspects of the discriminatory ability of the scale in both longitudinal and cross-sectional studies. Here the focus is on the clinical validity for cross-sectional studies. The authors used multivariate analysis of variance (MANOVA) with all eight scales of the SF-36 (including the mental health scale) as the outcomes and patient groups as the independent variables. The patient groups used were four types of chronic medical conditions, two types of hypertension (based on severity), four types of diabetes, two types of congestive heart failure (based on severity), the presence of 16 comorbid conditions (and a count of 10 others) and frequency of acute symptoms in four symptom clusters (ear, nose and throat; central nervous system; musculoskeletal; and gastrointestinal tract/ genitourinary tract), age effects, longitudinal comparisons of physical, mental and general health (one year follow-up), and finally cross-sectional and longitudinal comparisons of patients with clinical depression. The mental health scale was successful in distinguishing patients in all but the hypertension, diabetes and musculoskeletal categories, in a linear model. Again, the ability of the mental health scale to discriminate between these conditions does not necessarily provide evidence for convergent validity (in the absence of literature demonstrating that these groups have distinct mental health distributions). The mental health scale was much less successful at identifying physical change than the physical scale, it was the best at identifying mental change, and it was fifth best (out of the eight scales) at identifying general health change. Moreover, the mental health scale was the best at detecting significant differences in three of the four mental health tests (being second best at the fourth). Again, this merely demonstrates that, of the SF-36 scales, the mental health scale measures mental health constructs the best.

The next paper explicitly assessed both the divergent and convergent validity of the SF-36 (Roberts et al., 1997). It did this in three steps. Firstly, it tested whether items from the SF-36 have equivalent variances, then whether items within a scale were substantially related to the total score computed from other items in the scale (convergent validity) and finally it assessed whether an item correlated more highly with other items from within its own scale than with items from other scales (divergent

validity). The authors reported no results for the convergent validity of the mental health scale but indicated that items within the mental health scale were substantially correlated with the total score computed from the other items in the scale. There was a small amount of evidence for a lack of divergent validity with an item from the mental health scale being equally correlated with the vitality scale and its own scale. The paper went on to assess the clinical validity of the scale using a cohort of 10,308 civil servants. The mental health scale distinguished significantly between those with angina and those without, those with a zero CAGE (an instrument measuring severity of alcohol use) score and those with a maximum score, people who reported absence from work greater than one week in the past year and those who reported no absence due to sickness, medical group and psychiatric group, medical group and medical and psychiatric group, diabetes group and angina plus GHQ caseness group, and finally angina group and diabetes plus GHQ caseness group. In fact the only groups the mental health score failed to differentiate between were the angina only group and the diabetes only group. In fact, no scale differentiated between these two. The authors concluded that this was supporting evidence for the validity of the mental health scale in discriminating between medical and psychological conditions as well as psychiatric morbidity. This paper also failed to supply any hypothesis as to why a mental health scale should discriminate between those with angina and without. Instead they seem to have taken the same approach as many researchers in the field and investigated the validity of each of the eight scales in each patient group, even when a given scale would not be expected to be related to the condition present in that patient group.

A follow-up paper to one of the aforementioned papers (McHorney et al., 1993), in the same setting of Boston, Chicago and Los Angeles, assessed the SF-36 in a sample of 3,445 against 24 different socio-economic subgroups, consisting of three age categories, sex, race (three categories), education (four categories), a poverty indicator (binary), six clinical conditions (hypertension, diabetes, congestive heart failure, myocardial infarction, clinical depression and symptomatic depression) and four disease severities (uncomplicated medical, complicated medical, psychiatric and uncomplicated medical, psychiatric and complicated medical) (McHorney et al., 1994). The percentages scoring the maximum and minimum for each scale in each category were reported in order to identify the presence of floor or ceiling effects. There was no evidence for a floor effect (the maximum percentage scoring zero in any category is 0.8 (those in poverty), and little evidence for a ceiling effect (the maximum percentage scoring 100 is 11.7 in the 75 and over age group). Item-discriminant validity was assessed in the same paper (1994). The correlations between each item (all 36) and each scale (all 8) were calculated. As evidence of divergent validity the authors noted that the correlation between each item and its own scale was greater than all correlations with the other scales by more than two standard deviations.

There is a considerable body of evidence now showing that the MHI-5 has a high degree of validity. It has been compared favourably with many other mental health scales as well as against DIS. Many different approaches have been used to demonstrate the validity of the SF-36, some more successfully than others. The consensus from the literature is that the SF-36 comprises an excellent blend of validity, discriminatory power and brevity, and can be used on diverse social and demographic groups.

3.3.2 Reliability

Cronbach's Alpha

A number of studies have examined the SF-36 using Cronbach's alpha and the general consensus is that the SF-36 has high internal reliability. One study reported Cronbach's alphas ranging between 0.75 and 0.85 for the eight subscales (mental health = 0.79) (Roberts et al., 1997). Three studies report alphas greater than 0.8 (except for the social functioning scale in one of the studies which scored 0.76)(Garratt et al., 1993; Jenkinson et al., 1993; McCabe et al., 1996). The most recent of these studies is based on Version 2 of the SF-36 and reports an alpha of 0.84 for the mental health scale (Jenkinson et al., 1999). Another study quoted an alpha greater than 0.85 (Brazier et al., 1992), while a final study reported a Cronbach's alpha of between 0.82 and 0.90 in 24 socioeconomic and clinical subcategories (McHorney et al., 1994). This last paper also quoted item-scale correlations (correlations between each item and the scale it belongs to) ranging between 0.65 and 0.81 (mean: 0.758). They also suggested that only the physical functioning scale "*consistently met minimum standards of reliability for use on an individual patient level*". That scale reported minimum Cronbach's alpha coefficients of 0.90, up to a maximum of 0.94.

Test-retest

A paper by Brazier et al (1992) examined the test-retest reliability using two methods. Firstly, they used simple correlations between scores from the same people taken two weeks apart. Correlations for all scales were high, with the mental health scale achieving a correlation of 0.75. The authors correctly point out however, that scores can be correlated without being anywhere near equal in magnitude. So for instance if the scores after two weeks were consistently lower than the baseline, the two scores could be highly correlated. To overcome this shortcoming the authors also examined the difference between the scores. For the mental health scale there was a statistically significant difference between the two scores, with the scores a fortnight later being on average 0.71 higher than baseline (the highest difference of all the scales). However, since the scale was measured on a 100 point range, this difference was not deemed

clinically significant. A 95% reference interval (calculated using the differences on for a given scale and assuming a normal distribution) was calculated for each scale. This reference range was not reported, however the authors do state that the percentage of differences that lay within this reference interval was between 91% and 98% for all dimensions. The mental health scale reference range contained the lowest percentage of differences, indicating that the mental health scale is the least Normally distributed scale in the SF-36. While it appears that the mental health scale has the worst test-retest reliability of the eight SF-36 scales, it still demonstrates high levels of reliability. It is also worth noting that it is possible that mental health is genuinely more temporally variant than the other dimensions.

Another paper, by Roberts et al (1997), examined the test-restest reliability in a similar fashion. A sample of 289 people were given the SF-36 (version 1) twice, once at baseline and a second time a month later. The correlation for the mental health scale was 0.83 (95% conf. int : 0.76-0.89). The average difference between these time points was 0.79 (with the baseline being smaller). While this difference was statistically significant, it is small compared to the 100 point range of the score. Again, differences were calculated for all eight scales, standard deviations calculated for these differences and 95% reference ranges calculated. For the mental health scale, 93.1% of the differences lay within the 95% reference range.

Finally the SF-36 Manual and Interpretation Guide (Ware et al., 2000b) references a paper by Nerenz et al (1992), which also examines test-restest reliability. This paper was excluded from the original search because the population it dealt with was too specific (patients with diabetes mellitus). They found a slightly smaller correlation between the two time points of 0.795. The follow-up interval was six months here, much longer than either of the other two studies. This correlation is almost certainly an underestimate of the true reliability of the scale however since such a long follow-up means that there will be changes in scores due to both random error and genuine changes in mental health states.

All in all, the evidence for the reliability of the mental health scale of the SF-36 is remarkably consistent. Cronbach's alpha coefficents for the MHI-5 are repeatedly reported around 0.85, indicating high levels of internal consistency. Test-retest investigations also produce similar results with correlations being reported about 0.75. Taken together this evidence indicates that the reliability of the SF-36 is, if not extremely high, at the very least acceptable.

3.3.3 Suitability for Elderly Populations

The collection of data from older people has long been a problem in the field of health sciences research. Older age groups consistently have a lower response rate than the

population in general, and when they do respond, they tend to have more missing values than the average. Response to the SF-36 is no exception to this general trend. A number of papers have examined the issue of how reliable the SF-36 is for older populations.

The Brazier et al paper (1992) restricted its sample to those in the age range 16-74. They reported that *"the extent of missing data was significantly associated with increasing age in three of the eight scales"*, however mental health was not one of these. They noted that the 65-74 year olds had a higher level of missing data than the other age categories (twice as many missing items as the 55-64 category) and that further studies would be required to ascertain the validity of the SF-36 for use in that population.

A paper by Hayes et al (1995) specifically set out to examine whether the SF-36 was suitable for use with older adults. They used 90 patients aged 65 and over from two general practices and 100 from two outpatient sites. Some were given the SF-36 to self-administer and some were interview-administered. The results for the self-administered group were poor with 42 (43%) of the respondents unable to self-complete the questionnaire. This was due to visual problems, writing difficulties or a general unfamiliarity with completing questionnaires. This figure of 43% comprises 37% aged 75 and over, and 7% aged under 75. There were also problems with missing data for those that could complete the questionnaire, with 34 (61%) of the self-administered respondents omitting one or more of the 36 questions. The majority of these 34 were from the 75 and over category. The main reasons for the missing data was due to confusion over questions relating to work, e.g. "During the past 4 weeks have you had any of the following problems with your work or other regular daily activities as a result of your physical health". Many respondents thought this question inapplicable to them since they were either retired or didn't work. The authors noted however that the difficulties in self-completion here are likely to be an artefact of the study population instead of being specific to the SF-36, and would probably be observed in most measures completed by older people. Despite this high level of missing respondents and data, 177 (91%) of the respondents regarded all or most of the questions as clear and easy to understand. Also, the SF-36 was able to distinguish between the outpatients and the general practice respondents (with the outpatients having poorer health as expected). The authors suggested some minor word changes (particularly focussed on downplaying the work aspect of some of the questions) would make the SF-36 more suitable for use with older respondents.

A paper by McHorney et al (1994) reports that item completion is significantly lower for older populations. For the mental health scale however, the decline is quite mild, with under 65s having 91.7% completion, 65-74 having 88% completion and over 75s having 84.3% completion. They give some suggestions to improve item completion

among the elderly. They suggest a larger type size for the questionnaire to facilitate those who are visually impaired as well as telephone interview administration.

Another study (Parker et al., 1998) investigated response rates and completion rates for the SF-36 version 1 in elderly people (at least 65 years of age) who were medical or surgical inpatients (1,016 individuals). A small number of outpatients and visitors to general practice were also examined for the purposes of comparison with the inpatient group (80 outpatients and 40 GP patients). They found that response rates were related to age, disability, and cognitive impairment. They concluded that the self-completed version of the form was not suitable for routine use in hospital inpatients due to low response rates (533 individuals, 53%). Also, the response rate was related to physical and cognitive co-morbidity (measured independently) meaning that non-response bias is a serious problem for elderly patients.

Another paper (Hill et al., 1996) found that the SF-36 was unlikely to be suitable for older populations. A sample of forty-seven older patients who had been referred to mental health or continence services were given the SF-36 within a week of their referral and then again three months later. The authors found no statistically significant changes for any of the scales (except the pain scale). Qualitative reports of how their referral affected them found that most people had positive feelings of support, confidence, changed outlook and reduced sense of burden. These changes were not detected by the SF-36. This was taken as evidence that the SF-36 is not suitable for older populations. Since qualitative reports of change are by definition not numerical in nature, it is difficult to determine whether the change reported by the individuals was of sufficient magnitude that the SF-36 should have detected a change. The authors note that the patients themselves tended to talk about how they felt and their general mood and outlook, while the SF-36 focusses on individual's ability to perform tasks. The authors concluded that the SF-36 was not a suitable measure for older individuals because the high level of comorbidity in these groups masks change and the SF-36 focusses on functional tasks which are often inappropriate or inapplicable to older patients.

A paper from 1996 (Andresen et al., 1996) set out to examine the test-retest reliability of the postal version SF-36 in older adults. Four hundred and twenty two adults aged 65 and over and living in the community were sent SF-36 forms. Of these 253 (60%) returned questionnaires. Missing values further reduced this figure to 223 (53%). A month later, the process was repeated. Completed questionnaires for both mailings were received for 186 (45%) individuals. Internal consistency for the scales ranged between 0.802 and 0.924 with the mental health scale scoring 0.876. Test-retest reliability was also performed for each scale producing correlations ranging between 0.648 and 0.868. The mental health score had a correlation of 0.845 between time points. This study concluded that the SF-36 had good internal consistency and test-

retest reliability. The response rate for these older adults was not sufficiently high, with the authors recommending that further work be done to increase the response rate. The authors also found that healthier individuals were more likely to respond.

There is a final justification of the exclusion of the over 75s age group which is unrelated to the psychometric properties of the MHI-5. Older people are more likely to have cerebrovascular pathology, including cognitive decline and dementia, as a cause of lower mental health status (Brayne et al., 1995). The effect of the presence of these conditions on self-reported mental health is undetermined. For this reason, those respondents over 75 years of age were excluded.

The message from the literature is that it is unclear whether the SF-36 is a reliable tool for assessing elderly populations. A number of studies have reported worse response rates and more missing item information for the over 75 age group (a trend that is mirrored for the CHSNS dataset, as described in chapter 2). There is some debate over whether the tool can be modified for use in the older age groups, but in the original version 2 format, it seems that the SF-36 may not be as reliable as it is for the younger age groups. This message was supported by the CHSNS dataset. The proportion of missing mental health and socio-demographic response data increased with the age of respondents (Fone, 2005; Fone & Dunstan, 2006). This relationship was particularly evident over the age of 75 years.

3.3.4 Version 1 versus Version 2

There are a number of differences between version 1 and version 2 of the SF-36. Generally, these changes were to improve the layout and simplify the wording of the scale. The most important difference for the mental health scale is that version one had six possible answers to each question, while version 2 has only five. The version 1 order went as follows: "all of the time", "most of the time", "a good bit of the time", "some of the time", "a little of the time" and "none of the time", as in table 3.2. In version 2 the "a good bit of the time" category was dropped (as described previously in table 3.1) because it was found that this category was not consistently ordered in relation to the other possible responses. The removal of this category resulted in little or no loss of information, compared with version 1 (Ware et al., 2000a).

Table 3.2: Mental Health Inventory (MHI-5) items Version 1

	How much of the time during the past four weeks	**Responses**	**Score**
1.	have you been a very nervous person?	all of the time	1
2.	have you felt so down in the dumps	most of the time	2
	that nothing could cheer you up?	a good bit of the time	3
3.	have you felt downhearted and low?	some of the time	4
		a little of the time	5
		none of the time	6
4.	have you felt calm and cheerful?	all of the time	6
5.	have you been a happy person?	most of the time	5
		a good bit of the time	3
		some of the time	3
		a little of the time	2
		none of the time	1

A study by Jenkinson et al (1999) examined the SF-36 version II in a UK setting. Data on 8,889 individuals, collected via a postal questionnaire, was used to assess how the changes to version II affected its performance. The authors were particularly focussed on the PCS and MCS, however they do note that all eight dimensions of the SF-36 display good internal reliability as measured by Cronbach's alpha. The PCS and MCS show increased reliability over the equivalent domains scored by the SF-36 version 1.

3.3.5 Conclusions

The SF-36 mental health scale has been demonstrated to quantify mental health at least as well as any other well-known scale, and in some cases even appears to be a better measure of mental health. It is a remarkably short questionnaire, but its brevity does not appear to be costly in terms of accuracy. Reliability and validity have been investigated for the SF-36, and appear to be sufficiently high to justify its use. Evidence suggests that its application to assess elderly patients, particularly in postal questionnaire format, is not to be recommended due to decreased response rates, inapplicable questions and problems with comorbidity masking change. Version 2 represents an improvement over version 1 in terms of internal consistency, however both versions perform similarly.

Table 3.3: Summary of mental health scores

MHI-5	Frequency	MHI-5	Frequency
0.0	42	55.0	527
5.0	45	56.25	18
6.25	2	58.33	6
8.33	2	60.0	632
10.0	72	62.50	16
12.50	5	65.0	672
15.0	98	66.67	6
16.67	4	68.75	31
18.75	6	70.0	738
20.0	135	75.0	911
25.0	160	80.0	1049
30.0	187	81.25	29
31.25	9	83.33	6
33.33	6	85.0	1142
35.0	218	87.50	36
37.50	14	90.0	1527
40.0	342	91.67	4
41.67	3	93.75	13
43.75	24	95.0	599
45.0	438	100.0	379
50.0	500		

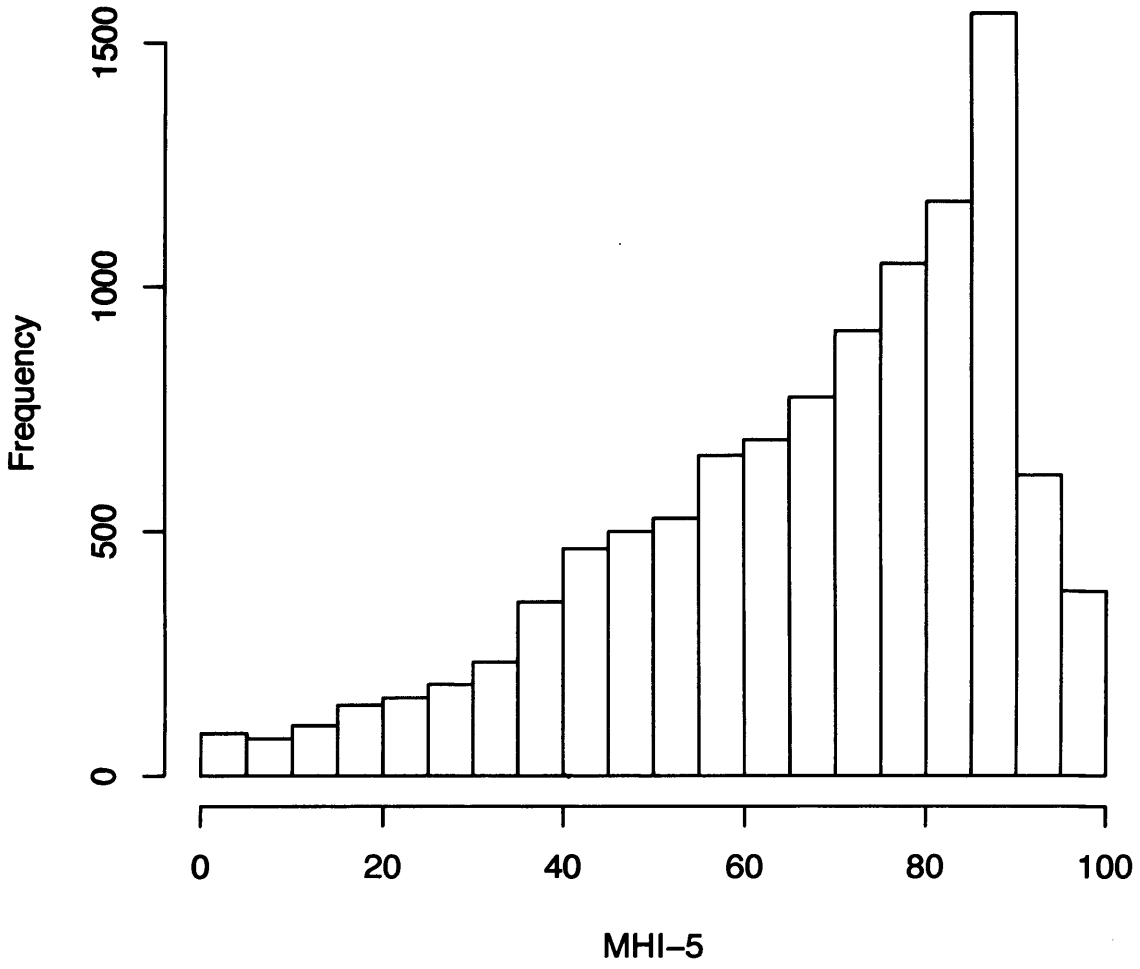
3.4 Methods of Analysis

In this section various ways to model the mental health scale from the SF-36 are considered. The mental health score (as formulated in equation 3.1) is frequently treated essentially as a continuously distributed normal variable, however, there are a number of other ways to model it. Firstly, some transformation of the data can be performed in order to make it follow a normal distribution more closely. Secondly, the mental health score can be treated as an ordinal variable and modelled using ordinal regression. Finally, the mental health scale can be dichotomised in some meaningful way, and the resulting variable treated as a binary one. Unlike the General Health Questionnaire (GHQ-12), the SF-36 mental health scale does not have a validated cutpoint to define a case of common mental disorder.

Because of the short number of questions and the multiple choice nature of the responses there are a relatively small number of scores that the MHI-5 can possibly produce. Excluding imputed scores there are 21 possible values. All values were transformed into a scale that ranged from 0 to 100 using equation 3.1. See table 3.3 for a summary of the possible values of the mental health score. There were a few values with quite low frequencies (e.g. 11.282, 31.234, 59.394 etc). These values were

imputed scores for those individuals who neglected to answer one or two of the MHI-5 questions. Any questions with missing answers were replaced with the average score for those questions with answers (provided that at least three of the five questions were answered). This is the standard imputation method (Ware et al., 2000b). Figure 3.2 shows the distribution of this score. Clearly, this variable was highly negatively

Figure 3.2: Distribution of MHI-5



skewed. Other studies which use the MHI-5 to measure mental health have simply ignored the problem of skewness. It is possible that this is a valid way to proceed as some statistical methods are quite robust to normality assumption violations (e.g. ANOVA, t-test, linear regression) (Lumley et al., 2002). However, the effect of such violations on hierarchical modelling is less well understood. It is prudent therefore to consider other approaches to dealing with this problem

3.4.1 Box-cox transformation

A natural starting point for this was to try using a simple Box-Cox (1964) transformation. The Box-Cox transformation is of the form

$$T(Y) = \frac{(Y^{\lambda} - 1)}{\lambda} \tag{3.4}$$

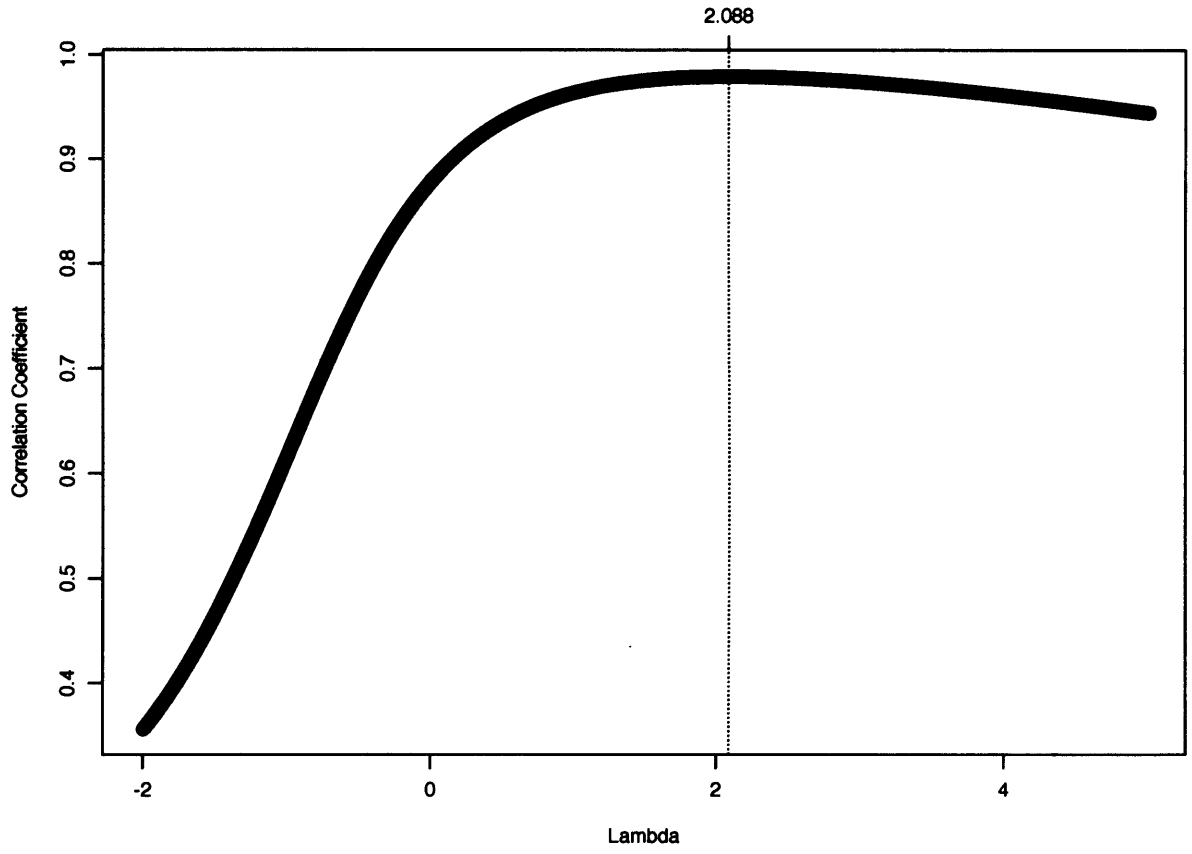
where Y is the response variable and λ is a constant known as the transformation parameter. In this method, λ is chosen so that the transformed variable is distributed as normally as possible, typically by plotting the log-likelihood as a function of λ. There are various measures of normality that can be used. Two methods will be investigated, firstly the correlation between the quantiles of the standard normal and the quantiles of the transformed response and secondly the log-likelihood of the transformed response. The higher the correlation or log-likelihood, the better the transformation. Figure 3.3 shows the relationship between λ and this correlation with quantiles of the standard normal. The dotted vertical line indicates the value of λ which provided the maximum correlation. Negative powers have a disastrous effect on the correlation with the standard normal, providing correlations much lower than no transformation at all ($\lambda = 1$). The λ which produces the best correlation is 2.088, but as figure 3.3 shows, there are a wide range of values about 2 which produce similar correlations.

A histogram of the transformed response shows the effect of the Box-Cox transformation (see figure 3.4). While this new variable was clearly still not normal, it was an improvement over the original in many respects. The transformation particularly reduced the skewness of the data (from -0.857 to -0.163). The large trough on the right hand side of the plot is a result of small numbers of imputed values and the transformation. Another way to measure normality is to use the log-likelihood. This was performed using a standard package in S-Plus (MASS library). This function takes a linear model and computes profile log-likelihoods for simple power transformations of the response, e.g. y^{λ}. Figure 3.5 illustrates the effect of λ on the log-likelihood. The λ that gives the largest likelihood is 1.763. Again, negative powers do not help in normalising the response, and the optimal λ provides only minimal improvement over the original untransformed response. The corresponding histogram of the transformed score is shown in figure 3.6. The skewness under this transformation is -0.322. There are two troughs present in this histogram, again attributable to small numbers of imputed values.

Summary

The application of the Box-Cox transformation to the MHI-5 scores collected in the CHSNS dataset indicate that the square transformation is approximately optimal for

Figure 3.3: Relationship between lambda and correlation coefficient



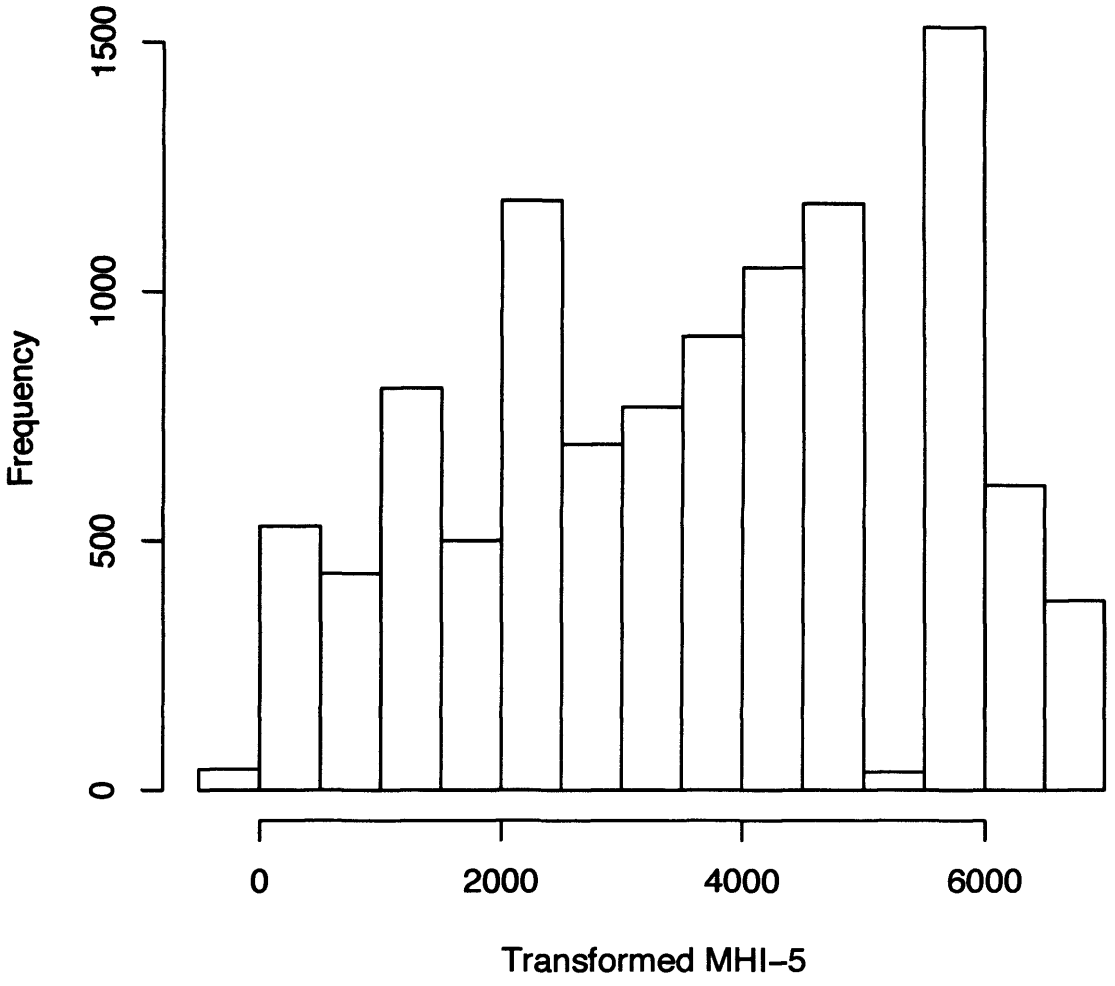
reducing the skewness of the scores. Figure 3.4 shows however that the resulting transformed distribution still does not appear to be Normally distributed. The square transform did not produce different conclusions to treating the MHI-5 scores as coming from a Normal distribution (Fone, 2005). Once a transformation is applied to the response variable in any analysis the interpretation of the parameters from that model become more complicated and so the square transform is not justified.

3.4.2 Ordinal Regression

Motivation

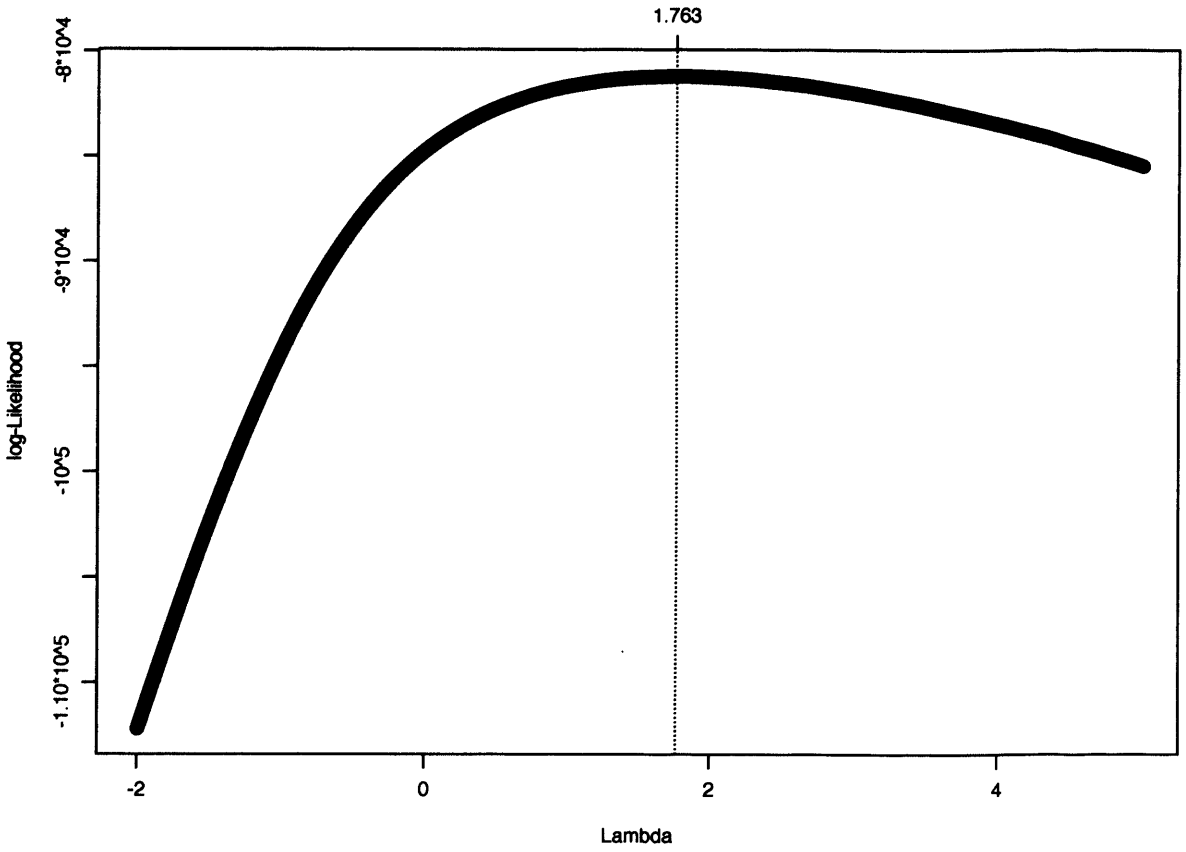
One of the ways to overcome the problem of a skewed response is to use ordinal regression. An ordinal variable, as the name suggests, is a discrete variable with an inherent ordering. However, it differs from an interval variable, in that the difference between categories is unknown or unknowable. An example of an ordinal variable is the educational achievement of an individual (on a very simple level). People might be categorized into three categories: those who left school at 15, those who left school at 18, and those who went on to further education. It is clear that there is a trend

Figure 3.4: Distribution of the transformed MHI-5, $\lambda = 2.088$



of increasing education in these categories, with those who left school at 15 having the lowest level and those who went on to further education having the highest level. However, quantifying the difference in educational level between any of the categories is not possible. So ordinal variables carry more information than nominal variables (where there is no ordering), but less information than interval variables (where there is a calculable and meaningful distance between the values of the variable). Here the MHI-5 mental health score will be treated as an ordinal variable. The MHI-5 is used here merely to rank individuals, without placing any meaning on the numerical distance between individual scores, e.g. it is not assumed that the difference between a mental health score of 20 and 40 is the same as the difference between scores of 40 and 60. There are a number of ways to model such a variable. In general, however, ordinal regression works by modelling logits of probabilities of the response belonging

Figure 3.5: Relationship between λ and log-likelihood



to certain categories, given covariate information. More detail is given below. Instead, if the ordinal variable has n categories, then n logistic regressions are applied, with the covariate coefficients constrained to be constant across all these logistic regressions.

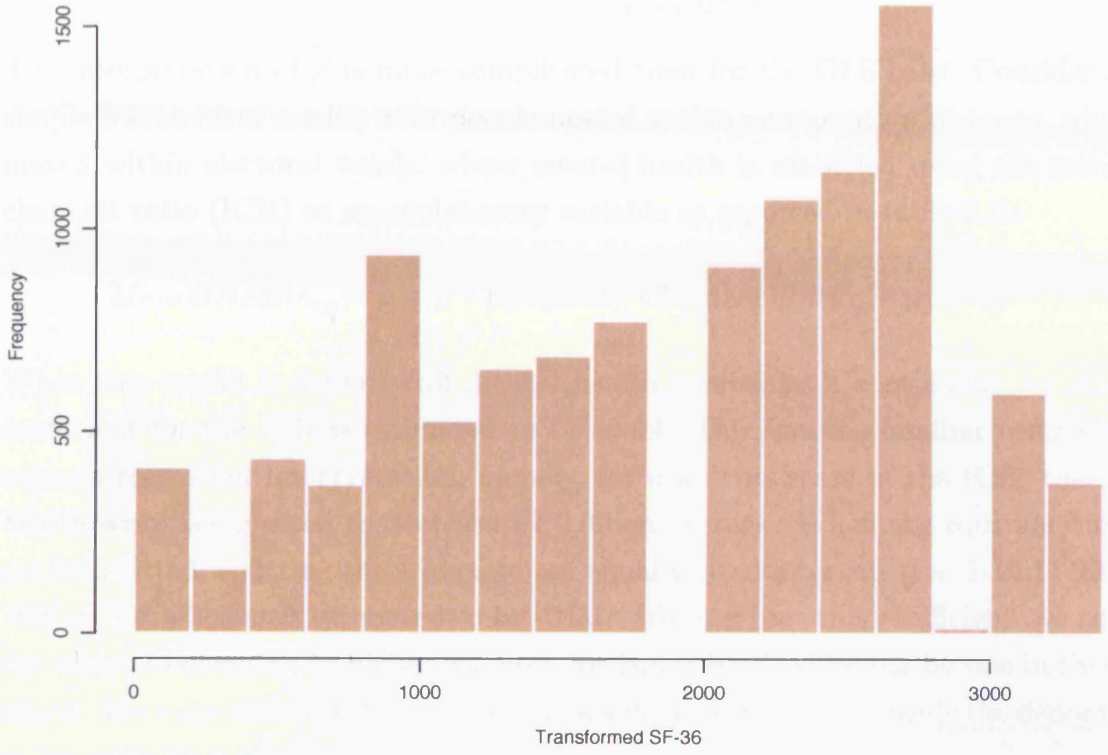
Method

There are several choices when the response has a number of levels including; Cumulative (or accumulated) logit, Continuation-ratio logit and Adjacent-categories logit. Each of these reduces the problem to a binary one in different ways. Imagine a response variable with c categories and let π_i denote the population probability that a given subject belongs to response category i. The cumulative logit for a category j in the response compares the probability of being above this category with being in or below that category as shown in equation 3.5.

$$L_j = \log \frac{\pi_j + \ldots + \pi_c}{\pi_1 + \ldots + \pi_{j-1}} \tag{3.5}$$

The continuation-ratio logit for each category j compares the probability of being in the category just above the cut point with being anywhere below it, as described by

Figure 3.6: Distribution of the transformed MHI-5, $\lambda = 1.763$



equation 3.6.

$$L_j = \log \frac{\pi_{j+1}}{\pi_1 + \ldots + \pi_j} \tag{3.6}$$

The adjacent-categories logit compares the probability of being in category $j+1$ with being in category j as shown in equation 3.7.

$$L_j = \log \frac{\pi_{j+1}}{\pi_j} \tag{3.7}$$

The first of these will be focussed on, the cumulative logit (equation 3.5). The logit of belonging to a higher category will be modelled as described in equation 3.8, where Y is the response, β is a vector of covariate coefficients and X is the covariate matrix. Without loss of generalisability only one covariate is modelled.

$$L_j = \text{logit}\left(Pr(Y \geq j)\right) = \alpha_j + \beta X \tag{3.8}$$

This approach shares a disadvantage with logistic regression in that the covariate coefficients can no longer be interpreted in terms of the original units of measurement, but instead, must be thought of in terms of probabilities and odds. Equation 3.8 leads

directly to equation 3.9, which allows the probabilities of lying in each category to be calculated for each individual with given covariates.

$$Pr(Y \geq j) = \frac{e^{\alpha_j + \beta X}}{1 + e^{\alpha_j + \beta X}} \tag{3.9}$$

The interpretation of β is more complicated than for the OLS case. Consider a very simple hierarchical model, with people nested within enumeration districts, which are nested within electoral wards, where mental health is modelled using the incapacity claimant ratio (ICR) as an explanatory variable as reported in table 2.10.

$$MentalHealth_{ij} = \mu + \beta * \text{Incapacity Claimant Ratio}_j + \mu_j + \epsilon_{ij} \tag{3.10}$$

When this model is fitted (with mental health treated as a continuous variable) the coefficient for the ICR is estimated to be -0.09. This has the familiar ordinary least squares regression interpretation, namely, for a unit increase in the ICR, the mental health score is expected to decrease by 0.09 on average. When the equivalent ordinal model is fitted, splitting the scale into ten equally sized intervals (i.e. 1-10,11-20,...,91-100), the coefficient is estimated to be -0.007. To interpret this coefficient, we examine the odds of belonging to higher category for two subjects differing by one in their ICR scores. Here the numerator is the odds for a subject with $X = x$, while the denominator is the odds for a subject with $X = x + 1$.

$$\frac{Pr(Y \geq j|X = x)}{Pr(Y < j|X = x)} \Big/ \frac{Pr(Y \geq j|X = x+1)}{Pr(Y < j|X = x+1)} = \frac{\text{expit}\,(\alpha_j + \beta x)}{1 - \text{expit}\,(\alpha_j + \beta x)} \Big/ \frac{\text{expit}\,\big(\alpha_j + \beta(x+1)\big)}{1 - \text{expit}\,\big(\alpha_j + \beta(x+1)\big)} \tag{3.11}$$

The right hand side of equation 3.11 reduces to e^{β}. This implies that for a unit increase in the ICR, the odds ratio of belonging to a higher category is e^{β}. In this situation we have $e^{-0.007} = 0.993$. This means that for a unit increase in the ICR, the odds ratio given in equation 3.11 that an individual lies in a higher category decreases by 0.7%, for all categories. Essentially, this means that when the incapacity claimant ratio increases, the odds ratio of belonging to a higher category increases (for all categories). For example, when the ICR is at 10 the odds of belonging to category 8 or higher is 2.18. When the ICR increases to 11, this odds decreases by 0.7% to 2.17. The probability of a person with given covariates belonging to each ordinal category can be calculated, and these can be used in various ways to provide predictions. Two of these methods will be illustrated with reference to the wards with the highest (Aberbargoed) and lowest (St. Martins) incapacity claimant ratios in Caerphilly, as illustrated in table 3.4. The first method is to calculate the expected value for a given individual. This is especially useful if the categories from which the response are derived are based on a continuous scale, as the value of this expectation can range between all values of

Table 3.4: Ordinal predictions for the most and least deprived wards

				Ordinal model	
Ward Name	Observed Mean Mental Health	Observed Mode Mental Health	Linear Prediction	Predicted Mean Mental Health	Predicted Mode Mental Health
Aberbargoed	63.24	80.01-90	64.28	61.95	80.01-90
St. Martins	76.12	80.01-90	74.24	71.08	80.01-90

the original response. The second method is to simply choose the category associated with the largest probability. These are referred to as the predicted mean and predicted mode respectively in table 3.4.

Since the ordinal model approach gives information on the probability of an individual lying in each category we can compare the probability distributions for these two wards. As figure 3.7 shows, the ordinal prediction does not fit the observed data perfectly. This is to be expected since the ordinal model does not have all of the observed information, only information on each individual's category. The fit is reasonably close however. A similar situation is observed for the ward with the lowest ICR, St Martins, in figure 3.8. Here the prediction is perhaps even closer to the observed probability density. It is clear from table 3.4 however that ordinal regression of the MHI-5 with 10 categories produces less accurate results than ordinary least squares regression in this case. This can be interpreted as implying that the skewness of the mental health scale is not sufficient to seriously bias the OLS estimates, making the use of ordinal regression unnecessary in this situation.

3.4.3 Binomial Modelling

Dichotomising a skewed continuous response is one way to sidestep the problem of non-normality, however, it can be criticised for being an inefficient method since it does not use all of the information that the original scale provides. In this respect the ordinal regression approach is certainly more methodologically sound. However, there is another reason to dichotomise the response, namely that cutpoints for scales such as the SF-36 are extremely useful in public health settings. They often facilitate the interpretation of the results of a scale. For instance, the SF-36 transformed scale ranges from 0-100, but it is unclear what the practical implications of a unit difference on this scale might be, let alone what size difference is clinically important. A well validated cutpoint, however, can divide scores into two meaningful categories. Also, while continuous scales undoubtedly convey more information, clinical management decisions are often based on simple yes/no categorisations. For practical purposes then,

Figure 3.7: Comparison of predicted and observed probability densities for Aberbargoed

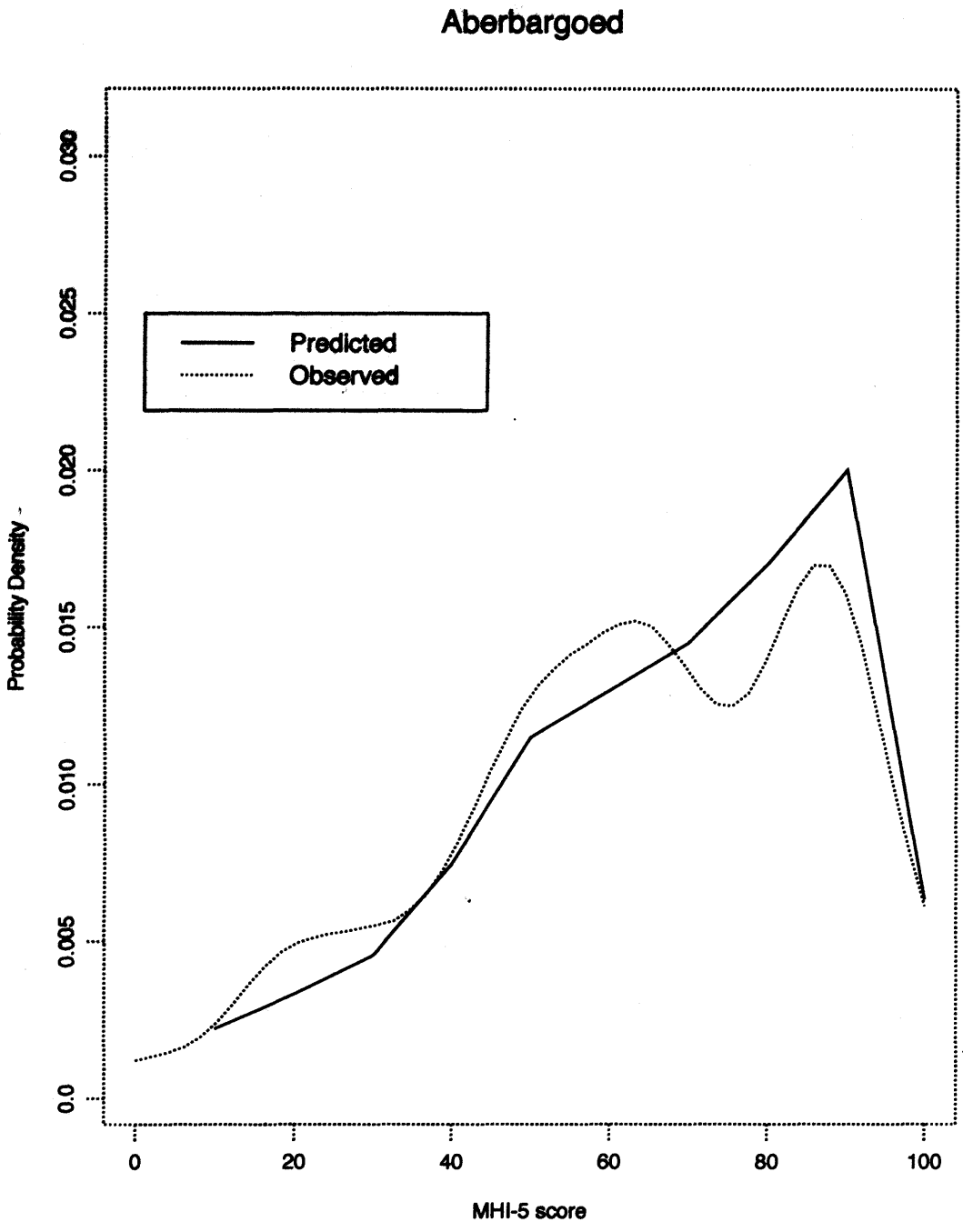
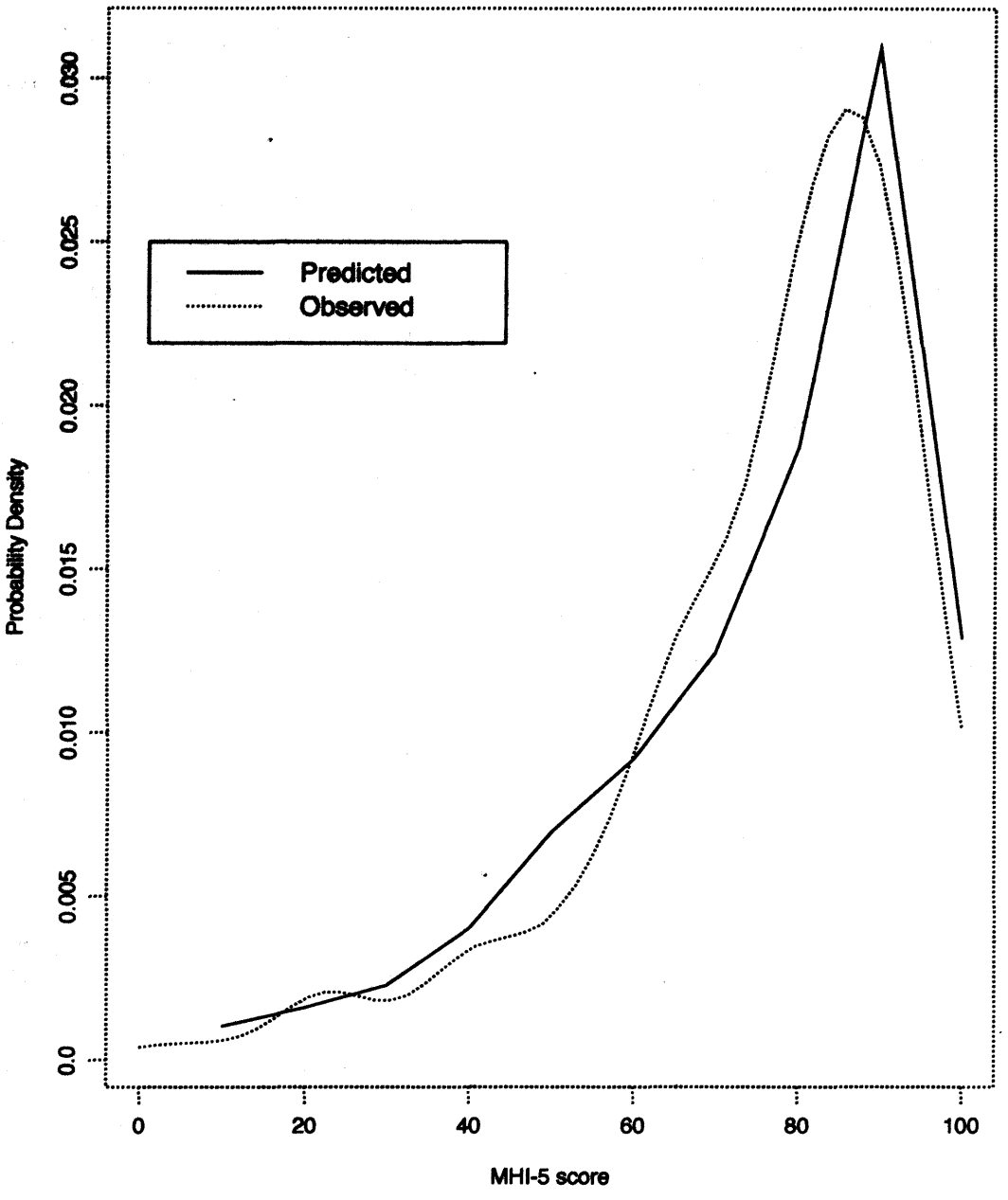


Figure 3.8: Comparison of predicted and observed probability densities for St. Martin's



cutpoints can be invaluable. In order to identify a cutpoint on the SF-36 which can identify a case of common mental disorder, it needs to be compared with a scale which has such a cutpoint. The General Health Questionnaire is such a scale. Unfortunately, the CHSNS did not incorporate the GHQ-12 into the postal questionnaire survey. The British Household Panel Survey (BHPS) wave 9 (Taylor et al., 2005) did include both the SF-36 and the GHQ-12, and data from this wave will be used to identify a cutpoint for the MHI-5.

Derivation of a cutpoint for the MHI-5 to identify a case of common mental disorder

The first wave of the BHPS was carried out in 1991 with a nationally representative sample of 5,500 households. The BHPS follows households through time, with an annual interview of every member of the household aged 16 and over. Individuals interviewed in the first sample who subsequently set up their own household continued to participate in the survey, as well as every individual in the new household. All waves of the BHPS include the GHQ-12, but wave 9 of the study (2000) also included the SF-36 version 1. There is complete information on both of these instruments for all 14,669 individuals in the dataset. The GHQ-12 comprises twelve questions, each with a set of Likert scale responses which score the question as 0, 1, 2 or 3. These are summarised in table 3.5. There are two ways of scoring the GHQ-12. Either the sum of these responses is used to provide a score ranging between 0 and 36 or alternatively, the response to each question is deemed positive if it is greater than one and the number of positives provides the score. This results in a score between 0 and 12 for each individual. This latter method is used in this study. Different studies use different cutpoints between 2 and 4 to define a case of common mental disorder. The most widely accepted convention is that a score of three or more is defined as a case (Goldberg et al., 1997). In that paper, the authors also state that the GHQ-12 is suitable for use as a case detector. The SF-36 version 1 was the version of the SF-36 used in this sample, as given in table 3.2. The MHI-5 questions were extracted from the SF-36 and used to construct the MHI-5 score.
The mental health component summary (MCS), introduced earlier in section 3.1, was also constructed. It was calculated in the standard way (Ware et al., 2000a), using UK norms (Jenkinson et al., 1997) and factor loadings (Jenkinson et al., 1999).

It should be noted that the MHI-5 is designed for use in investigating population mental health and not as a diagnostic tool. As mentioned previously, the GHQ-12 however does have a validated cutpoint to identify a case of common mental disorder (Goldberg et al., 1997). It has been demonstrated that the use of depression/anxiety or case finding instruments has no impact on the recognition, management or outcome

Table 3.5: General Health Questionnaire (GHQ-12) Items

	Have you recently	Responses	Score
1.	been able to concentrate on what you're doing?	better than usual same as usual less than usual much less than usual	0 1 2 3
2.	lost much sleep over worry?	not at all	0
3.	felt constantly under strain?	no more than usual	1
4.	felt you couldn't overcome your difficulties?	rather more than usual much more than usual	2 3
5.	been feeling unhappy or depressed?		
6.	been losing confidence in yourself?		
7.	been thinking of yourself as a worthless person?		
8.	felt that you are playing a useful part in things?	more so than usual same as usual	0 1
9.	felt capable of making decisions about things?	less so than usual much less than usual	2 3
10.	been able to enjoy your normal day to day activities?		
11.	been able to face up to your problems?		
12.	been feeling reasonably happy, all things considered?		

of depression/anxiety in primary care (Gilbody et al., 2001, 2006). As such, a cutpoint on either scale is only appropriate for use in research on populations and would not be suitable to clinically diagnose individuals.

Statistical Methods

Sensitivity and Specificity In order to identify a cutpoint for any new measure, it needs to be compared to another scale which can classify people as a case or a non-case. Ideally, this scale would be a gold standard and would produce no misclassifications. In the field of mental health however, no such scale exists. A well validated scale, such as the GHQ-12, with an associated cutpoint to distinguish cases from non-cases (albeit with error) is a good alternative. The GHQ-12 classifies each individual in the dataset as a case or a non-case. Our aim is to find the cutpoints on the MHI-5 and MCS that imitate the gold standard as closely as possible. Individuals with mental health scores less than or equal to the cutpoint on the MHI-5 or MCS will be defined as cases. The evaluation of a cutpoint involves the twin concepts of sensitivity and specificity. The sensitivity of a test is the probability of a case testing positive (i.e. a true positive).

The specificity of a test is the probability of a non-case testing negative (i.e. a true negative). Clearly a good test has a large sensitivity, but a test which automatically classifies everyone as a case has a sensitivity of one (the maximum possible), even though it is completely uninformative. There is a trade-off to be made, then, between sensitivity and specificity. As the cutpoint is decreased, the sensitivity decreases, while the specificity increases.

Receiver Operating Characteristic Curves As described in section 3.3.1 for each possible cutpoint on the SF-36 there is an associated sensitivity and specificity. These can be summarised using a receiver operating characteristic (ROC) curve. A ROC curve plots the sensitivity (i.e. true positive) against one minus the specificity (i.e. false positive). Each point on the curve represents a different cutpoint on the SF-36. A diagonal line at 45 degrees represents a completely uninformative test, or the line of chance.

Optimisation Criteria There are several approaches to choosing a cutpoint on a ROC curve. Five of these will be investigated in this study. Each method focuses on optimising a different criterion and so may produce a different cutpoint. The five methods are: 1. the Youden Index (1950), 2. the point closest to the upper left corner, coordinates (0,1), as used by Holmes (1998) 3. the misclassification rate, 4: the minimax method (Hand, 1987) and 5. prevalence matching, as used by (Hoeymans et al., 2004). Only the first two have a graphical interpretation on the ROC curve.

Youden Index In general, a good cutpoint is one which produces both a large sensitivity and a large specificity. An intuitive method therefore, is to maximise S, the sum of the sensitivity and specificity.

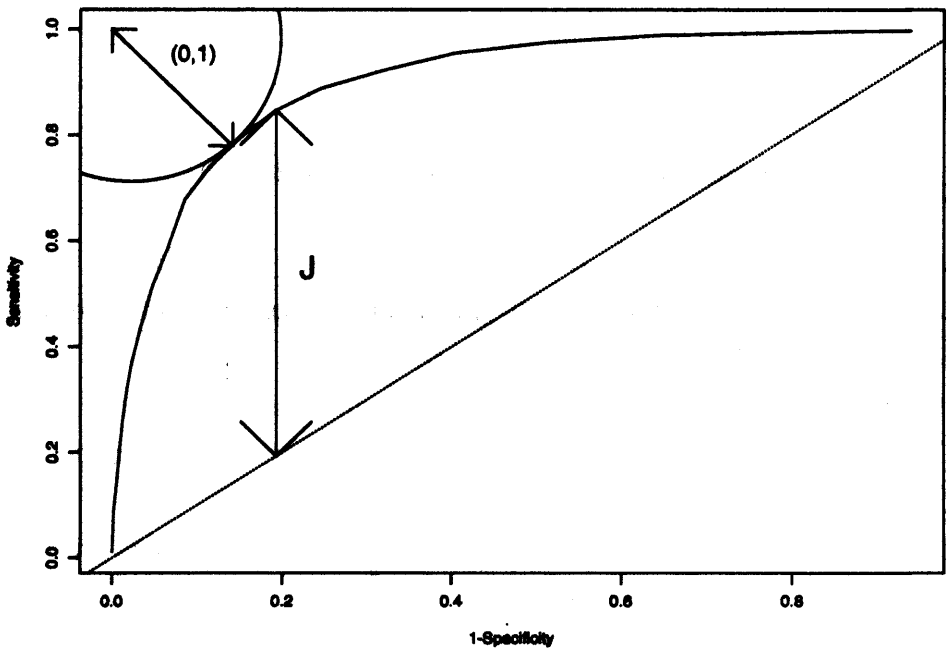
$$\mathrm{S} = max(\text{Sensitivity} + \text{Specificity}) \tag{3.12}$$

This approach assumes that sensitivity and specificity are equally important. This is exactly equivalent to the Youden index, J, shown in equation 3.13, since subtracting a constant does not affect the optimal cutpoint. This can be interpreted as the point on the ROC curve with the largest vertical distance from the line of chance.

$$J = max(\text{Sensitivity} + \text{Specificity} - 1) \tag{3.13}$$

Shortest distance to upper left corner The second optimisation method investigated in this paper is to choose the cutpoint associated with the point on the ROC curve closest to the upper left corner. This entails finding the cutpoint which min-

Figure 3.9: Graphical illustration of the Youden Index (J) and the (0,1) criterion



1. (0,1) refers to the minimum distance between the point (0,1) and the ROC curve
2. J refers to the Youden Index in equation 3.13

imises d in equation 3.14. This method also places equal weight on the sensitivity and specificity.

$$d = \sqrt{\left(1 - Sensitivity\right)^2 + \left(1 - Specificity\right)^2} \tag{3.14}$$

The rationale behind this is that a perfect ROC curve would pass through the point (0,1) (i.e. $Sensitivity = 1$ & $Specificity = 1$ for some cutpoint). Selecting the point on the curve which is closest to this point of perfection is one way to choose a cutpoint. The Youden index and the (0,1) criterion are illustrated in Figure 3.9.

Misclassification rate Alternatively, the misclassification rate, or error rate, could be minimised. For this we define the false positive rate (FPR) to be

$$\text{FPR} = \text{(Non-case Prevalence)x(1-Specificity)} \tag{3.15}$$

and the false negative rate (FNR) to be

$$\text{FNR} = \text{(Case Prevalence)x(1-Sensitivity)} \tag{3.16}$$

and it is the sum of these two terms that is minimised. This essentially gives weights to the sensitivity and specificity based on the prevalence of cases. If the population

has a very low prevalence of cases, then more weight would be given to specificity. If the prevalence is high, then sensitivity takes precedence. This presupposes that the penalty incurred for a false positive is equal to that incurred for a false negative. If this does not hold, the sum can be weighted according to the penalties incurred for false positives and negatives, i.e. minimise

$$\theta\text{x(1-Sensitivity)} + (1-\theta)\text{x(1-Specificity)} \tag{3.17}$$

where θ is the weight attached to the sensitivity. Choosing this weight may not be straightforward. For instance, in this study it is difficult to compare the consequences of both types of misclassification. Expression 3.17 is equivalent to equations 3.12 and 3.13 with θ set to 0.5, and equivalent to prevalence matching (defined later) with θ set to the population prevalence.

Minimax Criterion The minimax criterion involves minimising the frequency of the most common error. In a two by two classification table, this is equivalent to minimising the maximum of the off-diagonal elements. This involves minimising M in equation 3.18.

$$M = max\Big(\text{FPR}, \text{FNR}\Big) \tag{3.18}$$

This is similar to minimising the misclassification rate, except instead of the sum of FPR and FNR being minimised, the maximum of the two terms is minimised.

Prevalence Matching The final optimisation criterion we consider is to choose a cutpoint which results in the proportion of the screened population classified as positives (or screened prevalence) being closest to the gold standard case prevalence (i.e. the true prevalence). Those classed as positives comprise both true and false positives and so expression 3.19 is minimised, where the True Positive Rate (TPR) is the sensitivity multiplied by the case prevalence and $P(\text{Case})$ is the true prevalence.

$$|\text{TPR} + \text{FPR} - P(\text{Case})| \tag{3.19}$$

It is important to clarify at this point that, unlike in other studies which employ ROC curves, the area under the curve is not a meaningful criterion to use when attempting to identify a cutpoint on a scale. The area under the curve summarises the performance of an entire measure across all cutpoints. It is appropriate when two new measures are being compared against a gold standard in order to determine which of the new measures performs most similarly to the gold standard. It cannot, however, be used to determine an optimum cutpoint on a scale.

Since the method uses the same dataset both to define cutpoints and assess the

performance of those cutpoints, there is the possibility of overestimating the sensitivity and specificity. This potential source of bias is investigated by repeating the analysis using 75% of the dataset (randomly selected), and then assessing the performance of the cutpoints produced on the remaining 25% of the dataset.

Results of previous studies investigating a cutpoint for the MHI-5

One study of 7,359 adults representative of the Dutch general population used the GHQ-12 to derive a MHI-5 cutpoint using the prevalence matching method (Hoeymans et al., 2004). They used a less severe CMD case criterion of two or more on the GHQ-12, giving a screened case prevalence of 22.8%. The MHI-5 cutpoint which matched this prevalence most closely was 72, resulting in a case prevalence of 20.6%. To illustrate how this approach can lead to different results in different populations, we carried out the equivalent procedure in the BHPS dataset. Using a GHQ-12 caseness criterion of two or more classifies 32.9% of the dataset as cases. The MHI-5 cutpoint which best matches this prevalence is 76 (providing a case prevalence of 36.2%).

One small study compared four psychiatric case-finding instruments in 69 patients presenting to general practice in Wales and chose cutpoints which provided an undefined "similar sensitivity and specificity values for each instrument" (Winston & Smith, 2000). The Revised Clinical Interview Schedule was used to define a case of CMD. This study identified an MHI-5 cutpoint quoted as 71/72.

A report published in Dutch compared the MHI-5 with the GHQ-12 in order to ascertain a cutpoint (Perenboom et al., 2000). They sampled 7,065 independently living individuals aged 18 to 64 from the general population. A score of two or more on the GHQ-12 was used to define caseness, which classified 24.4% of the population as a case. They used the Youden Index and prevalence matching methods. The Youden Index indicated an MHI-5 cutpoint of 72, leading to a case prevalence of 22.8%. The Composite International Diagnostic Interview (CIDI) was used to determine whether individuals suffered from any of the following disorders: depression, bipolar disorder, dysthymia, panic disorder, agoraphobia, specific phobia, social phobia, generalised anxiety disorder, obsessive compulsive disorder, schizophrenia, anorexia and bulimia. The percentage of the population diagnosed with at least one of these disorders was found to be 12.2%. The MHI-5 cutpoint which matched this prevalence most closely was 60, producing a case prevalence of 11.2%.

Four other studies have defined a cutpoint by comparing MHI-5 scores with a range of different validated clinical interview schedules. These are summarised in turn below. The wide range of cutpoints found reflects the wide variety in sample sizes, study settings and outcomes of interest.

A study of 95 non-psychiatric patients who were HIV seropositive used the Struc-

Table 3.6: Summary of Previous SF-36 cutpoint Studies

Author	Cutpoint	Size	Population	Gold Standard
Hoeymans et al	72	7,539	Dutch general population	GHQ-12
Winston & Smith	71 or 72	69	Adult Welsh GP presenters	CIS-R
Perenboom et al	72	7,065	Dutch general population	GHQ-12
Holmes	52	95	American HIV+ outpatients	SCI
Rumpf et al	65	4,036	German general population	M-CIDI
Strand et al	52 or 56	6,875	Norwegian general population	MHI-5
Friedman et al	59 or 60	1,444	American functionally impaired community dwelling elderly	MINI-MDE

tured Clinical Interview for DSM-III-R (SCID-NP-HIV) psychiatric disorders (Holmes, 1998) and found a cutpoint of 52 using the (0,1) method. This study investigated more severe disorders than the CMD and so produced a very low cutpoint. Applying this cutpoint to the BHPS dataset would identify only 8.3% of the individuals as cases.

A study of 4,036 German nationals resident in an area of approximately 50 km in diameter surrounding Lübeck used the Munich Composite International Diagnostic Interview (M-CIDI) and found a cutpoint of 65 (Rumpf et al., 2001). This study used the (0,1) method. This low cutpoint can be attributed to the fact that the M-CIDI is used to screen for DSM-IV Axis 1 psychiatric disorders.

A Norwegian study used the MHI-5 as the gold standard to define cutpoints for other measures (three different versions of the Hopkins Symptom Checklist: SCL-25, SCL-10, SCL-5) (Strand et al., 2003). Postal questionnaire surveys with MHI-5 information were returned by 6,865 (70.5% response rate) individuals and cutpoints of 52 and 56 were used successively. The reason why two different cutpoints were used is that the authors cite the Holmes paper (Holmes, 1998) as well as a Dutch paper supporting the use of 52, and a poster session abstract and another unpublished study supporting the use of 56. This paper is included since they use a cutpoint on the MHI-5.

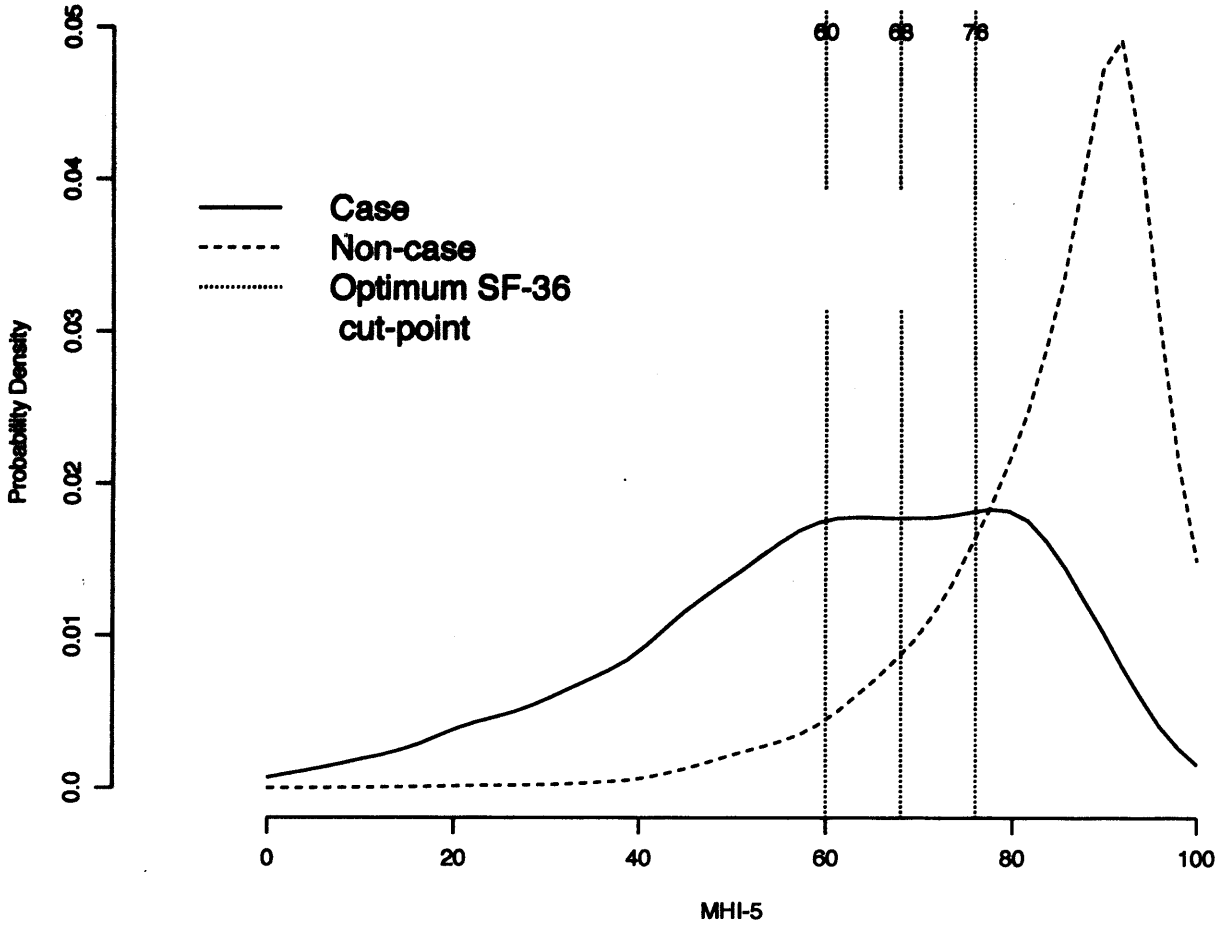
Another study (Friedman et al., 2005) investigated the validity of the MHI-5 for assessing major depression using 1,444 functionally impaired, community dwelling elderly Americans. The gold standard against which the MHI-5 was compared was the MINI-International Neuro-Psychiatric Interview Major Depressive Episode (MINI-MDE) module. The Youden index optimisation criterion produced a cutpoint quoted as 59/60. Again, the study focussed on major depression and so produced quite a low cutpoint.

To our knowledge no study has attempted to identify a cutpoint of the MCS.

Results

First consider the MHI-5. Figure 3.10 compares the MHI-5 score probability distributions for cases and non-cases (as defined by the GHQ-12) within this dataset. There is considerable overlap in MHI-5 scores for these two populations. The non-cases have

Figure 3.10: Probability distribution of cases and non-cases for the MHI-5, with optimal cutpoints



higher MHI-5 scores in general, but the cases have MHI-5 scores spread right throughout the range of possible scores, resulting in a relatively flat distribution. Consequently, no cutpoint on the MHI-5 will lead to complete separation of the two underlying populations, as is typically the case in ROC curve analyses. Maximising the Youden index leads to a cutpoint of 76 (a case of common mental disorder is defined by a score of less than or equal to 76) for the MHI-5. This results in a sensitivity of 0.756, a specificity of 0.771 and a misclassification rate of 23.3%. Figure 3.11 shows the distribution of MHI-5 scores in the two underlying populations (cases and non-cases as assessed by the GHQ-12). This graph is similar to Figure 3.10, but is weighted by the observed prevalence (note that the y-axis is population frequency and not population density). Again there is considerable overlap in MHI-5 scores for these two populations.

Using the shortest distance from (0,1) criterion the optimal cutpoint is also 76. In

Figure 3.11: Population distribution of cases and non-cases for the MHI-5, with optimal cutpoints

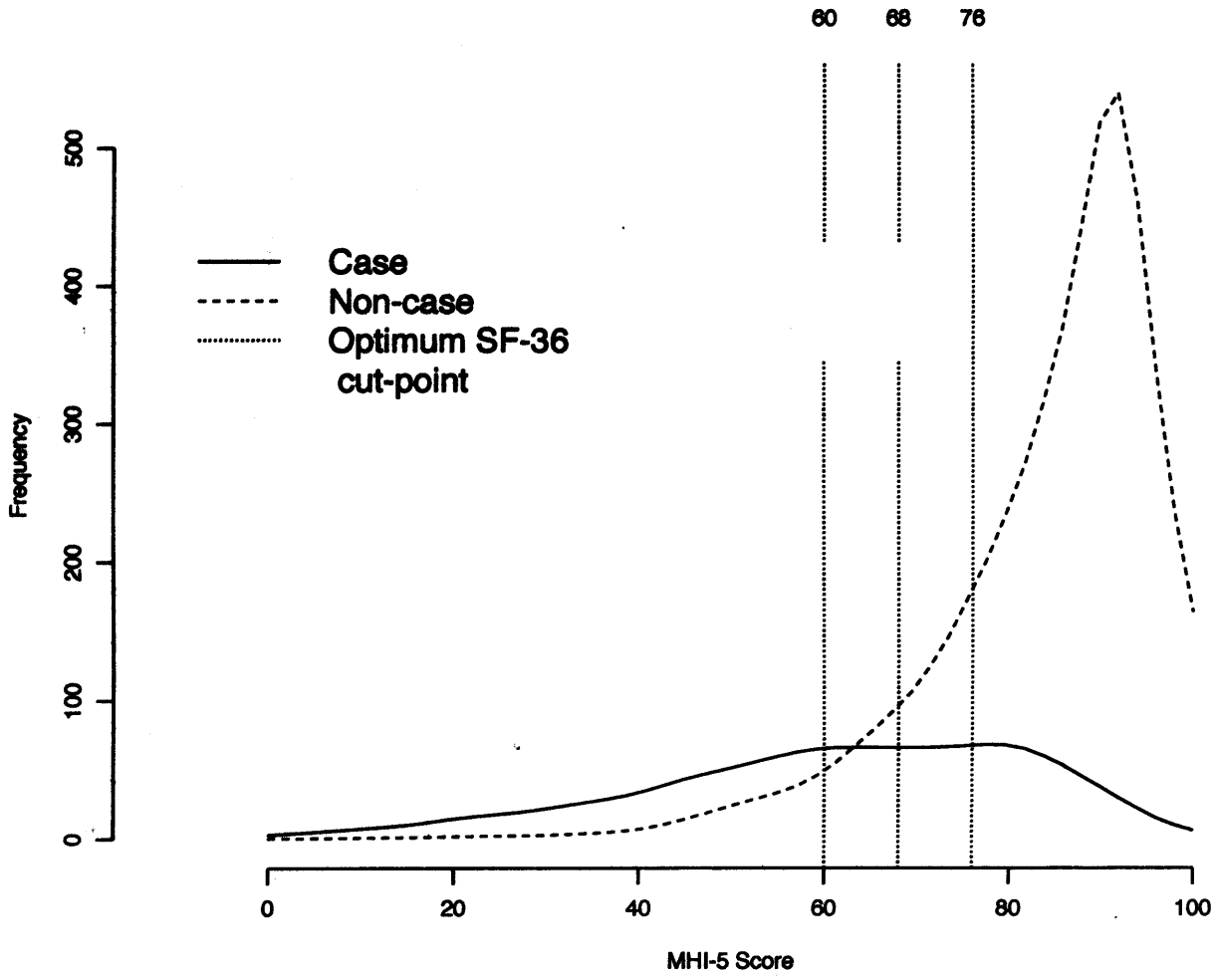


Table 3.7: MHI-5 and MCS cutpoints and associated test characteristics for five optimisation criteria

Scale	Optimisation Criterion	Cutpoint	Sensitivity	Specificity	Positivity[a] Rate	Error Rate[b] %
MHI-5	Youden Index	76	0.756	0.771	0.362	23.3
	(0,1) [c]	76	0.756	0.771	0.362	23.3
	Misclassification Rate	60	0.473	0.943	0.163	17.6
	Minimax method	68	0.615	0.882	0.244	18.5
	Prevalence Matching	68	0.615	0.882	0.244	18.5
MCS	Youden Index	51.7	0.745	0.787	0.348	22.4
	(0,1)	52.1	0.759	0.772	0.362	23.1
	Misclassification Rate	44.8	0.476	0.941	0.164	17.6
	Minimax method	48.9	0.630	0.874	0.253	18.8
	Prevalence Matching	48.9	0.630	0.874	0.253	18.8

[a]Positivity rate refers to the proportion of the sample defined to be a case using each cutpoint
[b]Error rate refers to the proportion of the sample classified differently to the GHQ-12. This comprises both false negatives and false positives
[c](0,1) refers to the criterion which minimises the distance between the point (0,1) and the ROC curve

general, these two optimisation methods will not always give the same cutpoint though the discrete nature of both scales means that in practice they often will.

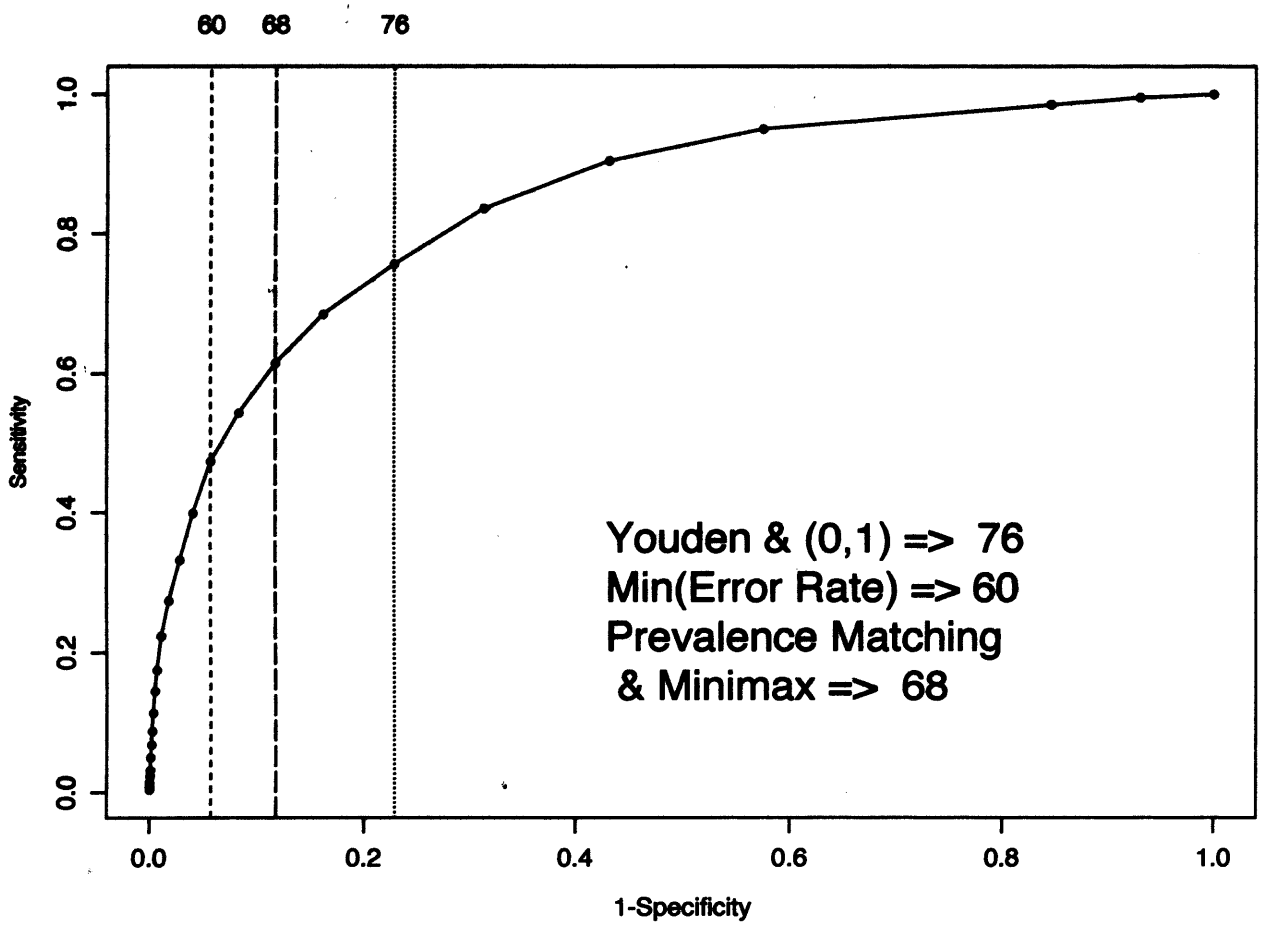
Using the sample prevalence of 25.3% (according to the GHQ-12) to weight the sum of the sensitivity and specificity (thereby minimising the error rate) the corresponding cutpoint is 60. This results in a low sensitivity of 0.473, and a high specificity of 0.943. The misclassification rate is 17.6%.

Using the prevalence matching method of choosing a cutpoint the optimal cutpoint is 68. This produces a case prevalence of 24.4%, which is the closest to the GHQ-12 case prevalence of 25.3%. The minimax method yields the same cutpoint as prevalence estimation. The sensitivity and specificity at this cutpoint are 0.615 and 0.882 respectively.

These cutpoints are plotted on the ROC curve in figure 3.12, and summarised in table 3.7. Two by two crosstabulations between the GHQ-12 and the MHI-5 are provided for all of the cutpoints in tables 3.8 - 3.10.

Next, we examine the MCS, where the situation is similar, with the distribution of cases and non-cases overlapping. MCS probability distributions for cases and non-cases are shown in figure 3.13. This figure should be compared with 3.14 which shows

Figure 3.12: MHI-5 ROC curve using a GHQ caseness criterion of 3 or more



1. ROC curve based on a GHQ-12 caseness criterion of 3 or more. Vertical lines indicate the optimum cutpoints using the five different optimisation criteria.

the population distribution of the MCS for cases and non-cases. The Youden index and the (0,1) methods produce very similar cutpoints of 51.7 and 52.1, respectively. Minimising the error rate produces a cutpoint of 44.8 while both prevalence matching and the minimax method indicate a cutpoint of 48.9. Table 3.7 summarises the results and illustrates the trade off that must be made between sensitivity and specificity, while figure 3.15 illustrates the ROC curve. Two by two crosstabulations between the GHQ-12 and the MCS are provided for all of the cutpoints in tables 3.11 - 3.14.

Assessment of bias

Using three-quarters of the data to derive cutpoints resulted in no change of optimum cutpoints for the MHI-5. When these were applied to the unused 25% of the data in order to assess their performance the sensitivities all decreased, while the specificities increased, as shown in table 3.15. In this table, numbers bounded by transparent boxes indicate an decrease over the equivalent value reported in table 3.7, numbers bounded by shaded boxes indicate a increase, and numbers surrounded by no box indicate no change. The differences for the sensitivity range between 0.01 and 0.029, while the differences for the specificity range between 0.002 and 0.004. A similar split analysis was performed for the MCS. Only one method (the missclassification rate) produced a different cutpoint to the when the full dataset was used. Applying these cutpoints to the remaining 25% of the data produced sensitivities and specificities close to those achieved when the entire dataset is used. This procedure was repeated using one half and one quarter of the dataset to derive cutpoints (the training set) and the remaining data to assess the sensitivities and specificities. None of the MHI-5 cutpoints changed, while some of the MCS cutpoints changed. The sensitivities and specificities were as likely to increase as decrease. This provides evidence that the sensitivities and specificities for the optimum cutpoints are not overestimated.

Discussion

Main Findings For the MHI-5 the five methods produce three distinct cutpoints. Both the Youden Index and the point closest to the upper left corner methods produce

Table 3.8: Crosstabulation of MHI-5 (cutpoint 76 for Youden Index and (0,1)) and GHQ-12 (cutpoint 3)

	MHI-5					
GHQ-12	**Non-case**	**%**	**Case**	**%**	**Row Total**	**%**
Non-case	8,450	57.6	2,510	17.1	10,960	74.7
Case	904	6.2	2,805	19.1	3,709	25.3
Column Total	9,354	63.8	5,315	36.2	14,669	100.0

Table 3.9: Crosstabulation of MHI-5 (cutpoint 60 for Min(Error Rate)) and GHQ-12 (cutpoint 3)

	MHI-5					
GHQ-12	**Non-case**	**%**	**Case**	**%**	**Row Total**	**%**
Non-case	10,331	70.4	629	4.3	10,960	74.7
Case	1,953	13.3	1,756	12.0	3,709	25.3
Column Total	12,284	83.7	2,385	16.3	14,669	100.0

Table 3.10: Crosstabulation of MHI-5 (cutpoint 68 for prevalence matching and minimax)) and GHQ-12 (cutpoint 3)

	MHI-5					
GHQ-12	**Non-case**	**%**	**Case**	**%**	**Row Total**	**%**
Non-case	9,669	65.9	2,281	15.5	10,960	74.7
Case	1,428	9.7	1,291	8.8	3,709	25.3
Column Total	11,097	75.6	3,572	24.4	14,669	100.0

Table 3.11: Crosstabulation of MCS (cutpoint 51.7 for Youden Index) and GHQ-12 (cutpoint 3)

	MCS					
GHQ-12	**Non-case**	**%**	**Case**	**%**	**Row Total**	**%**
Non-case	8,622	58.8	2,338	15.9	10,960	74.7
Case	946	6.4	2,763	18.8	3,709	25.3
Column Total	9,568	65.2	5,101	34.8	14,669	100.0

Table 3.12: Crosstabulation of MCS (cutpoint 52.1 for (0,1) method) and GHQ-12 (cutpoint 3)

	MCS					
GHQ-12	**Non-case**	**%**	**Case**	**%**	**Row Total**	**%**
Non-case	8,465	57.7	2,495	17.0	10,960	74.7
Case	893	6.1	2,816	19.2	3,709	25.3
Column Total	9,358	63.8	5,311	36.2	14,669	100.0

Table 3.13: Crosstabulation of MCS (cutpoint 44.8 for Min(error)) and GHQ-12 (cutpoint 3)

	MCS					
GHQ-12	**Non-case**	**%**	**Case**	**%**	**Row Total**	**%**
Non-case	10,317	70.3	643	4.4	10,960	74.7
Case	1,943	13.2	1,766	12.0	3,709	25.3
Column Total	12,260	83.6	2,409	16.4	14,669	100.0

Figure 3.13: Probability distribution of cases and non-cases for the MCS, with optimal cutpoints

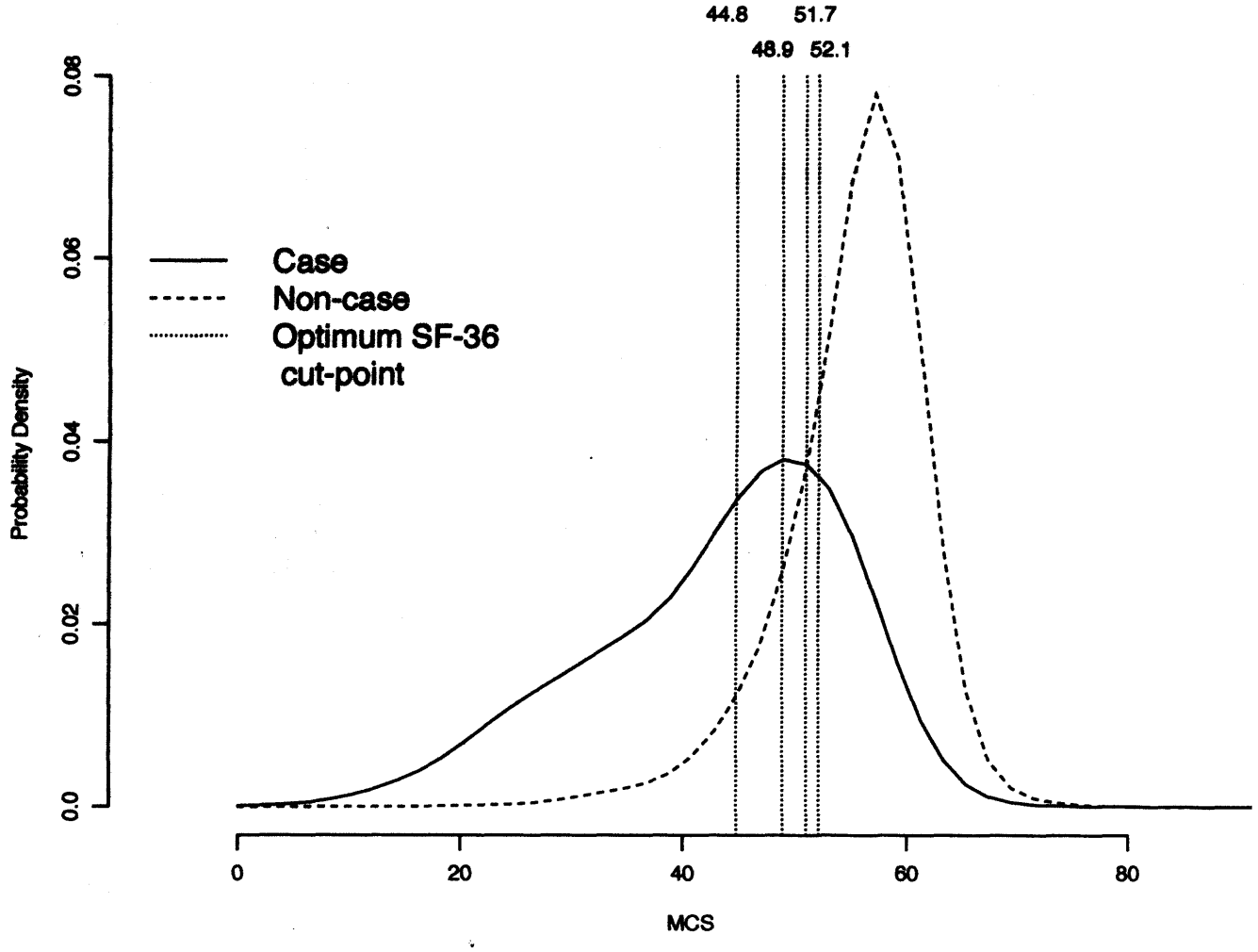


Table 3.14: Crosstabulation of MCS (cutpoint 48.9 for prevalence matching and minimax methods) and GHQ-12 (cutpoint 3)

	MCS					
GHQ-12	**Non-case**	**%**	**Case**	**%**	**Row Total**	**%**
Non-case	9,581	65.3	1,379	9.4	10,960	74.7
Case	1,374	9.4	2,335	15.9	3,709	25.3
Column Total	10,955	74.7	3,714	25.3	14,669	100.0

Figure 3.14: Population distribution of cases and non-cases for the MCS with optimal cutpoints

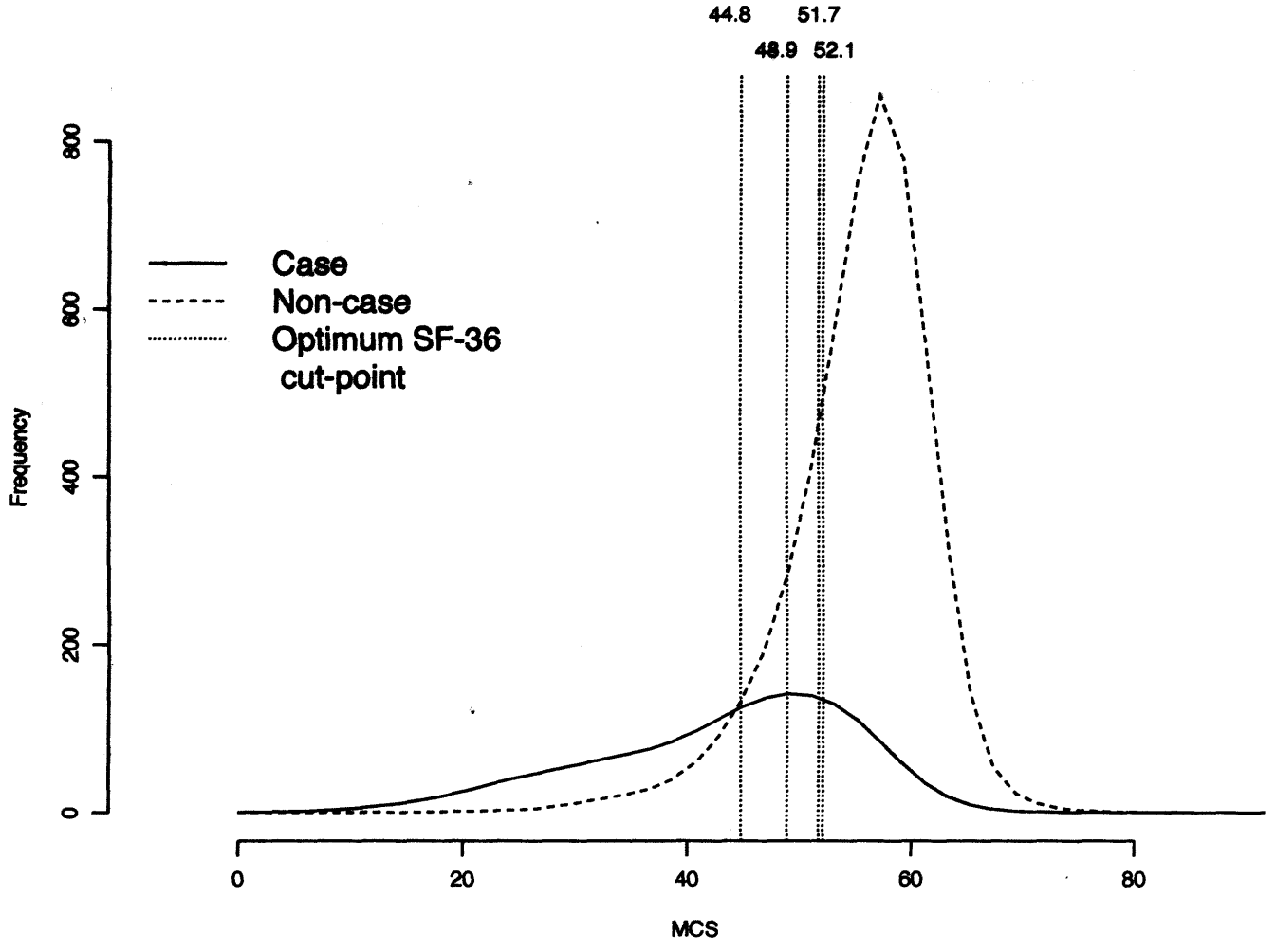
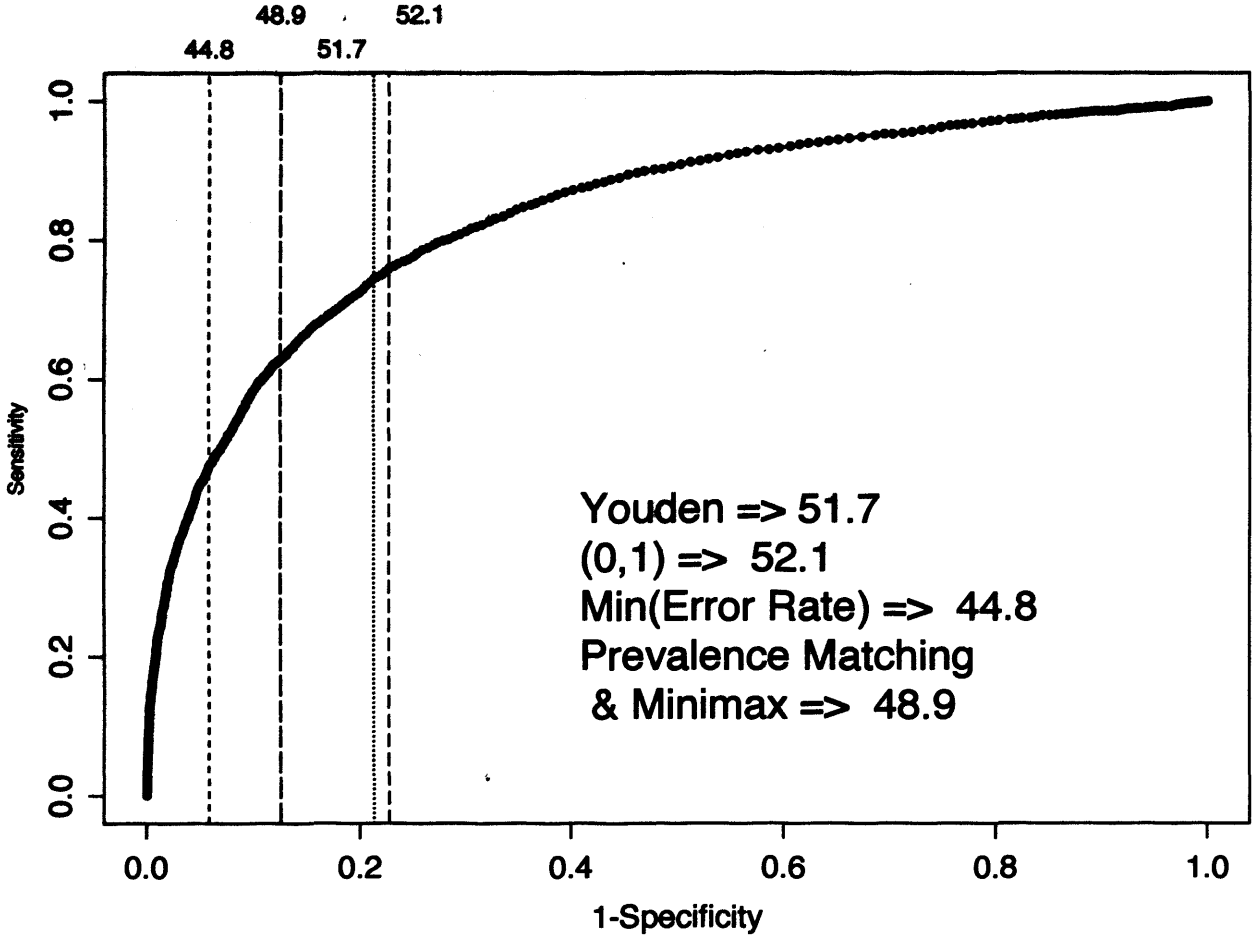


Figure 3.15: MCS ROC curve using a GHQ caseness criterion of 3 or more



1. ROC curve based on a GHQ-12 caseness criterion of 3 or more. Vertical lines indicate the optimum cutpoints using the five different optimisation criteria.

Table 3.15: Sensitivity and specificity for split analyses

Mental Health Scale	Method	75% training dataset Cutpoints	Sens	Spec	50% training dataset Cutpoints	Sens	Spec	25% training dataset Cutpoints	Sens	Spec
MHI-5	Youden Index	76	0.727	0.775	76	0.749	0.773	76	0.753	0.775
	(0,1)	76	0.727	0.775	76	0.749	0.773	76	0.753	0.775
	Misclassification Rate	60	0.463	0.946	60	0.475	0.940	60	0.480	0.943
	Minimax Method	68	0.588	0.884	68	0.614	0.879	68	0.616	0.885
	Prevalence Matching	68	0.588	0.884	68	0.614	0.879	68	0.616	0.885
MCS	Youden Index	51.7	0.738	0.787	52.1	0.754	0.773	51.8	0.748	0.786
	(0,1)	52.1	0.754	0.772	52.1	0.754	0.773	51.8	0.748	0.786
	Misclassification Rate	45.1	0.492	0.942	45.1	0.481	0.937	44.5	0.466	0.944
	Minimax Method	48.9	0.633	0.872	48.8	0.624	0.878	48.6	0.622	0.883
	Prevalence Matching	48.9	0.633	0.872	48.8	0.624	0.878	48.6	0.622	0.883

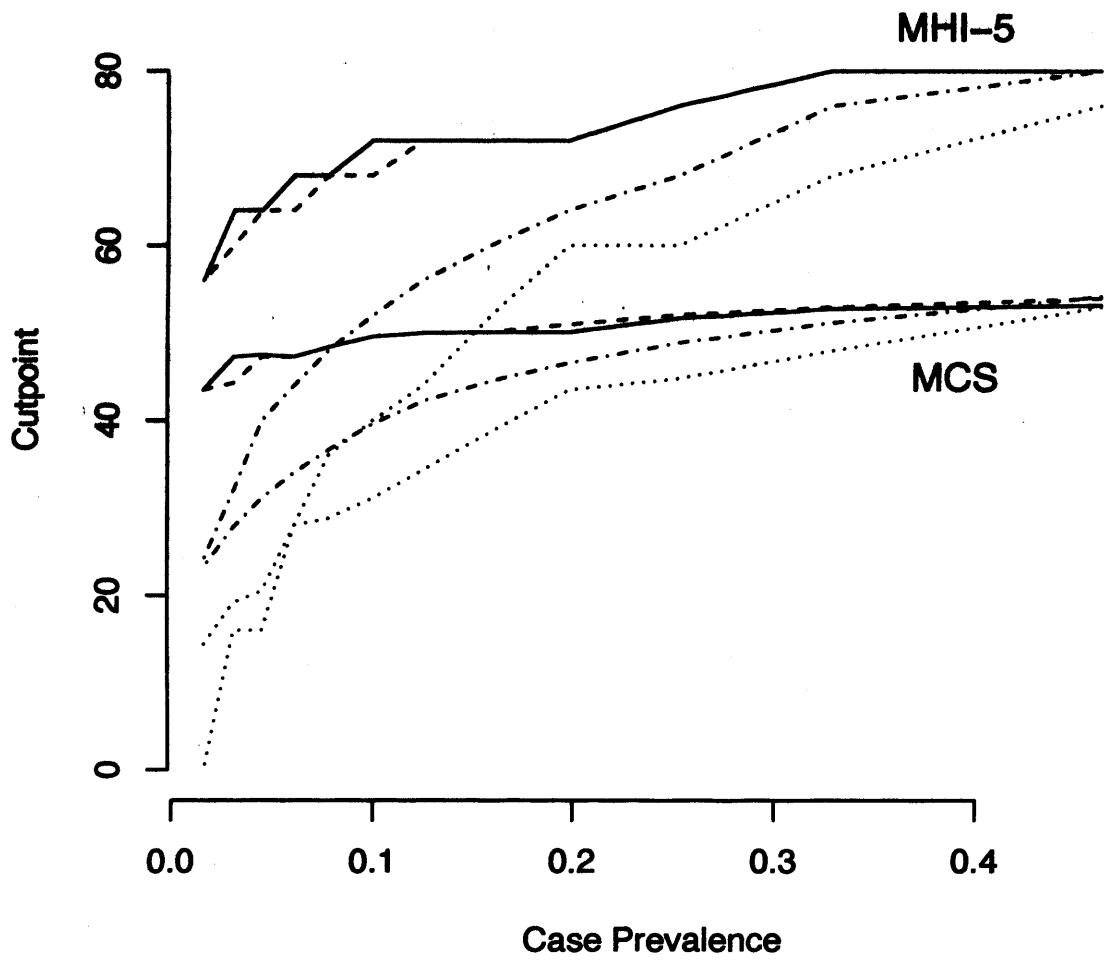
1. Cells bounded by a transparent box indicate an decrease from the equivalent value when the entire dataset is used
2. Cells bounded by a shaded box indicate an increase from the equivalent value when the entire dataset is used
3. Unbounded cells indicate no change from the equivalent value when the entire dataset is used

a cutpoint of 76. Prevalence matching and the minimax method both indicate that 68 is the optimal cutpoint, while minimising the misclassification rate provides an optimal cutpoint of 60. For the MCS, the Youden Index and the point closest to the upper left corner produced cutpoints of 51.7 and 52.1 respectively. Both prevalence matching and the minimax method gave an MCS cutpoint of 48.9 and minimising the missclassification rate produced an MCS cutpoint of 44.8. It is important to point out that the reason the prevalence matching, minimax, and minimising the misclassification methods give lower cutpoints than the Youden or (0,1) criteria is due to the fact that the case prevalence is less than 50% (25.3%). If the case prevalence were greater than 50% this situation would be reversed and the three aforementioned methods would give cutpoints greater than the Youden or (0,1) criteria. It is also worth noting that the shorter MHI-5 performs remarkably similarly to the longer MCS. Table 3.7 shows that the error rates produced by the five optimisation methods are very similar, for the two scales. The correlation between the GHQ-12 and the MHI-5 is high (-0.65). The correlation between the MCS and GHQ-12 is also (-0.65). The correlation between the MHI-5 and the MCS is 0.88. Table 3.7 shows that the MCS is only marginally more efficient at discriminating cases of CMD than is the MHI-5, despite employing over seven times as many questions.

Figure 3.16 compares the relationship between the optimum cutpoint and the population case prevalence for four of the five methods (the minimax method was excluded since it was largely coincidental with prevalence matching), and for both the MHI-5 and the MCS. For the minimising the error rate and prevalence matching methods the optimal cutpoint varies greatly with population prevalence, while the Youden index and (0,1) methods are relatively independent of population prevalence. This is the case for both scales. This invariance under different population prevalences is a property that is extremely useful for studies that span large and heterogenous areas, such as international comparisons. Both methods also have intuitive interpretations as described earlier, and so there is very little to choose between them.

When the misclassification rate was minimised there was still a error rate of 17.6% for both the MHI-5 and MCS, which may imply that they measure slightly different constructs to the GHQ-12. This finding is echoed by Hoeymans et al (2004) who noted that the MHI-5 was uncorrelated with age, whereas older age groups scored higher on the GHQ-12 (indicating worse mental health). Weinstein et al (1989) drew attention to the fact that the comparative nature of the GHQ-12 response choices is not conducive to detecting chronic disorders. A subject suffering from chronic anxiety disorder may well answer the question "Have you recently lost much sleep over worry?", with the response choice "no more than usual" if their condition is a long-standing one. The MHI-5 and MCS avoid this problem by employing less comparative response choices. Another explanation for the lack of complete agreement between the GHQ-12 and the

Figure 3.16: Relationship between prevalence and MHI-5 cutpoint for five optimisation methods



1. Case prevalence is altered by varying the cutpoint used to define caseness on the GHQ-12 from 1 to 12
2. Solid line denotes the Youden Index
3. Dashed line denotes the (0,1) method
4. Dotted line denotes the minimising the error rate method
5. Dashed and dotted line denotes the prevalence matching method
6. The minimax method is excluded since it is predominantly coincidental with the prevalence matching method

two SF-36 mental health measures is that they were designed differently. The MHI-5 includes one or more questions on each of the following mental health dimensions: anxiety, depression, loss of behavioural/emotional control and psychological well-being (Ware et al., 2000b), while the MCS is a weighted sum of all eight health dimensions of the SF-36. The GHQ-12 on the other hand includes items on depression, anxiety, social performance and somatic complaints (Goldberg & Williams, 1988). However, the high correlations between the GHQ-12 and both the MHI-5 and the MCS indicate that, despite these differences, the three scales perform very similarly, as shown by table 3.7.

More generally, this study has found that the minimax method and prevalence matching methods give very similar results. Indeed, in this study they produce identical cutpoints. This is not a coincidence, as the two criteria become equivalent if the scale in question is continuous (and the probability of caseness is calculated from the same dataset).

Investigators should give careful consideration to which of these cutpoints is most appropriate for their study, since selecting which criterion should be optimised depends primarily on the intended application of the resulting cutpoint. For instance, a study whose primary goal is to identify cases in a given locality might do well to minimise the misclassification rate. However, a study interested in comparing CMD internationally should consider utilising the Youden Index or the (0,1) method, as these methods are most appropriate when the study area encompasses regions with different case prevalences. Prevalence matching has the advantage of simplicity but will inevitably lead to different cutpoints in different populations. The minimax method approximates to prevalence matching when the scale in question is continuous.

Summary

The User's Guide for the GHQ recommends that investigators who want to optimise the trade-off between sensitivity and specificity should carry out a validity study to determine the optimum GHQ for their population. This is clearly excellent advice, and is borne out by this study. Of the five optimisation methods used in this study, the Youden Index and the (0,1) method are the most suitable for the determination of a generalisable cutpoint, since they are least dependent on the population case prevalence. Both approaches indicate that the best cutpoint to define a case of CMD using the MHI-5 is less than or equal to 76, while for the MCS the Youden Index indicates a cutpoint of less than or equal to 51.7 and the (0,1) method a cutpoint of less than or equal to 52.1. The MHI-5 has the advantage over the GHQ-12 of brevity, consisting of only five multiple choice questions and performs very similarly to the longer MCS. Further validation studies, ideally using a clinical interview schedule and

a large population, spanning different countries, are required to confirm these findings.

3.5 Conclusion

This chapter has investigated the validity and reliability of the MHI-5 as well as examining various different approaches to modelling it. Firstly the background of the SF-36 was summarized. Next, some standard techniques for assessing validity and reliability were assessed. The literature surrounding the validation of the SF-36 was appraised critically under the following sub-headings: Comparison with other scales, Comparison across different subject groups, Cronbach's Alpha, Test-retest, Suitability for Elderly Populations, and Version 1 versus Version 2. The validity and reliability of the MHI-5 were shown to be of a high standard. The MHI-5 compared favourably with many other mental health scales. It was not deemed suitable for use with the over 75's due to missing values. A comparison between versions 1 and 2 of the SF-36 show that they perform similarly (Ware et al., 2000a). In summary, the MHI-5 appears to be a well-validated and reliable tool to measure mental health status.

Transforming the data reduced the skewness of the mental health score, but complicated the interpretation of the results of any regression performed on the transformed data. The transform most effective at reducing the skewness was a square transform. The reduction in skewness did not justify the increased difficulty of interpreting the results.

Ordinal modelling suffered from a similar problems of interpretation of results. The mental health scale would have to be split into a relatively small number of categories in order to be ease the interpretation of the results. This means a reduction in the response information. The method also involves strict assumptions and it is difficult to assess if these assumptions are satisfied.

The final method investigated regards deriving a cutpoint to define a case of common mental disorder on both the MHI-5 and MCS from the GHQ-12 using information from the ninth wave of the BHPS. ROC curve analysis was employed for this purpose. Five different optimisation methods were applied to the data and the results compared and contrasted. The best cutpoint for the MHI-5 was less than or equal to 76. While this cutpoint had a high error rate, it was the least dependent on population prevalence. For the MCS, the best cutpoint to define a case of CMD was less than or equal to 51.7. Moreover the best optimisation criterion appears to be either the Youden Index or the point closest to the upperleft corner. These methods depend least on the underlying population prevalence. Further work needs to be done to validate these cutpoints, ideally incorporating comparison with a clinical interview schedule. Normal modelling is used in the rest of the thesis, since the increased complication of the interpretation of the results from the other methods is not deemed justified.

Chapter 4

Introduction to Bayesian Modelling

4.1 Background

Classical inference begins with the assumption that the parameters to be estimated in a statistical method are fixed (but unknown) quantities. Hypothesis testing is performed by assuming a null hypothesis, then assessing the probability that the data could have been observed under this null hypothesis. In many situations this is an entirely reasonable and practical approach.

Bayesian inference treats the problem in a slightly different way. Here, both the parameters to be estimated and the data itself are treated as random quantities. In Bayesian modelling, the data are used to make inference about the distribution of the underlying parameter of interest. A way of contrasting the two methods is that, in classical inference, statements are made about the probability of observing the data given a parameter (and a model), while in Bayesian inference, probability statements are made about the parameter given the data.

There is another fundamental difference between the two methods. In classical inference, the parameters are estimated solely from the data. Bayesian inference, on the other hand, incorporates so-called "prior information" into the analysis. Prior information is any information that is known (or believed to be known) about the parameters before the data are collected. This could be as simple as the knowledge that a proportion must be positive, or as detailed as specifying a range of possible values that a parameter can assume. Prior information is usually based on expert opinion, previous analyses, or sometimes chosen for convenience of computation. It is typically incorporated into the analysis by specifying a probability density function (PDF) for a parameter of interest. For instance, if a given parameter is expected to be Normally distributed and much is known about the true value of that parameter, then the prior PDF might be a Normal distribution, centred around the value the parameter is suspected to be, with a small variance. Conversely, if little is known about a parameter,

then the prior PDF might have a large variance. The former is called an informative prior, while the latter is an example of a non-informative prior.

The legitimacy of including such information and the form such information should take are frequently disputed. Bayesian methods are seen as more subjective than frequentist techniques (Efron, 1986). Bayesian proponents can counter this criticism in a number of ways. Firstly, non-informative priors are used in many analyses, negating much of the need to justify their use. Secondly, if the prior chosen is informative, then appropriate sensitivity analyses should be performed to give an indication of how dependent the results are on the prior. Finally, there are many situations where it could be considered unethical to disregard information from previous studies, and so incorporating them as prior information could be viewed as an advantage.

Another crucial difference between the frequentist and Bayesian methods involves how each method expresses uncertainty about the parameter of interest. Confidence intervals are used in frequentist analysis while credible intervals are used in Bayesian analysis. A 90% confidence interval around a parameter, indicates that if the study were repeated a large number of times, with different data being collected each time, then 90% of the confidence intervals estimated would contain the true parameter. It should not be interpreted to mean that we are 90% sure that the true value lies within the interval, since the true parameter is considered to be a fixed quantity. A 90% credible interval on the other hand means that the posterior probability that the parameter lies within the interval is 0.9. The 90% here can be interpreted as the percentage probability that the parameter lies in the interval.

To summarise, in certain situations Bayesian methods can provide a much more natural way of approaching hypothesis testing and inference. By allowing the parameters of interest to belong to probability distributions (instead of being fixed constants) probability statements can be made about where those parameters might lie. This is in contrast with the frequentist approach which treats parameters of interest as fixed quantities, and makes statements about the probability of observing the data under various null hypotheses. The next section will provide more detail on Bayesian methods.

4.2 The Bayesian Method

As discussed in the previous section, Bayesian analysis utilises two types of information: observed data and prior information. These two sources of information are combined via Bayes' Theorem (giving the method its name), to provide a posterior probability for the parameter being estimated, as described in equation 4.1, where A and B are

two events.

$$Pr(A|B) = \frac{Pr(B|A)Pr(A)}{Pr(B)} \tag{4.1}$$

Here, $Pr(A|B)$ is the posterior probability, $Pr(A)$ is the prior information, $Pr(B)$ is the probability of event B from the data itself. This formula therefore, allows the probability for event A to be updated, given some additional information on the related event B. In other words we have refined our knowledge about event A using additional information about event B. For continuous distributions the theorem can be restated so that the PDF of a parameter θ is updated given observed information x as in equation 4.2

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{f(x)} \tag{4.2}$$

Invoking the law of total probability in the denominator gives equation 4.3.

$$f(\theta|x) = \frac{f(x|\theta)f(\theta)}{\int f(x|\theta)f(\theta)d\theta} \tag{4.3}$$

Here, $f(x|\theta)$ is the usual probability model of x given some parameter (or parameters) θ, $f(\theta)$ represents the prior information for θ, and $f(\theta|x)$ is the posterior distribution for θ revised in light of the data x. Once the posterior distribution has been estimated, it is often summarised for ease of use. Typically, the mean of the distribution is used as a summary statistic. It should be noted here that equations 4.2 and 4.3 abuse conventional notation, since the f does not represent a single function, but is instead used to denote the density function of whatever parameter it references. The posterior distribution can be evaluated, provided that this integral has an analytical solution, which is not guaranteed. In the past, the prior information was often chosen so that the resulting posterior distribution could be easily calculated. Choosing a prior with a functional form such that the resulting posterior distribution has the same functional form is known as choosing a conjugate prior. If the data are Gaussian, the choice of a Gaussian prior results in a Gaussian posterior distribution. If the data are binomial, the choice of a beta distribution results in a beta posterior distribution.

It is worth examining a simple analytical case in order to illustrate the role of the prior. Suppose a sample of size n is taken from a normal distribution, X, with known standard deviation,τ, but unknown mean. The conjugate prior is a normal distribution. Suppose a normal prior is assumed with mean μ and standard deviation σ. Using equation 4.3, the prior information and the data are combined, producing a normal posterior distribution, with mean and standard deviation as given in expression 4.4.

$$\text{Mean} = \frac{n\bar{x}\sigma^2 + \mu\tau^2}{n\sigma^2 + \tau^2} \qquad \text{Standard Deviation} = \frac{\sigma\tau}{\sqrt{n\sigma^2 + \tau^2}} \tag{4.4}$$

To further clarify the way information from both the prior and the data are combined to form the posterior distribution, the mean can be rewritten as in equation 4.5.

$$\lambda\bar{x} + (1-\lambda)\mu \qquad \text{where} \quad \lambda = \frac{n\sigma^2}{n\sigma^2+\tau^2} = 1 - \frac{\tau^2}{n\sigma^2+\tau^2} \to 1 \quad \text{as} \quad n \to \infty \tag{4.5}$$

This reformulation shows that as the sample size of the data increases, the posterior mean is dominated by the sample mean. The posterior mean is a weighted average of the prior mean and the mean of the data, with the latter given increasing weight as the sample size increases. Increasing σ (the prior distribution standard deviation) also leads to more weight being given to the data, which is sensible since increasing the prior standard deviation means that there is greater uncertainty about the prior mean. In summary, the data will dominate the posterior distribution if there is a large sample size, or if the prior information is uncertain.

The posterior distribution variance can be similarly reformulated to illustrate the effect of both the prior and the data. The standard deviation given in equation 4.4 is rewritten in equation 4.6.

$$\frac{\tau}{\sqrt{n+m}} \qquad \text{where} \quad m = \frac{\tau^2}{\sigma^2} \tag{4.6}$$

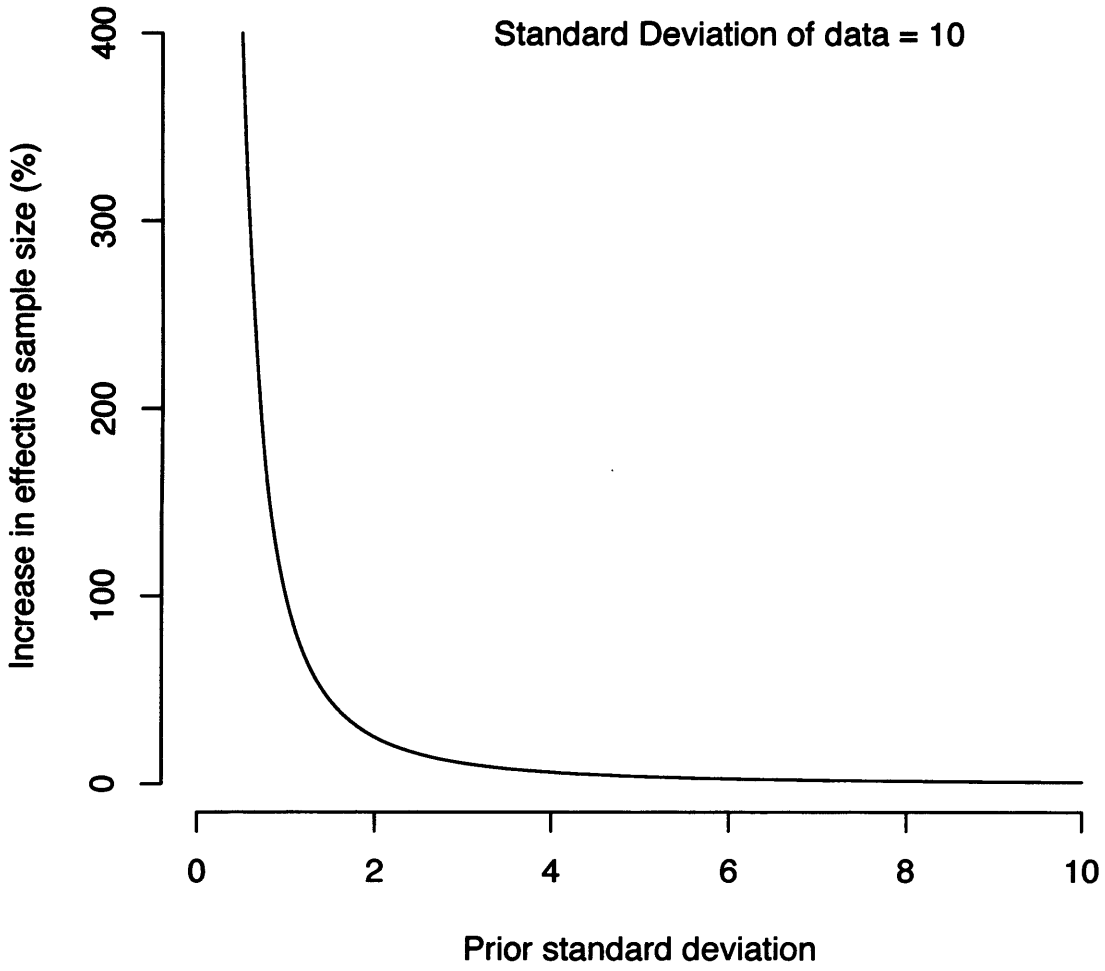
This is the expression for the standard deviation of a sample of size $n+m$ from the distribution of X. Essentially, the extra information from the prior has augmented the sample size of the data by m. As shown in equation 4.6, m is the ratio of the variance of the data and the prior. If the variance of the prior is large (indicating that the prior information is not very informative), then m will be small, and will provide only a small increase in the effective sample size. If the information on the prior is more exact, the variance will be small, and so the extra effective sample size will be large. This relationship is illustrated in figure 4.1, for the case where the standard deviation of X is 10.

If either the sample size of the data or the standard deviation of the prior distribution are large then the contribution of the prior is weakened. In such situations the resulting posterior distribution will predominantly be a result of the data itself, and the results of the Bayesian analysis should mirror classical frequentist methods.

Until recently the choice of prior distribution was a trade-off between choosing a distribution that was a reasonable representation of reality (or, more precisely, beliefs about reality) and choosing a distribution that would simplify the calculation of the posterior distribution.

Recent advances in computing power and software have produced a method called Markov Chain Monte Carlo (MCMC) estimation which uses a simulation method to estimate the posterior distribution. This means that the prior information no longer

Figure 4.1: Relationship between the standard deviation of the prior and the resulting increase in the effective sample size



needs to belong to a conjugate distribution since the evaluation of complex integrals is no longer necessary. This method is a considerable improvement over the analytical approach in many situations, since the choice of prior information distribution is no longer restricted by the distribution of the data. The next section, 4.3, will summarise the MCMC method.

Once a large enough sample from the posterior distribution for a given parameter has been taken it can be used to make inferences about that parameter. Having the distribution of a random variable provides all of the information about that variable and so the posterior distribution can be used to make inference about that parameter, e.g. credible intervals can be calculated, hypothesis tests can be performed and summary statistics can be extracted.

So, the Bayesian method can be summarised as follows. Observed data are combined with prior information via Bayes' theorem, to produce a posterior distribution for an unknown parameter (or parameters) which summarises both prior beliefs and current information. Often, the analytical procedure for combining the two types of information is complicated or intractable and so other methods have been developed to avoid this problem, as will be described in the next section.

4.3 Estimation of Bayesian Models

4.3.1 Monte Carlo Markov Chains

It is instructive to begin with a description of Markov chains and their properties. A Markov chain *"describes an idealised pattern of movement or transitions through a set of states"*(Congdon, 2001). In other words, a Markov chain comprises a set of possible states and a set of probabilities of transitions between these sets (transition matrix). The probability of any transition depends solely on the current state occupied, and not on any historic transitions (known as the no memory or Markov property). Suppose the Markov chain undergoes a number of transitions. The probability a given state is occupied after these transitions depends on the initial probabilities and the transition matrix. It can be shown that under certain conditions (Gamerman, 1997), these state probabilities tend to a limiting distribution, which coincides with the stationary distribution. The key step in MCMC methods is to construct a Markov chain whose stationary distribution is the posterior distribution of interest. Instead of analytically deriving the posterior distribution from the data and the prior information, a sample from the posterior distribution can be obtained by simulating a Markov chain with the required stationary distribution. In summary, since limiting distributions of Markov chains are well understood, this method provides easy access to the posterior distribution of analytically intractable problems. The key issues therefore are how to construct the transition matrix that leads to the desired stationary distribution, and how many transitions are sufficient so that the resulting limiting distribution is close enough to the stationary distribution. These will be described in the next two sections: Metropolis-Hastings and Gibbs Sampling, and Convergence, respectively.

4.3.2 Metropolis-Hastings and Gibbs sampling

The Metropolis-Hastings algorithm (?) is one way to generate a Markov chain. The basic premise is quite simple. Given that state i is occupied, a next possible state (called a candidate) is selected from a distribution of possible states called the proposal distribution. The candidate state is accepted with a probability which depends

on both the current state and the candidate state. Careful construction of this acceptance probability ensures that the resulting Markov chain has the desired stationary distribution.

The choice of proposal distribution has an impact on how quickly the limiting distribution is attained. Proposal distributions which favour candidate states close to the current state have high acceptance probabilities but also move slowly. A slow moving chain will take a long time to attain the limiting distribution. On the other hand a proposal distribution that allows large jumps between successive states will permit the chain to move quickly but such large jumps will have a low acceptance probability.

A special case of the Metropolis-Hastings algorithm is called the Single Component Metropolis-Hastings algorithm. This is used when the situation calls for multiple parameters to be estimated simultaneously, as is the case in most models. So in a situation with k parameters of interest each state represents a k-dimensional point. The single component Metropolis-Hastings algorithm provides an efficient way to generate suitable Markov chains in these situations.

Specialising even further, Gibbs sampling (which is the method employed in this chapter), is a special case of the single component Metropolis-Hastings algorithm. Again this is used when there are multiple parameters to be estimated. Under Gibbs sampling the proposal distribution for each parameter is the full conditional distribution given the current values of all of the other variables and the candidate state is accepted with probability 1. In other words the candidate state is always accepted. Each component is adjusted in turn at each iteration, and the proposal distribution updated for the next parameter.

4.3.3 Convergence

Once the Markov chain has been constructed, there remains the practical consideration of assessing when the chain has reached convergence thereby allowing the limiting distribution to be estimated. This is not straightforward since determining when the process is sampling from a distribution is difficult. The time until the chain has converged is referred to as the burn-in period. The end of this period must be chosen so that values after this time are being drawn from the limiting distribution. Values produced by the chain prior to this are discarded.

There is no definitive test to determine when convergence has occurred, but there are a number of diagnostics which can be performed to assess various other proxies. The simplest of these involves the autocorrelation function. Essentially, this involves calculating correlations for simulated values that lie set numbers of iterations apart (also known as lags). If these correlations are large for large lags, this indicates successive iterations of parameters are very similar, meaning that the chain is converging

(or mixing) slowly.

More formal techniques to measure convergence have been developed. These typically involve comparing variances between different chains with different initial starting values, or between different time intervals for the same chain. The starting value of a chain will influence the results in the short term, but not in the long term provided the chain mixes well. If different starting values produce chains which converge to the same stationary distribution, this provides evidence that the process is not dependent on the initial starting values. The Brooks-Gelman-Rubin (1998) statistic relies on the former. The method involves simulating multiple Markov chains with overdispersed initial values (i.e. providing initial values both larger and smaller than might be expected from the posterior distribution), discarding the burn-in period, then comparing the within chain variance with the pooled variance from all of the chains. When the ratio of within chain variance and pooled variance is close to one, the chain is said to have converged. WinBUGS (Lunn et al., 2000) provides a graphical representation of this test, by plotting the within variance, the pooled variance and then the ratio of the two. The first two should coincide and the ratio should approach one.

The next section will illustrate some of these concepts with reference to a simple example from the CHSNS dataset.

4.4 Illustration of Bayesian Methods

The example will involve modelling the mental health scores of males and females in the ward with the lowest average mental health score (Twyn Carno) and the ward with the highest average mental health score (St. Martins) in order to demonstrate Bayesian methods. This trimmed down dataset will be used in order to better demonstrate the effect of the prior information (if the entire dataset were employed, the large sample size would swamp out the impact of all but the most extreme forms of prior information). MCMC methods will also be illustrated.

We assume the following:

$$\begin{aligned} \mathrm{X}_i &\sim \mathrm{N}(\mu_i, 470) \\ \mu_i &= \alpha + \beta \times \mathrm{Gender}_i + \gamma \times \mathrm{Ward}_i \end{aligned} \tag{4.7}$$

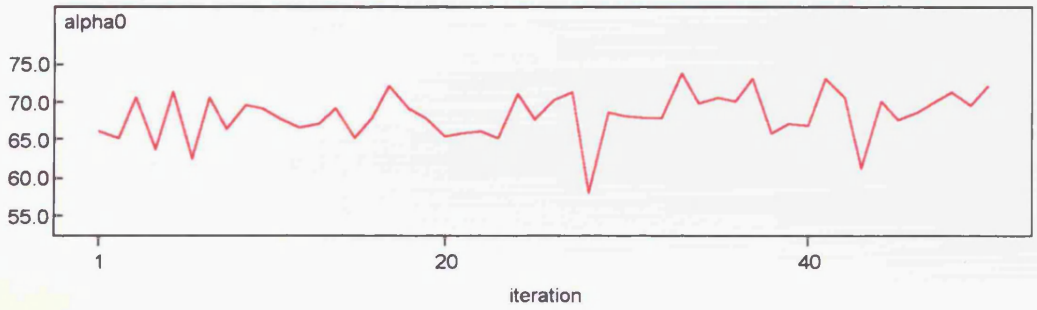
The X_i are the mental health scores. The variance of X_i is 470 since that is observed variance of the mental health scores. Gender is zero if male, and 1 if female so that β is the effect of being female compared to being male. Ward is zero for those in St Martins and 1 for those in Twyn Carno.

Functional forms for the distributions of the parameters of interest can be specified. Uninformative priors will be illustrated first. The prior distribution chosen for α

is what is known as an improper prior. It is essentially a uniform distribution without boundaries, meaning that any value is equally likely as any other. It is called improper since the integral of this distribution is infinite. This will not prevent the posterior distribution from integrating to one. Both the β and γ parameters (indicating the gender and area effects respectively) are assumed to be Normally distributed with mean zero and large variance $(1,000)$. All of these priors are vague, representing very little information about their respective parameters. Starting values for all of these parameters must be specified. For this example, the starting values for α, β and γ were all zero. Sensitivity analyses were performed with different starting values (ranging between -100 and 100) and indicated that the results were not dependent on the choice of initial values.

WinBUGS offers a few tools for graphical exploration, and these will now be presented. Each iteration produces a new parameter value (for each parameter), and these can be plotted to show how they behave across different iterations (called the history of the parameter). These history plots can provide evidence for whether the chain is sampling from the stationary distribution. If the history shows a trend, then it suggests that the chain has not converged to the stationary distribution. Even if there is no trend then it is possible that the chain is mixing slowly and is caught in one particular part of the posterior distribution. In this situation convergence would not have occurred but may appear to have from the history plot. There are other ways to test whether the chain has converged, such as the Brooks-Gelman-Rubin test statistic (1998) which will be described presently. Since MCMC estimation in WinBUGS

Figure 4.2: History of the first 50 estimated values for the intercept in equation 4.7



is computationally efficient, and this is a simple example, it is trivial to have large numbers of iterations and assess the stability of the estimates more thoroughly. Figure 4.3 displays the history of the intercept estimates for the first 1,000 iterates and again shows no upward or downward trend, suggesting the stationary distribution has been attained. The equivalent plots for the β and γ parameters are given in figures 4.4 and 4.5, and show the same pattern. Again, the process is not expected to converge to a single point, but to a distribution.

Autocorrelation plots for each of these parameters are very similar and indicate

Figure 4.3: History of the first 1,000 estimated values for the intercept in equation 4.7

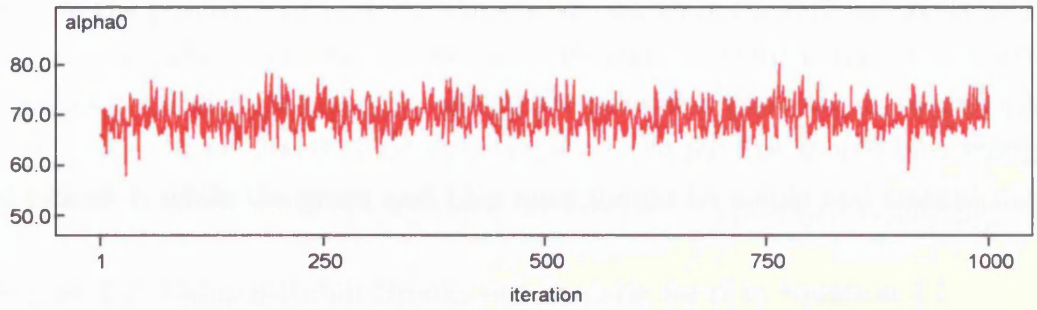


Figure 4.4: History of the first 1,000 estimated values for β in equation 4.7

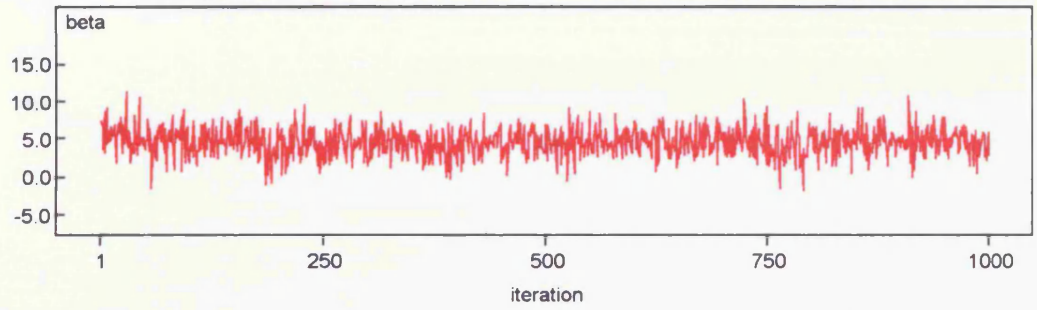
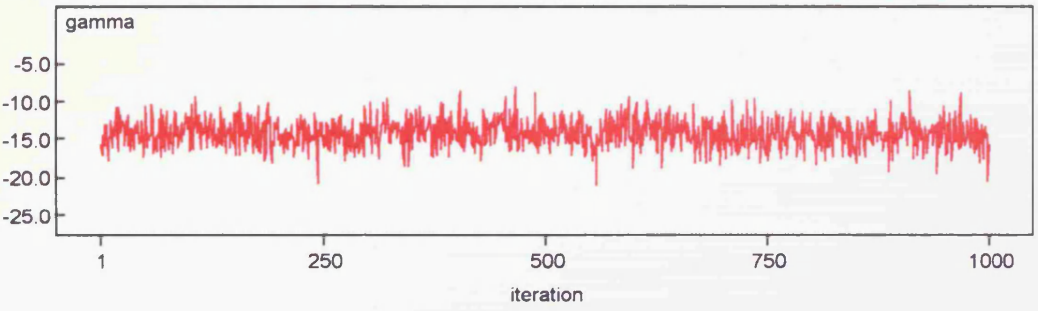
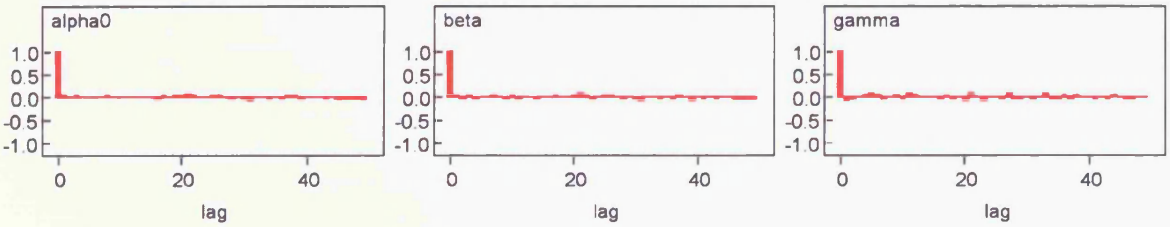


Figure 4.5: History of the first 1,000 estimated values for γ in equation 4.7



that the chain is mixing well, with the correlation between successive iterates dropping off rapidly with increasing lags, as shown in figure 4.6.

Figure 4.6: Autocorrelation plots for α, β and γ



In order to employ the Brooks-Gelman-Rubin test statistic (1998), at least two chains need to be run and then compared. If this is done, using overdispersed initial parameters (0 and 100), the between chain and within chain variability can be graph-

ically compared, as shown in figures 4.7-4.9. For these plots "the width of the central 80% interval of the pooled runs is green, the average width of the 80% intervals within the individual runs is blue, and their ratio R (= pooled / within) is red - for plotting purposes the pooled and within interval widths are normalised to have an overall maximum of one" (Lunn et al., 2000). Under convergence, the red line should be a straight horizontal line at 1, while the green and blue lines should be stable and coincidental.

Figure 4.7: Gelman-Rubin-Brooks test statistic for α in equation 4.7

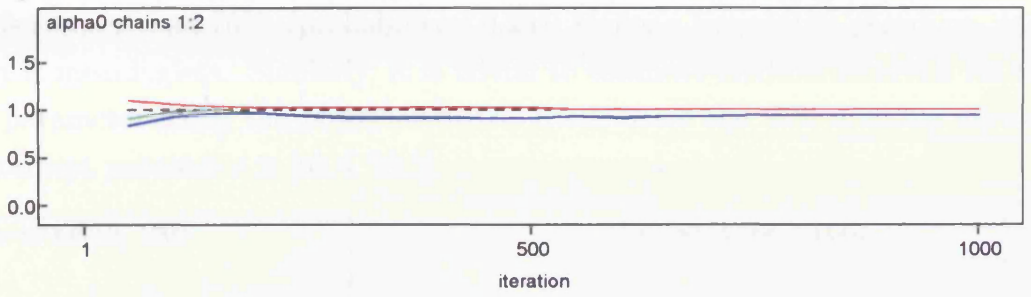


Figure 4.8: Gelman-Rubin-Brooks test statistic for β in equation 4.7

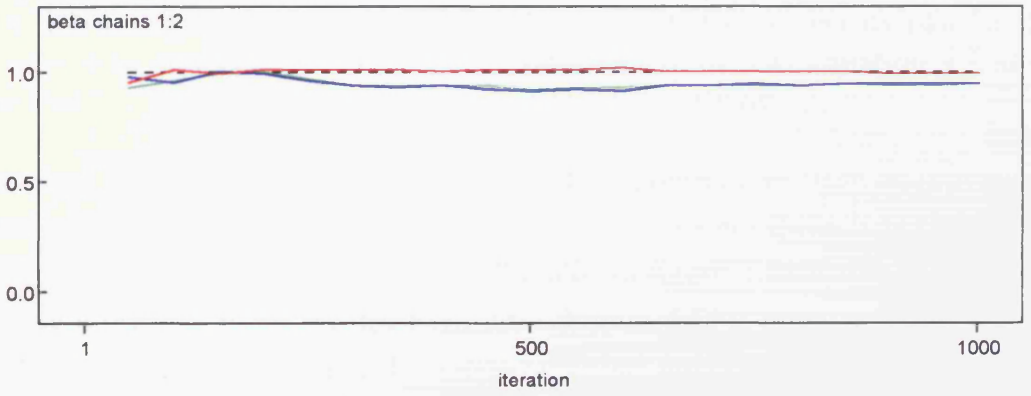
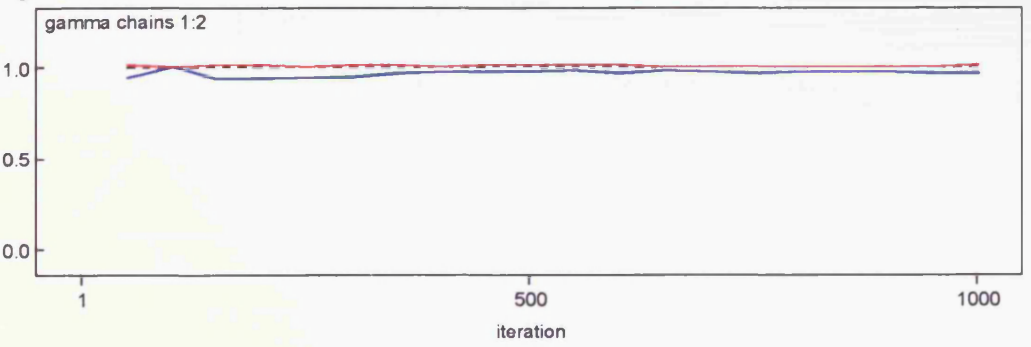


Figure 4.9: Gelman-Rubin-Brooks test statistic for γ in equation 4.7



Since the chain appears to converge extremely quickly, there may be less need for a burn-in period. Since this example is so small however, the chain runs very quickly

and so a burn-in period of 250 iterates is chosen. A further 250 are simulated so that the sample size for the posterior distributions of all parameters is 1,000. Figure 4.10 shows the kernel density plot of the posterior distribution for the intercept term (α) from equation 4.7. This figure shows that the distribution of the intercept term is centred just below 70. Being based on only 1,000 samples, the density is not very smooth. If the chain is increased to say, 100,000 in length the resulting kernel density plot is much smoother, as depicted in figure 4.11. Using the posterior distribution, it is straightforward to answer questions such as "what is the probability that the intercept is greater than 70?" (probability= 0.24). Under a frequentist paradigm, such a question is meaningless. Similarly, it is trivial to estimate credible intervals for the intercept parameter using the posterior distribution. Here the 95% credible interval for the intercept parameter is $[62.4, 73.5]$.

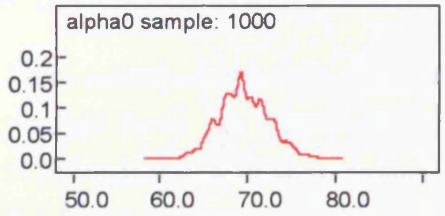


Figure 4.10: Kernel density plot for the intercept term α in equation 4.7 after 1,000 iterations

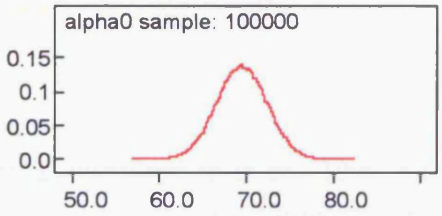


Figure 4.11: Kernel density plot for the intercept term α in equation 4.7 after 100,000 iterations

Similar plots for the posterior distribution of the β parameter from equation 4.7, are presented in figures 4.12 and 4.13. Here the distribution of the β parameter is centred just below 5. Again, the kernel density plot in figure 4.12 is not particularly smooth, being a product of only 1,000 iterations. Figure 4.13 is much smoother being based on 100 times as many points. This indicates that men in these two areas score five points higher than women on average on the MHI-5. Higher scores mean better mental health. Again, having access to the posterior distribution allows for probability statements to be made about where given parameters are likely to lie. Here the 95% credible interval for the β parameter is $[2.1, 9.0]$.

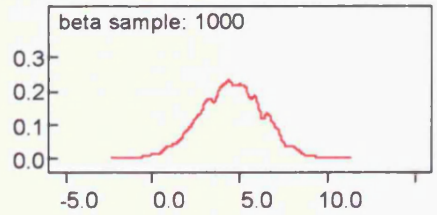


Figure 4.12: Kernel density plot for the regression coefficient β in equation 4.7 after 1,000 iterations

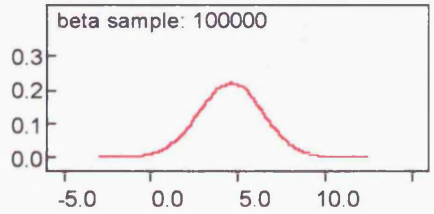


Figure 4.13: Kernel density plot for the regression coefficient β in equation 4.7 after 100,000 iterations

Finally, kernel density plots for the γ parameter are given in figures 4.14 and 4.15.

The area-level indicator associated with the γ parameter represents the observed effect of living in Twyn Carno relative to St Martin's. This explains why the γ parameter's posterior distribution is located around -15, as the mental health scores in Tywn Carno are lower than in St Martin's. Again, the kernel density in figure 4.15 is much smoother than 4.14 due to the much larger sample size.

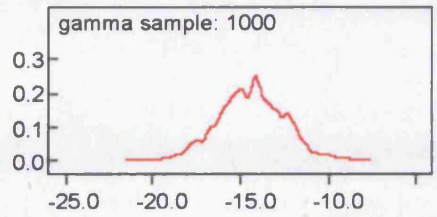


Figure 4.14: Kernel density plot for the regression coefficient γ in equation 4.7 after 1,000 iterations

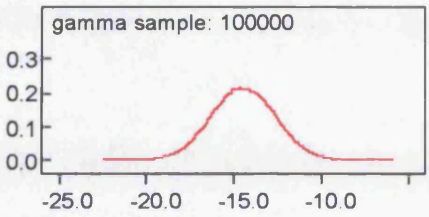


Figure 4.15: Kernel density plot for the regression coefficient γ in equation 4.7 after 100,000 iterations

More informative priors are now incorporated into the model to assess what impact they might have. As mentioned earlier, it is well documented that men report better levels of mental health than women (Weich et al., 1998; Bebbington et al., 1998; Emslie et al., 2002). This information could be included in the model, by building into the prior information that the β parameter will be positive. The approach illustrated here is to assume that β belongs to a uniform distribution ranging between 0 and 20. The model is now given in equation 4.8. Adjusting the model to include this information, similar plots to figures 4.12 and 4.13 can be obtained.

$$\begin{aligned} \mu_i &= \alpha + \beta \times \text{Male}_i + \gamma \times \text{Twyn Carno}_i \\ \text{Mental Health}_i &\sim \text{N}(\mu_i, 470) \\ \beta &\sim \text{U}(0, 20) \\ \gamma &\sim \text{N}(0, 1000) \end{aligned} \tag{4.8}$$

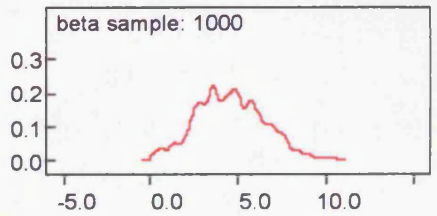


Figure 4.16: Kernel Density plot for the regression coefficient β in equation 4.8 after 1,000 iterations

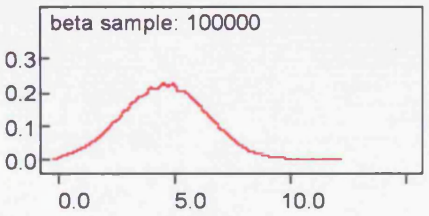


Figure 4.17: Kernel Density plot for the regression coefficient β in equation 4.8 after 100,000 iterations

Figure 4.16 shows the kernel density plot for β after 1,000 iterations, while figure 4.17 shows the same plot after 100,000. Notice that the assumption that this parameter is positive has truncated the distribution at zero. Again, the increased sample size

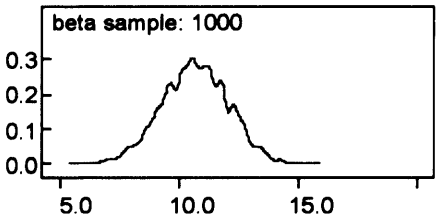


Figure 4.18: Kernel Density plot for the regression coefficient β in equation 4.9 after 1,000 iterations

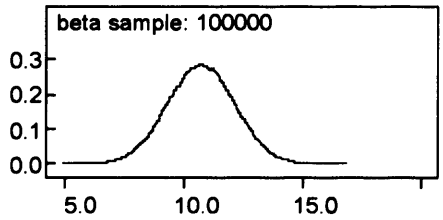


Figure 4.19: Kernel Density plot for the regression coefficient β in equation 4.9 after 100,000 iterations

smooths the posterior distribution. Including this more informative information has not changed the posterior distribution for this parameter a great deal, since the prior was still quite vague. The mean of β is 5.5, with a standard deviation of 1.8 (95% credible interval [2.1,9.0]). This is exactly the same confidence interval was estimated for the β parameter previously. Having specified more information did not alter this confidence interval. The extra information specified will prevent this parameter and confidence limits from being negative.

In order to illustrate the effect poorly chosen priors can have on the results of a Bayesian analysis, it is now assumed that prior information has led to a choice of N(20,5) for the prior distribution of β. This information is precise and incorrect. So now the model is as in equation 4.9

$$\begin{aligned} \mu_i = \alpha + \beta \times \text{Female}_i + \gamma \times \text{Twyn Carno}_i \\ \text{Mental Health}_i \sim \text{N}(\mu_i, 470) \\ \beta \sim \text{N}(20, 5) \\ \gamma \sim \text{N}(0, 1000) \end{aligned} \tag{4.9}$$

The posterior distributions for β from this model after 1,000 and 100,000 iterations are plotted in figures 4.18 and 4.19 respectively. The data indicates that the difference should be 5, but the prior indicates that it should be 20. The posterior distribution has incorporated both of these sources of information and has a mean of 10.7 and a standard deviation of 1.4 (95% credible interval for the mean [8.0,13.5]). Including informative priors therefore should be treated with caution since including the wrong information can have disastrous consequences. Informative priors should only be included if there is good reason to do so. These techniques are now applied to the full CHSNS dataset.

4.5 Spatial variation in cases of common mental disorder in Caerphilly county borough

This section will provide some motivation for the use of hierarchical modelling by examining the geographical variation in cases of common mental disorder throughout the study area. Table 4.1 gives some summary statistics for the underlying mental health score.

Table 4.1: Basic Summary of Mental Health Score

	Males	Females	Overall
Mean	71.9	67.4	69.4
Variance	20.8	22.2	21.7
N	4770	5883	10653

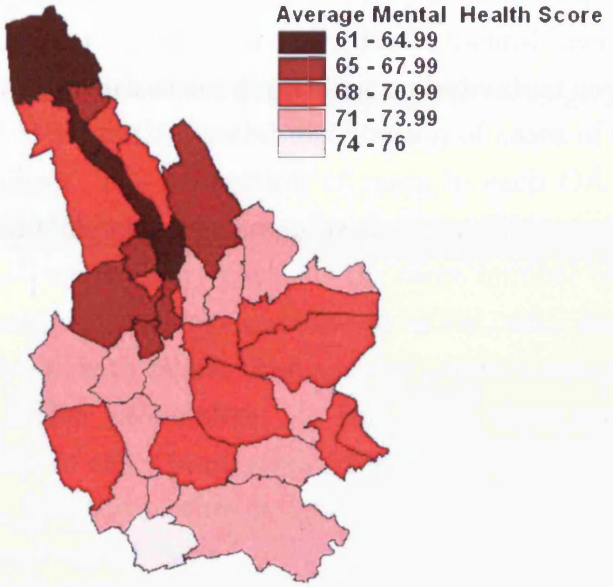
Males scored higher on average than females (indicating better mental health) and this result was significant (Mann-Whitney test, p-value$<$ 0.001). Table 4.2 shows average mental health, grouped by 1991 census electoral wards.

Table 4.2: Summary of Mental Health Score by Ward

Ward name	Mean	N	Std. Dev	Ward name	Mean	N	Std. Dev
Aber Valley	68.05	304	22.96	Moriah	67.44	263	22.14
Aberbargoed	63.24	238	23.92	Nelson	72.42	324	19.43
Abercarn	68.16	327	21.37	New Tredegar	67.65	272	20.86
Abertysswg	68.69	237	21.39	Newbridge	70.29	328	21.49
Argoed	66.20	225	22.34	Pengam	67.41	286	23.19
Bargoed	67.39	287	22.31	Penmaen	70.88	325	21.07
Bedwas and Trethomas	69.90	302	22.52	Penyrheol	71.16	368	21.32
Blackwood	72.21	345	21.12	Pontllanfraith	70.30	299	20.24
Cefn Fforest	69.37	319	20.13	Pontlottyn	63.55	312	23.88
Crosskeys	71.08	309	20.31	Risca East	69.58	337	20.98
Crumlin	71.45	292	22.18	Risca West	70.44	349	21.17
Darran Valley	67.65	281	21.80	St. Cattwg	66.41	299	21.91
Gilfach	69.73	231	22.00	St. James	73.15	328	19.83
Hengoed	65.92	277	23.11	St. Martins	76.12	348	18.66
Llanbradach	72.40	307	21.19	Tir-Phil	64.08	210	23.85
Machen	72.34	297	20.15	Twyn Carno	61.39	225	24.21
Maes Y Cwmmer	72.32	252	19.89	Ynysddu	69.70	317	21.95
Morgan Jones	70.90	330	20.44	Ystrad Mynach	70.95	303	21.41

Even from the table it is clear that there is considerable spatial variation in mental health scores within the borough. The ward with lowest average mental health score is

Figure 4.20: Map showing the spatial variation in mental health at Ward level



Twyn Carno at 61.39, while St. Martins has the highest average mental health score with 76.12. Twyn Carno is the northernmost ward, while St. Martins is one of the most southerly wards in the borough. A Kruskal-Wallis test confirms that (before controlling for compositional variables) there is significant variation in mental health between wards (p-value < 0.001). The differences between wards can be quite dramatic. Aberbargoed and Blackwood are adjacent wards which have an average difference in mental health scores of nearly nine points. Figure 4.20 maps the information in table 4.2 (figure 2.3 illustrates the location of the wards). The north/south difference is clear from this map, with average mental health scores being lower in the north of the borough. Particularly noticeable is the elongated ward at the north of the borough with mental health scores in the range 61-64. This picks out the settlements along the floor of the Upper Rhymney Valley. Clearly, this ecological analysis proves nothing about the effect of place of residence on mental health, as it is possible that the differences seen here are entirely attributable to the composition of these wards.

In order to examine the spatial variation at a smaller level than wards, which may be too large and heterogenous to adequately represent area of residence, 2001 census Output Areas (OAs) can be employed. Since cases of common mental disorders are of interest the original mental health scale is dichotomised to produce a binary variable indicating whether individuals are a case of CMD or not. The cutpoint is chosen is the cutpoint of 60 that minimises the misclassification rate from chapter 3. This is the same cutpoint used in a previous analysis of the data (Fone, 2005). There 60 was

chosen to give a proportion of cases of common mental disorder closest to 32% (a cut-point of 60).This was chosen since the 1996 Health in Wales survey (Kingdon et al., 1998) indicated that 32.4% of the Caerphilly county borough population satisfied the General Health Questionnaire (GHQ-12) (Goldberg & Williams, 1988) case definition for common mental disorder. Then each individual's mental health score is modelled as a Bernoulli distribution, with mean depending on individual covariates such as gender and age. Figure 4.21 plots the spatial distribution of cases of CMD using Output Areas (559 in Caerphilly). The proportion of cases in each OA is placed in one of seven categories. The middle category contains the overall borough proportion of 32% and is coloured white. Each category contains the same number of OAs. Areas with a lower proportion of cases than average are shaded in red, with darker hues indicating smaller proportions. Areas with higher than average proportions are shaded blue, with darker hues denoting higher proportions. In this figure there appears to be a trend of increasing proportions of cases (and so, decreasing mental heath scores) toward the north of the borough, but the picture is more complicated, with some areas in the north having low proportions of cases.

There is lots of variation in the observed proportions of cases of CMD in figure 4.21. Some of this variation is undoubtedly due to small sample sizes per OA (sample sizes for some of the OAs are as low as 2). In order to remove this random variation, while leaving true spatial patterns, smoothing can be performed.

One method of smoothing employs a model developed by Besag, York and Mollié (1991). This model incorporates information from adjacent areas in order to provide better estimates of the true proportion of cases when the sample size for an OA is small. These proportions based on small sample sizes will be smoothed toward the global average proportion of 0.32. Equation 4.10 shows the model assumed for the data, with O_i representing the observed number of cases for the i^{th} OA. A nice feature of the Besag, York and Mollieé model is that other explanatory variables can be included in the model in equation 4.10. This can be extremely useful in the field of epidemiology where basic socioeconomic variables need to be controlled for, making this model frequently suitable for public health research (Lawson et al., 1999). The H and the S components can be thought of as surrogates for unknown or unobserved variables. The S component represents variation that is due to structured variables, i.e. variables that, if observed, would display substantial spatial structure. These would be variables that would have similar values in adjacent OAs. The S component is distributed as a conditional autoregressive normal, where random effects for an area are influenced by the areas adjacent to it. Weights are attached to each pair of areas in the dataset, and while any weights can be used, here the weight is 1 if the two areas in question share a boundary and 0 otherwise. The conditional distribution of each S_i is Normal with mean equal to the mean of all the adjacent S_i's, and variance inversely

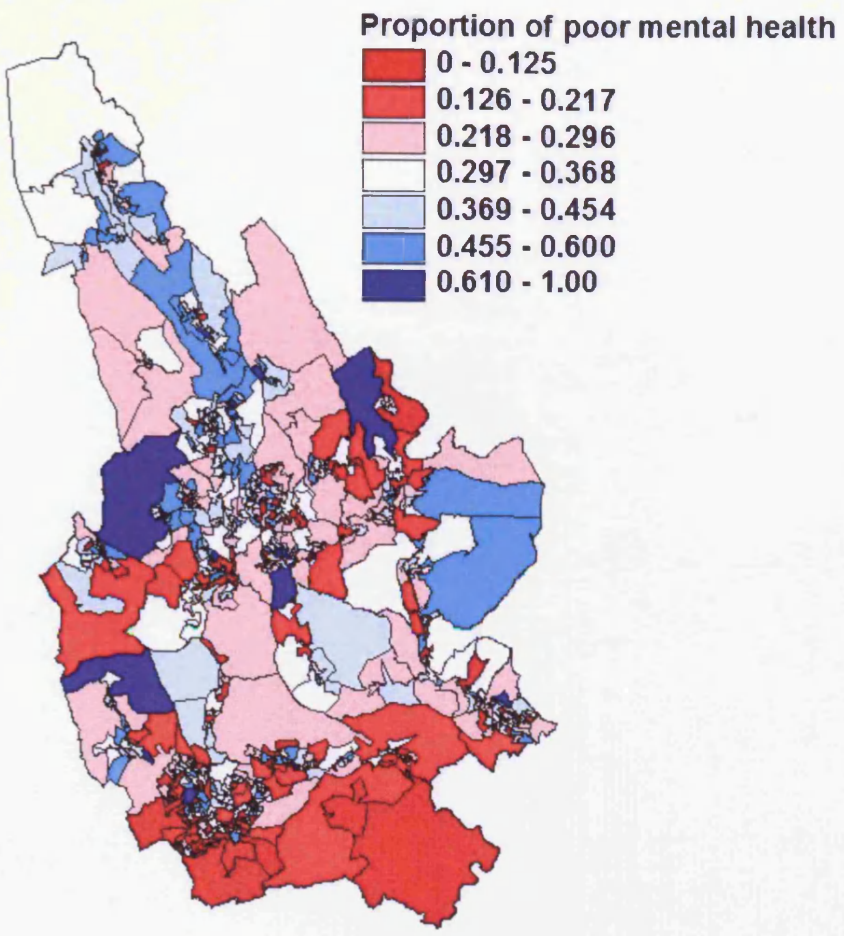


Figure 4.21: Map showing the unsmoothed proportions of cases at OA level

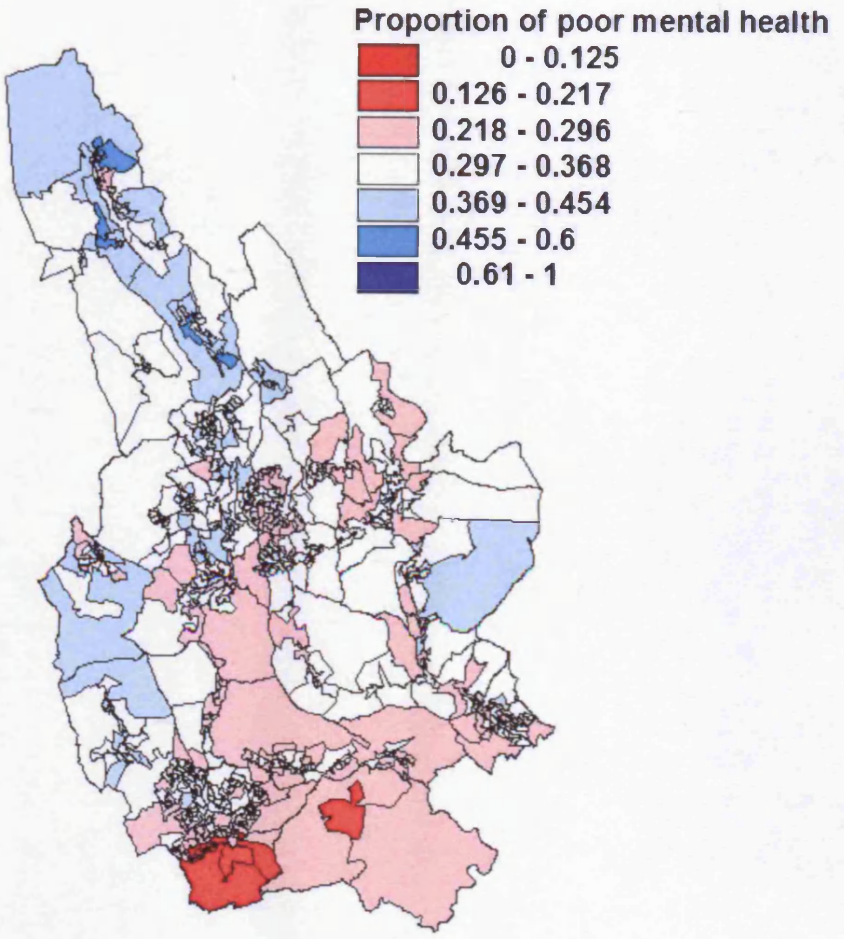


Figure 4.22: Map showing the smoothed proportions of cases at OA level

proportional to the number of observations in the adjacent areas. An adjacency matrix informs the model about which areas are adjacent.

The H component is the heterogeneity component and represents unstructured variables, i.e. variables that if observed would be unrelated with spatial position.

$$\begin{aligned} logit(\mu_i) &= \omega + H_i + S_i \\ O_i &\sim \text{Binomial}(\mu_i, n_i) \\ \omega &\sim \text{N}(0, 1000) \\ H &\sim \text{N}(0, \sigma_h^2) \\ S &\sim \text{CAR.N}(0, \sigma_s^2) \end{aligned} \tag{4.10}$$

Priors must be specified for ω, σ_s^2 and σ_h^2. Since we have no information about ω a N(0,1000) distribution was chosen. Another possible alternative would be to use a uniform distribution with a wide range. It is often difficult to rigorously justify the specific form of an uninformative prior for a paramter, since often the reason for using an uninformative prior is that there is little information available about the parameter that could be used to justify any choice. We know a little more about the σ_{h^2} and σ_{s^2} parameters. Being variances they are strictly non-negative. Uninformative priors for such parameters are typically Gamma or log-normal distributions. Here we choose the former and model both as Gamma distributions with mean 1 and variance 1,000. Again the justification of this specific choice of distribution is not strong, however, since they have large variances the impact of which distribution is chosen should not be large (in fact sensitivity analyses were performed to confirm that the choice of the Gamma distribution here does not radically alter the results compared with choosing a log-normal). The specification of the priors used in this analysis are shown in equation.4.11.

$$\begin{aligned} \omega &\sim \text{Normal}(0, 1000) \\ \sigma_h^2 &\sim \text{Gamma}(0.001, 0.001) \\ \sigma_s^2 &\sim \text{Gamma}(0.001, 0.001) \end{aligned} \tag{4.11}$$

The model was run for 11,000 iterations, with the first thousand discarded as a burn-in period. A full history for each parameter (including the 1,000 long burn-in period) is provided in figures 4.23-4.27. These history plots indicate that the chains converge quickly.

Densities for these parameters are provided in figures 4.28-4.32.

Initial values for all of these parameters were set to 1, however sensitivity analyses were performed with different initial values between 0 and 100. The results of the model did not change.

Figure 4.23: History plot of the H_1 for the BYM model

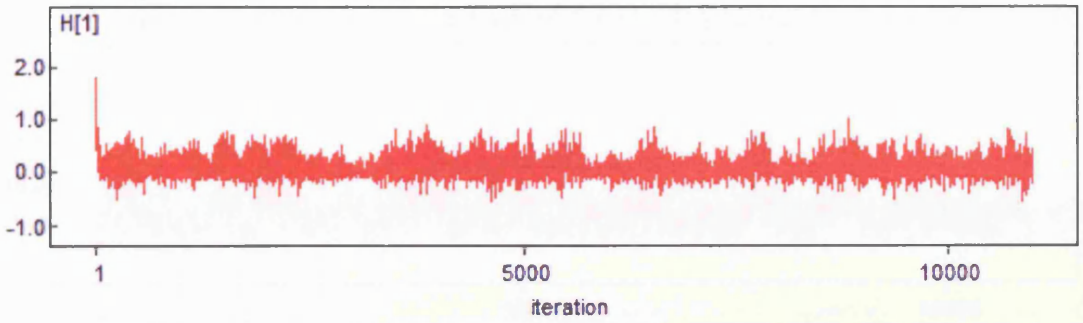


Figure 4.24: History plot of the S_1 for the BYM model

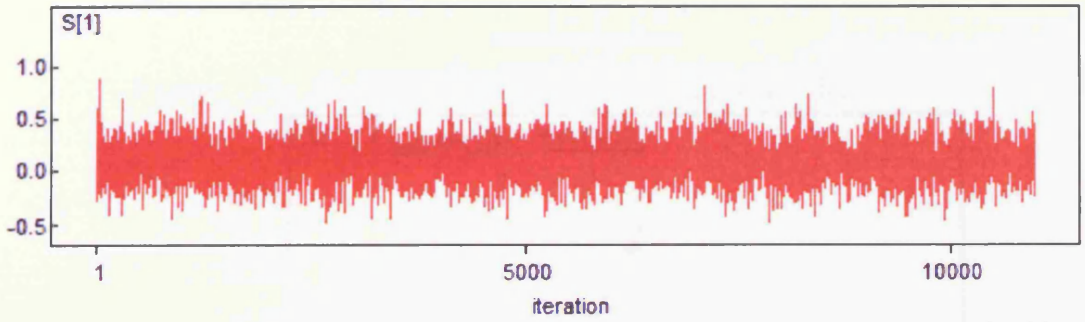


Figure 4.25: History plot of ω (the intercept term) for the BYM model

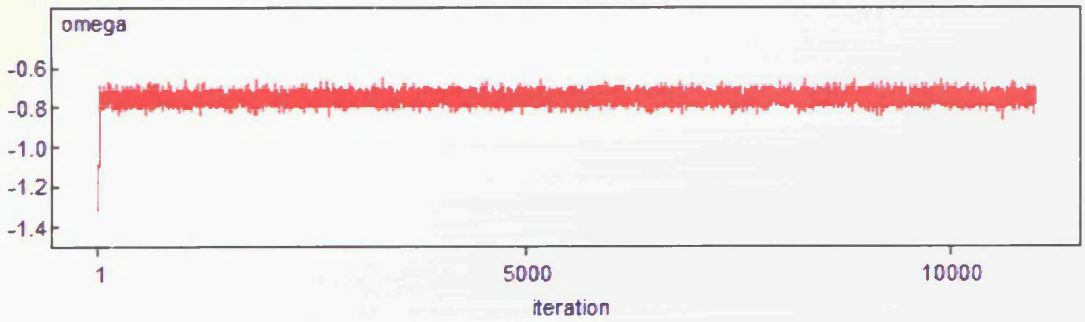


Figure 4.26: History plot of σ_h^2 for the BYM model

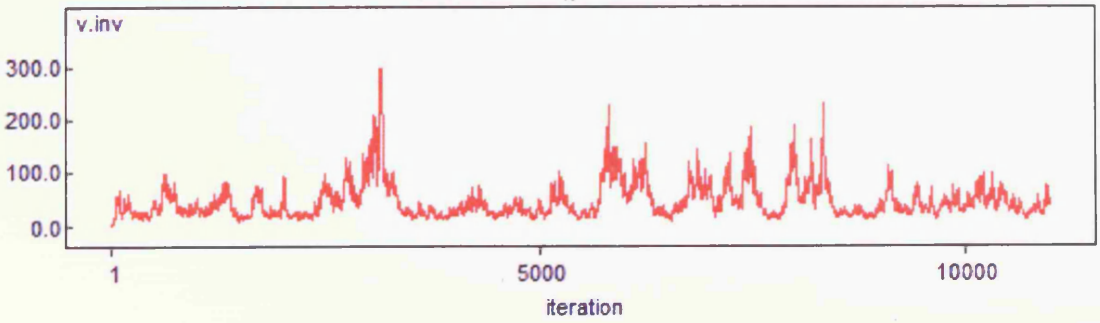


Figure 4.27: History plot of σ_s^2 for the BYM model

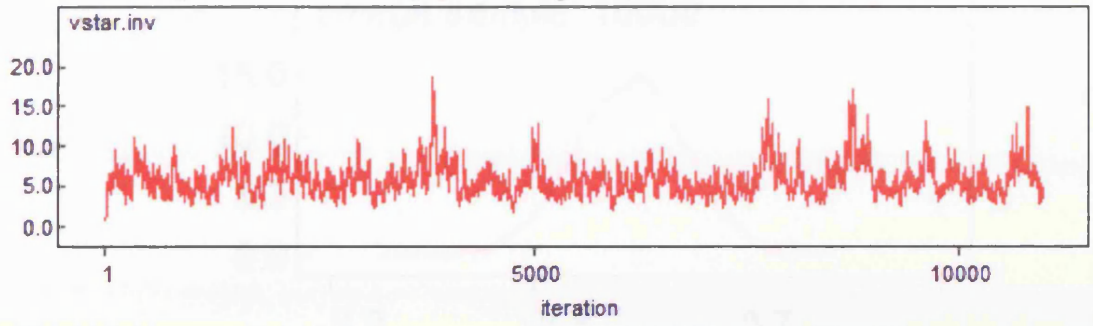


Figure 4.28: Density plot of H_1 for the BYM model

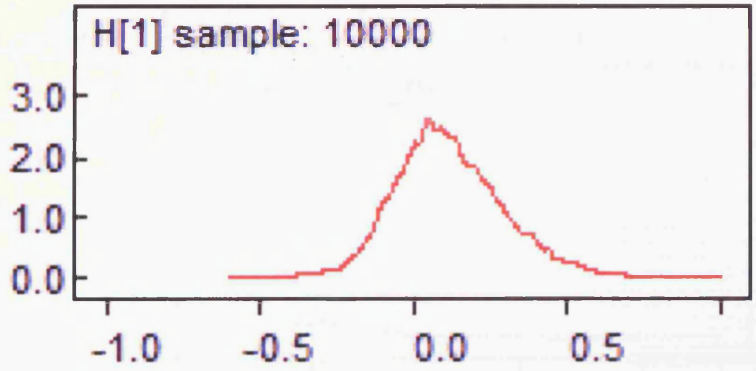


Figure 4.29: Density plot of S_1 for the BYM model

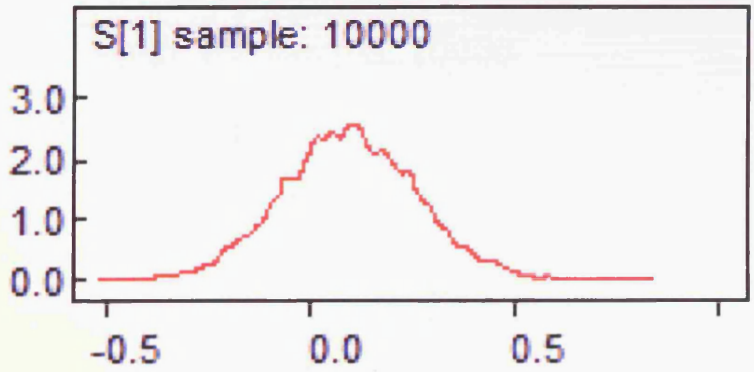


Figure 4.30: Density ω (the intercept term) for the BYM model

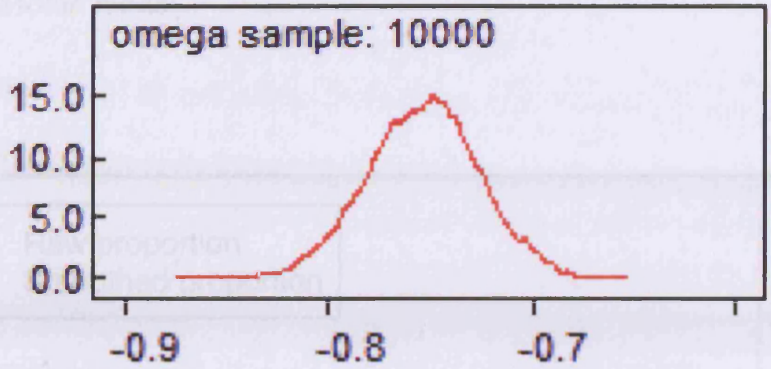


Figure 4.31: Density σ_h^2 for the BYM model

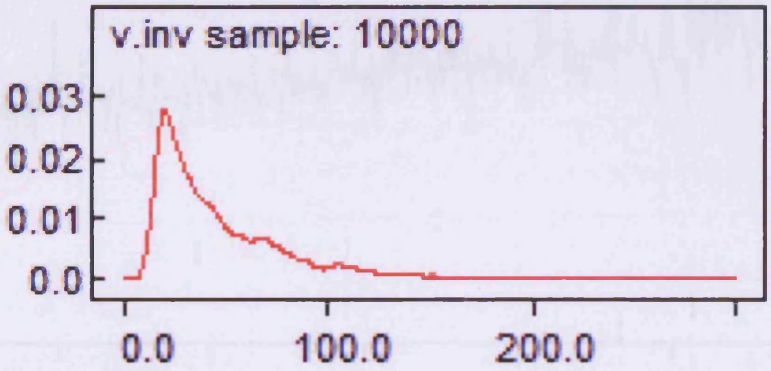


Figure 4.32: Density σ_s^2 for the BYM model

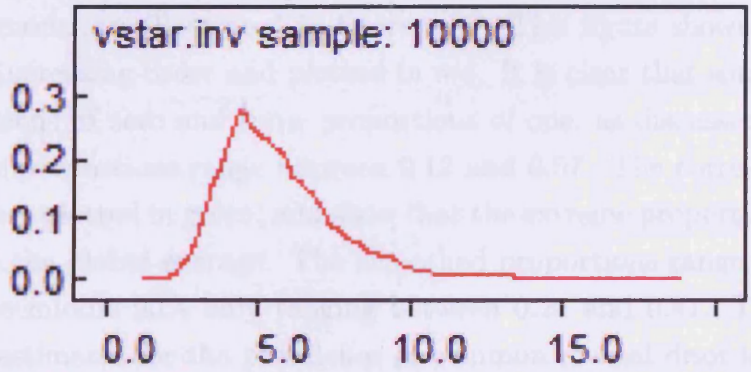
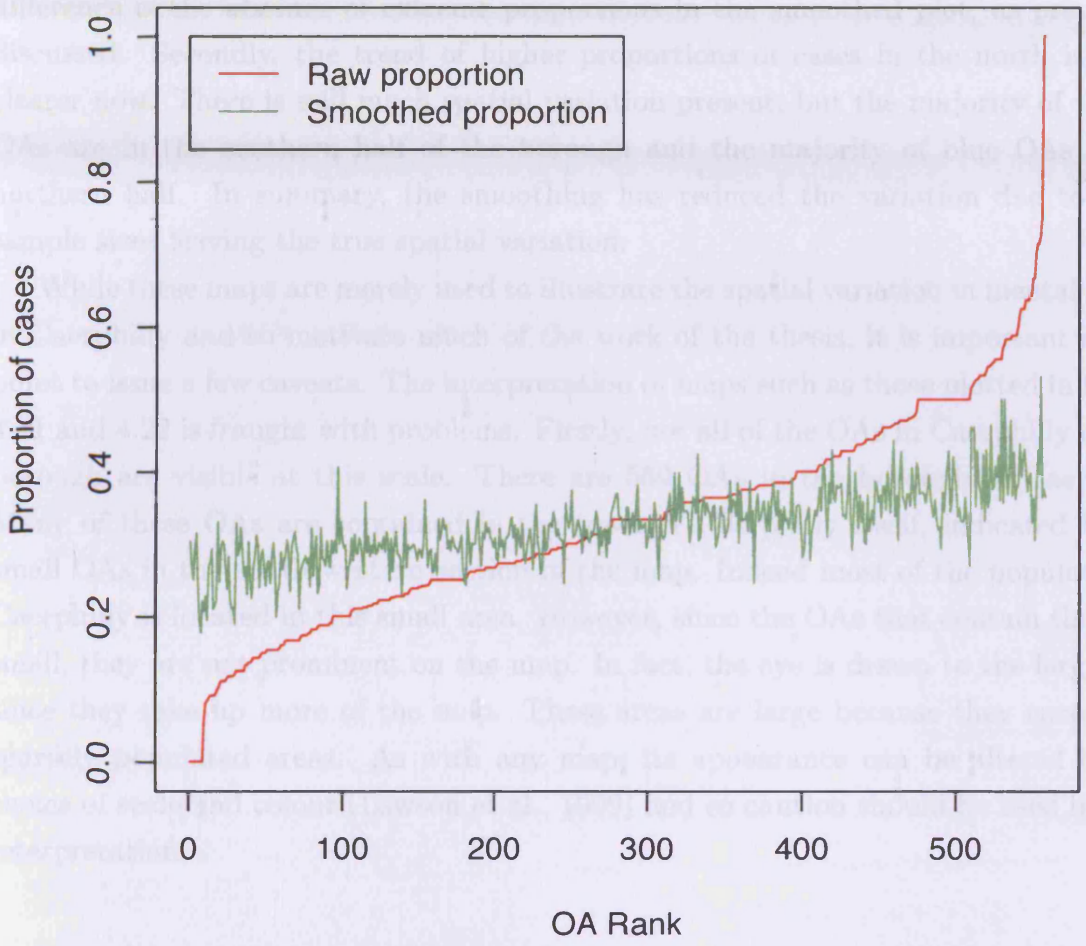


Figure 4.33: Ranked raw proportions and their corresponding smoothed values from the Besag, York and Mollié model



The results of the model are illustrated in figure 4.33. This figure shows the raw proportions ranked in increasing order and plotted in red. It is clear that some of the OAs have case proportions of zero and some proportions of one, as discussed earlier. The middle 90% of the proportions range between 0.12 and 0.57. The corresponding smoothed proportions are plotted in green, and show that the extreme proportions have been smoothed toward the global average. The smoothed proportions range between 0.18 and 0.53, with the middle 90% only ranging between 0.24 and 0.41. These are much more believable estimates for the prevalence of common mental disorders. The proportions have not merely been scaled toward the global average however, with some of the smoothed proportions being farther from the global average than their unsmoothed counterparts. This is indicated in figure 4.33 by the fact that the green

line is not merely a scaled down version of the red line, but instead displays random variation. This is due to the smoothing being calculated based on the sample size of the OA in question as well as information from adjacent OAs.

Figures 4.21 and 4.22 map the raw and smoothed proportions, respectively. They are mapped using the same colour scheme for ease of comparison. The first striking difference is the absence of extreme proportions in the smoothed plot, as previously discussed. Secondly, the trend of higher proportions of cases in the north is much clearer now. There is still much spatial variation present, but the majority of the red OAs are in the southern half of the borough and the majority of blue OAs in the northern half. In summary, the smoothing has reduced the variation due to small sample sizes leaving the true spatial variation.

While these maps are merely used to illustrate the spatial variation in mental health in Caerphilly and so motivate much of the work of the thesis, it is important at this point to issue a few caveats. The interpretation of maps such as those plotted in figures 4.21 and 4.22 is fraught with problems. Firstly, not all of the OAs in Caerphilly county borough are visible at this scale. There are 559 OAs in the borough of Caerphilly. Many of these OAs are contained in the town of Caerphilly itself, indicated by the small OAs in the south western section of the map. Indeed most of the population of Caerphilly is located in this small area. However, since the OAs that contain them are small, they are not prominent on the map. In fact, the eye is drawn to the large OAs since they take up more of the map. These areas are large because they encompass sparsely populated areas. As with any map, its appearance can be altered by the choice of scale and colour (Lawson et al., 1999) and so caution should be used in their interpretation.

4.6 Conclusion

This chapter has introduced and explained Bayesian inference. Using Bayesian techniques the spatial variation of mental health present in Caerphilly was investigated. Bayesian smoothing was used to demonstrate that this spatial pattern is not an artefact of low sample sizes, and represents a genuine question of interest. The rest of this thesis will investigate methodological issues surrounding the investigation of this spatial variation. A crucial tool in any such investigation is hierarchical modelling, which will be introduced in chapter 5.

Chapter 5

Hierarchical Modelling

5.1 Hierarchical models

This section will introduce and explain one of the statistical tools employed in analysing the Caerphilly Health and Social Needs dataset, namely hierarchical modelling. Hierarchical modelling in its simplest form is an extension of the simple linear regression model. The simple linear model can be expressed in the form given in equation 5.1

$$y_i = \alpha + \beta x_i + \epsilon_i \tag{5.1}$$

Here y_i is the response for the i^{th} individual, while α represents the overall average response for the entire dataset. The x_i term represents the covariate for the i^{th} individual, β represents the effect of a unit increase in the x covariate on the response. The ϵ_i term is the random component of the model. It is called the error term, and it has mean zero and variance σ_ϵ^2. There are a number of assumptions implicit in this simple model, one of which is that the responses are independent of one another. This means that the response given by the i^{th} individual is not affected by, or correlated with, the response given by any other individual in the dataset. When this assumption is violated, the standard errors for the βs under ordinary least squares or maximum likelihood regression are underestimated (Goldstein, 2003). This results in an increased chance of type I error, where a true null hypothesis is incorrectly rejected. This means significant associations may be reported, when in fact these associations are ascribable to chance.

There are many practical situations where the assumption of independence is violated. The classic example is in the field of educational research. Here a typical hypothesis might be that schools which teach reading using the "look-say" (also known as the "whole word" or "sight method") produce children with higher reading levels than schools which use the "phonetic method" method. Here reading level would be the outcome of interest with teaching type as an explanatory covariate. In this scenario an

ordinary least squares (OLS) linear model would not account for the fact that children from the same class in the same school are likely to be more similar to one another than they are to children in different classes, or different schools. Hence, observations from children in a given school will be correlated. This becomes even more obvious if each student's reading score is considered to provide a measure of the same thing (i.e. the efficacy of the reading method to which they are exposed). Using linear regression this might be modelled as in equation 5.2

$$y_{ij} = \beta_0 x_0 + \beta_1 x_{1ij} + \epsilon_{ij} \tag{5.2}$$

Here y_{ij} refers to the response from the i^{th} student in the j^{th} school. In this model there is a single intercept denoted by β_0. The x_0 term is present to keep the notation consistent and takes the value 1 for all individuals. The β_1 term represents the coefficient of the x covariate (this model has only one covariate). Hierarchical modelling treats each school as coming from a population of possible schools and allows each school to deviate from the overall average by a random amount. So, instead of investigating hypotheses related to these specific schools, a hierarchical model allows inference to be made about the population of schools from which were sampled. This is analogous to the case where individuals are sampled from a population, not to examine the individuals themselves, but to make statements about the population from which they were drawn. To do this a random error term τ_{0j} is included, denoting the deviation of school j from the overall average, β_0. It has expected value zero and variance σ_τ^2. Including this term in equation 5.2 gives equation 5.3

$$y_{ij} = \beta_0 x_0 + \beta_1 x_{1ij} + \tau_{0j} + \epsilon_{ij} \tag{5.3}$$

This explicit acknowledgement of the hierarchical clustering in the data allows the correct standard errors to be calculated for the effects using ordinary least squares regression. Equation 5.3 is called a variance components model, since it partitions the variation in the data into either the individual or school level. The assumptions neccessary for the estimation of parameters in this model are that the y_{ij}s are normally distributed, and that the τ_{0j}s and ϵ_{ij}s are normally distributed about zero and independent of each other. This model can be extended again if the β_1 terms are treated as random variables. This would be done if it were expected that the slope of the relationship between reading ability and time spent exposed to a given teaching method varied from school to school. Any covariate can be allowed to vary randomly at any level.

$$\beta_{1j} = \beta_1 + \upsilon_{1j} \tag{5.4}$$

Again the expected value of v_{1j} is zero and its variance is denoted by $\sigma_{v_{1j}^2}$. When this is substituted into equation 5.3 it results in equation 5.5.

$$y_{ij} = \beta_{0j}x_0 + \beta_1 x_{1ij} + v_{1j}x_{1ij} + \tau_{0j} + \epsilon_{ij} \tag{5.5}$$

One of the main reasons to model random slopes is to improve model fit. If the underlying cause of the slopes being different between areas is not known, fitting random slopes can account for this variation. In general, when there is the possibility that relationships between variables vary in different areas but there is no specific hypothesis (or information) regarding the mechanism by which this might happen, random slopes are employed. If there a specific mechanism of interest, say for instance, that the relationship between area deprivation and mental health depends on whether the area is rural or urban, then a cross-level interaction (described below) is used.

The advantages of hierarchical modelling do not end there however. To extend the earlier example of investigating the effect of "look-see" and "phonetic" methods of teaching reading on student reading levels, it would be important to control for individual and school level variables. Perhaps one school has an attached Montessori school, which most children attend before enrolling in the primary school. This may mean that these students have a head start on students who do not attend a pre-school. In this scenario it may be neccessary to measure children's reading level before they are exposed to either teaching method, to get a baseline reading level. This would be an individual-level variable. School level variables may also be available, such as average teacher:pupil ratio for a school. If this information were to be included in a standard linear model they would both be treated the same way. In hierarchical modelling however, each variable can be assigned to the correct level and treated accordingly. So pupil's reading level would be modelled at individual level, while the teacher:pupil ratio could be included at school level. The contribution each level makes in explaining the variation in the data can then be determined.

Cross-level interactions can also be included in the model. As the name suggests cross-level interactions are ones which combine variables from different levels. If for instance, there was reason to believe that there might be a gender effect in the above example a cross-level interaction could be included to assess that. The hypothesis could be that boys learn reading faster if they are taught using the "look-say" method, while girls learn faster using the "phonetic" method, with the gender of students being the individual-level variable and the teaching method being a class- or teacher-level variable.

Another interesting parameter to examine in hierarchical modelling is the Intra-class Correlation Coefficient (ICC). The ICC gives an estimate of the proportion of variation attributable to each level. This allows the relative importance of each level

to be assessed. For instance, if the question of interest were voting preference in a general election an important context to consider might be household, since residents of the same household are likely to share similar political views. In this situation, the surrounding constituency may also be an important context, since it is possible that in a given area there is a proven, trusted and popular politician who takes the majority of the votes. However, if this study were a national one, then the influence of country (England, Scotland, Wales or Northern Ireland) may not be important at all. The ICC gives a quantitative assessment of the relative importance of each context, and the formula for a two-level model (i.e. equation 5.3) is shown in equation 5.6

$$\rho = \frac{\sigma^2_{\mu 0}}{\sigma^2_{\mu 0} + \sigma^2_{\epsilon}} \tag{5.6}$$

Here $\sigma^2_{\mu 0}$ is the higher level variance, while σ^2_{ϵ} is the lower, or individual-level variance. It has been demonstrated that what may be seen as a modest ICC can be consistent with large area effect sizes. This sentiment is echoed by Merlo (2003) who says "*We need to understand that large odds ratios and a low intraclass correlation are not counterintuitive facts, but they give different and complementary information*". Small ICCs can also dramatically affect the design effect for a hierarchical study. When there is a small ICC this means that level-one units nested in the same level-two unit are only slightly more similar than level-one units in different level-two units, i.e. there is only a small amount of clustering. This clustering reduces the amount of information that the sample carries. So, if say 50 students are sampled from 10 schools, the sample size is 500 individuals. Since they are clustered however, the effective sample size is less than that. The design effect is the number by which the number of lower level units must be multiplied by, in order to achieve the same statistical power of a study with 500 independent individuals. It is given in equation 5.7.

$$D = 1 + (\bar{n} - 1)\rho \tag{5.7}$$

Here D is the design effect and $\bar{n}$ is the average number of individuals per higher level unit. So if the data has a relatively modest ICC of 0.005, and 40 individuals are sampled from 10 higher units, it produces a design effect of 1.195. This means that the number of lower level units sampled would need to be increased by nearly 20% in order to have the same effective sample size as the independent case. So to achieve the statistical power provided by 40 independently distributed individuals, it would be neccessary to sample 50 individuals.

Raw residuals are calculated in the usual OLS fashion by using the model to predict the outcome and then subtracting that prediction from the observed outcome. So for a two-level model, raw residuals could be calculated as in equation 5.8, where y_{ij} is the

observed value and $\hat{y_{ij}}$ is the fitted value from the model.

$$r_{ij} = y_{ij} + \hat{y_{ij}} \tag{5.8}$$

For hierarchical modelling however, such residuals are unsatisfactory since ideally it would be possible to decompose these raw residuals into the contributions from the levels modelled. In other words, it would be useful to be able to think of the raw residual for an individual from class A in school B, as being the sum of that individual's departures from the class A average, class A's departure from the school B average, and school B's departure from the overall average. This is intuitively appealing, since with multiple error terms multiple residuals would be expected. In order to calculate residuals for the second level, the following procedure is followed. All of the raw residuals for the j^{th} grouping are averaged (denoted by r_{+j}). Then the residual for the j^{th} group is calculated as in equation 5.9, where n_j is the number of lower level units in level-two unit j.

$$\hat{\tau_{0j}} = \frac{\hat{\sigma}_u^2}{\hat{\sigma}_u^2 + \frac{\hat{\sigma}_\epsilon^2}{n_j}} r_{+j} \tag{5.9}$$

Residuals calculated in this way are sometimes called shrunken residuals, since the multiplier of r_{+j} is always less than one (since all of σ_u^2, σ_ϵ^2 and n_j are strictly non-negative). When the individual-level variance (σ_ϵ^2) dominates, this multiplier is much smaller than one meaning that the higher level residuals calculated will be shrunken towards zero. This makes sense, since when the individual level dominates, information on the higher level units will be relatively scarce. The shrinkage could also be large if n_j, the j^{th} group sample size is, small. Again, in such a situation the higher level residual should be small, since there is little information for that group. The individual-level residual is whatever is left after the higher level residuals have been subtracted from the raw residual. So for the two-level model, the individual-level residual is calculated in equation 5.10. These ideas can be extended to include three or more levels.

$$\hat{\epsilon_{ij}} = r_{ij} - \hat{\tau_{0j}} \tag{5.10}$$

Binary response variables can also be modelled using hierarchical methods. The response is assumed to be distributed binomially as in equation 5.11, where y_{ij} is the response for the i^{th} individual in the j^{th} group, and π_{ij} denotes the probability that the i^{th} individual in the j^{th} group is a success.

$$y_{ij} \sim \text{Binomial}(1, \pi_{ij}) \tag{5.11}$$

Similar to the non-hierarchical case, some link function is used to transform the probabilities of success from the range $[0, 1]$ to the range $(-\infty, +\infty)$ and the result is modelled, as opposed to modelling the response indicator variable itself. The situation is modelled with random intercepts as in equation 5.12 using the logit link.

$$\text{logit}(\pi_{ij}) = \beta_{0j}x_0 + \beta_{1j}x_{1ij} + \tau_{0j} + e_{ij} \tag{5.12}$$

In the binomial modelling situation individual-level residuals cannot be calculated since it is group proportions (or probabilities) that are modelled. This makes the calculation of a meaningful ICC coefficient more difficult. It would appear that this problem is intractable, however if certain assumptions are made, then an estimate for the level-one variance can be found. The following example is paraphrased from *Multilevel Modelling* by Snijders and Bosker (1999). If Y is a binary variable, say passing or failing an exam, then the distribution of Y could be expressed as in equation 5.13, where Y' is the underlying continuous score for the individual.

$$Y = \begin{cases} 1 & \text{if } Y' \geq 40 \\ 0 & \text{if } Y' < 40 \end{cases} \tag{5.13}$$

If a multilevel model is fitted to this underlying variable Y', then the individual-level residual can be examined to see if it comes from a logistic distribution. If the underlying variable is unknown then this assumption cannot be tested, however if a logistic model is to be fitted for Y then it must be assumed that the individual-level residual has a logistic distribution. This implies that the cumulative distribution function of individual-level residuals is the logistic function, as in equation 5.14.

$$f(X) = \frac{exp[\frac{-(X-\alpha)}{\beta}]}{\beta[1 + \frac{exp[-(X-\alpha)}{\beta}]^2} \tag{5.14}$$

This distribution has mean α and variance $\frac{\pi^2\beta^2}{3}$. With $\alpha = 0$ and $\beta = 1$, the variance is $\frac{\pi^2}{3}$. This estimator for the individual-level variance *"may be reasonable where the (0,1) response is, say, derived from truncation of an underlying continuum such as a pass/fail response based upon a continuous mark scale"* (Goldstein et al., 2002). If the individual-level residual from the underlying continuous score is distributed as a standard normal distribution, a probit model should be used instead of a logistic one.

5.2 Application of Hierarchical Modelling

The Caerphilly Health and Social Needs Study dataset is hierarchical since each individual is nested within a household, which is nested within a postcode, which is nested

within an enumeration district, which is nested within an electoral ward. People in the same household are likely to have more in common with one another than they would with people from other households, e.g. the financial situations of residents of the same household are undoubtedly quite similar, as are their social classes. This non-independence of individual characteristics within households (and to a lesser extent postcodes, enumeration districts and electoral wards) is one of the justifications for the use of hierarchical modelling.

It is useful to investigate some simple hierarchical models. If the relationship between mental health and area deprivation is to be investigated there are a number of models that could be used. A basic approach would be a three-level model with individuals nested enumeration districts, nested within wards. Suppose it is of interest to predict the mental health of individuals using their area deprivation as a predictor. The area deprivation score used is the Townsend index, described earlier in section 2.2. Here the Townsend index is calculated at enumeration district level. Enumeration districts are smaller than wards (typically each ward contains about 10 enumeration districts). A simple model to investigate the relationship between the Townsend index and mental health is described in equation 5.15.

$$MentalHealth_{ijk} = \beta_0 x_0 + \beta_1 Townsendscore_{ijk} + e_{ijk} \tag{5.15}$$

Here $MentalHealth_{ijk}$ is the mental health score for the i^{th} individual in the j^{th} enumeration district, in the k^{th} ward. Notice the β_0 term has no j or k subscripts indicating it is a constant for all enumeration districts and all wards, and the same is true for the β_1 term. So, this is a constant intercept, constant slope model. The relationship fitted is displayed in figure 5.1. This shows that as the Townsend score increases (more deprivation), the mental health score decreases (worse mental health). Next, the intercept is allowed to vary between wards. Equation 5.15 becomes equation 5.16.

$$MentalHealth_{ijk} = \beta_{0k} x_0 + \beta_1 Townsendscore_{ijk} + e_{ijk} \tag{5.16}$$

Notice that the intercept term now has a k subscript, indicating it varies from ward to ward. Figure 5.2 now shows 36 parallel lines, indicating the relationship between mental health and enumeration district deprivation for each ward. In this model the relationship between ED measured deprivation and individual mental health is allowed to have different intercepts in different wards but is constrained to have a fixed slope (whether this particular example is a sensible approach to take is a different matter). This model can be further complicated by allowing the slopes to vary from ward to ward, as in equation 5.17.

$$MentalHealth_{ijk} = \beta_{0k} x_0 + \beta_{1k} Townsendscore_{ijk} + e_{ijk} \tag{5.17}$$

Figure 5.1: Relationship between Mental Health and Enumeration District Deprivation- constant intercept, constant slope

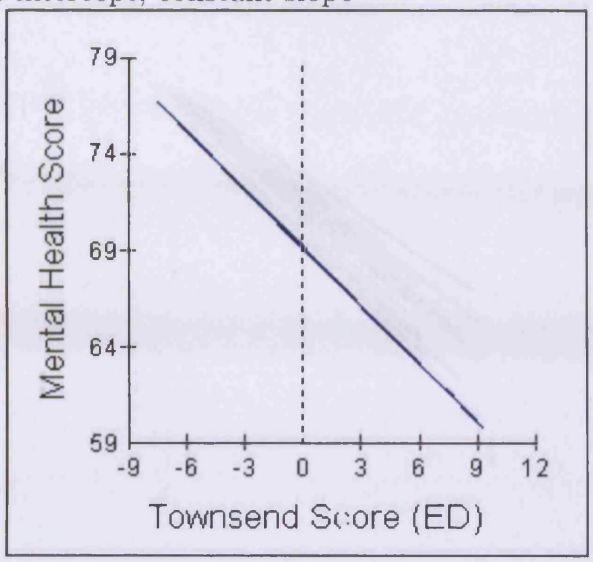
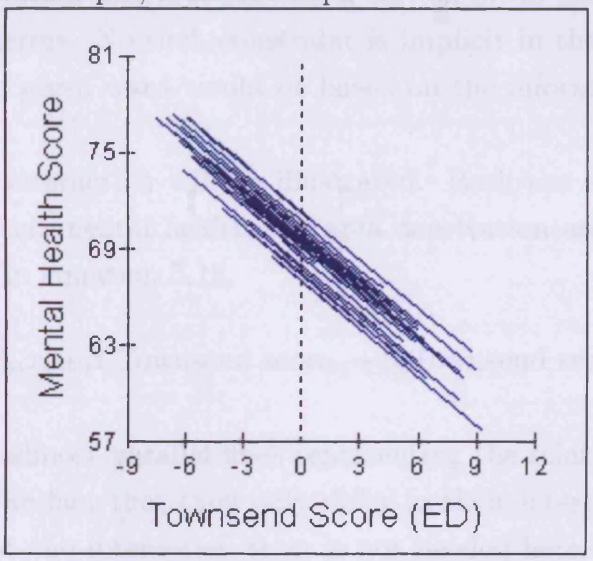
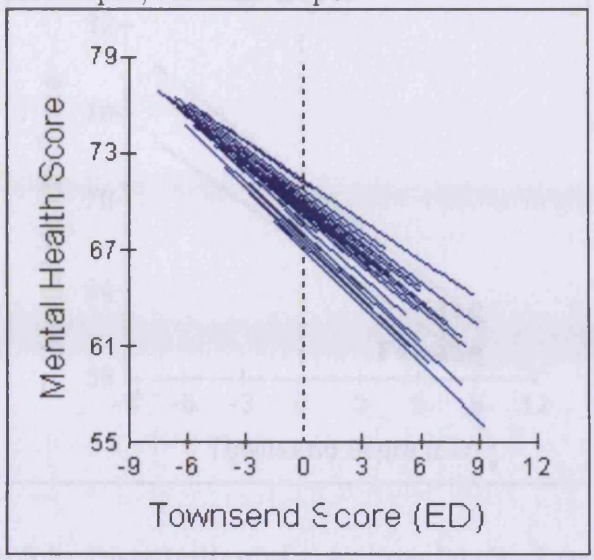


Figure 5.2: Relationship between Mental Health and Enumeration District Deprivation- random intercepts, constant slope



Here the slope coefficient acquires a k subscript. In this model the slope of the relationship between ED measured deprivation and individual mental health is allowed to vary between wards. This produces figure 5.3 This fanning out pattern indicates that the intercept terms in the model are positively correlated with the slope terms, so larger intercepts mean larger slope terms. This produces the figure 5.3 where the ward with the largest intercept term has the largest slope (i.e. the smallest negative slope). A similar style plot could, of course be produced using an OLS model, with thirty five intercept terms and thirty five slope terms. Hierarchical modelling however is fundamentally different from this approach in that the distributional assumptions

Figure 5.3: Relationship between Mental Health and Enumeration District Deprivation- random intercepts, random slopes



are placed on both the intercept and slope terms. The intercept terms are assumed to be drawn from a normal distribution with a certain mean and variance. The same is true for the slope terms. No such constraint is implicit in the OLS formulation, in which the model for a given ward would be based on the information from that ward only.

Now a cross-level interaction will be illustrated. Both the slope and intercept of the relationship between mental health and area deprivation are allowed to differ for males and females as in equation 5.18.

$$MentalHealth_{ijk} = \beta_{0j}x_0 + \beta_1 \text{Townsend score}_{ij} + \beta_2 \text{Townsend score}_{ijk} : \text{Female}_{ijk} + e_{ijk} \quad (5.18)$$

Figure 5.4 shows two almost parallel lines representing the relationship for males and females separately. The fact that they only differ in their intercepts, (i.e. they share a slope) indicates that the interaction term is not needed here. Essentially, this is a graphical illustration that the interaction term coefficient must be close to zero (since this coefficient indicates the difference between the male and female slope coefficients). The coefficient for the interaction term is -0.135, with a standard error nearly as big as that (0.124), thus reinforcing the conclusion that there is no evidence for a cross-level interaction here.

Since this is a three-level model there are three sets of residuals to examine. Figure 5.5 displays histograms of the three levels of residuals. Each level's residuals should be centred around zero and be normally distributed in order to satisfy the assumptions of hierarchical modelling. Essentially, any diagnostic that should be performed on the residuals from an OLS regression should be performed for every level of residuals in

Figure 5.4: Cross-level interaction between mental health, Townsend score and gender

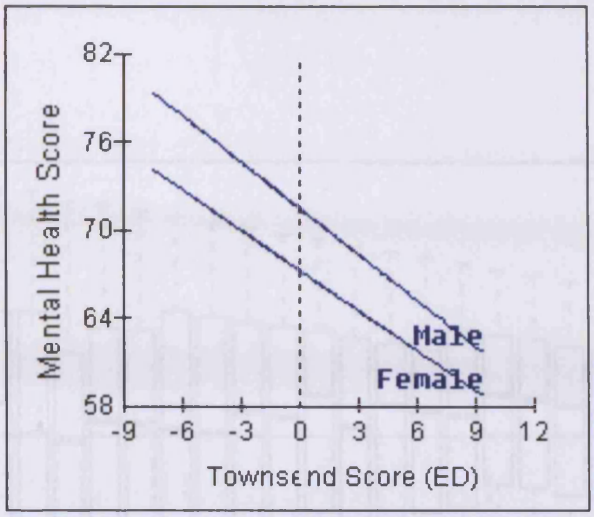
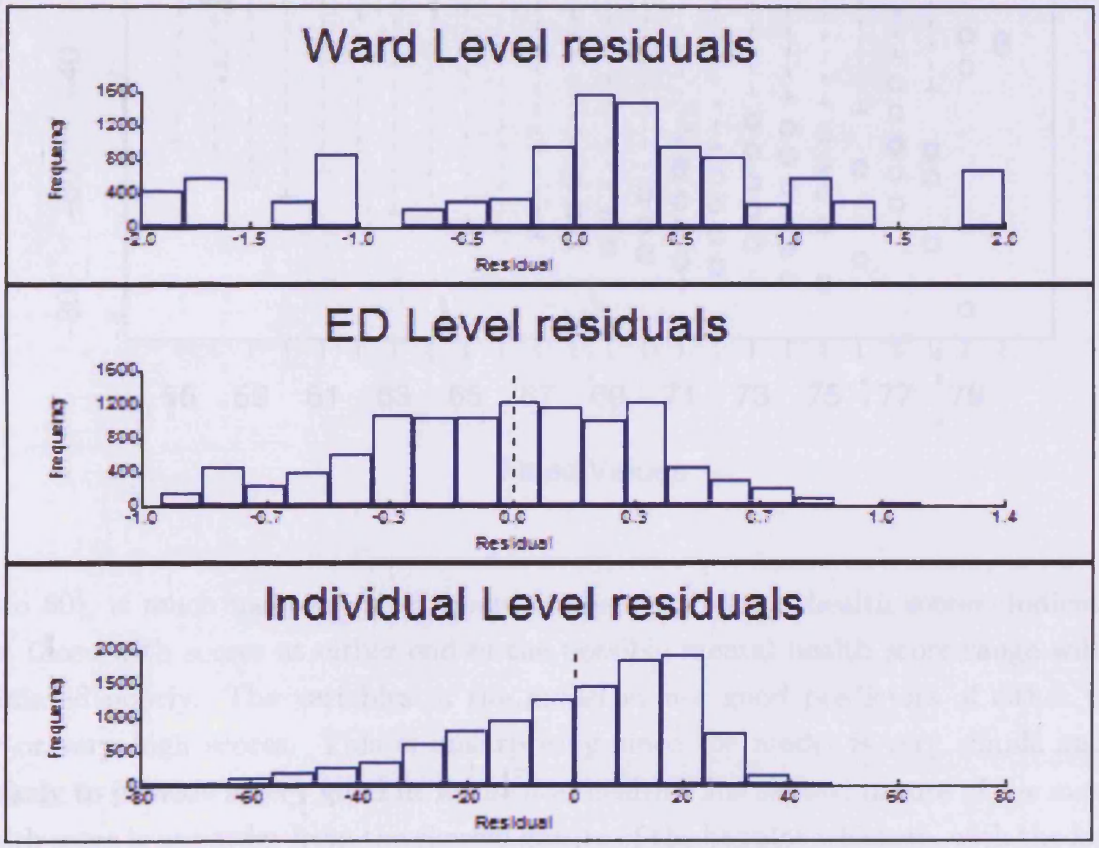
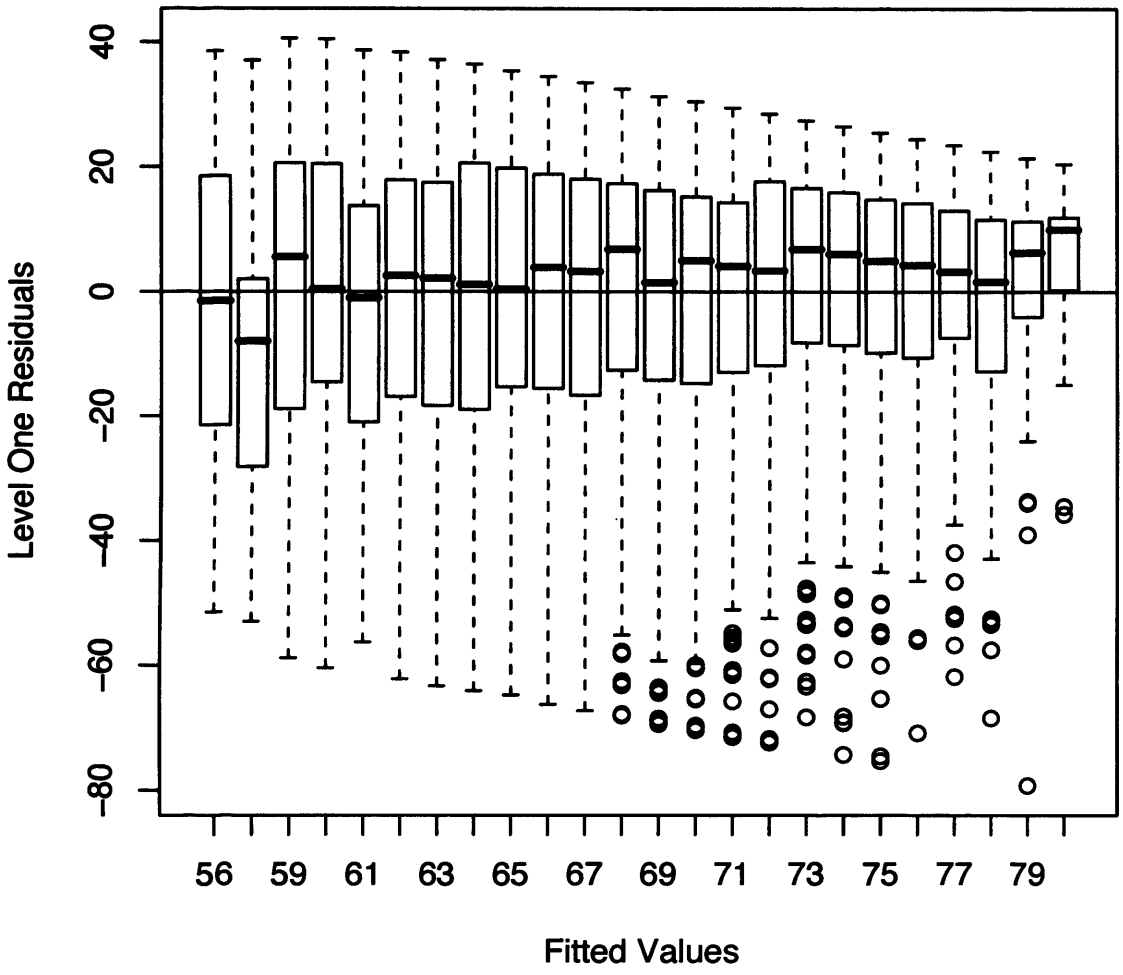


Figure 5.5: Distribution of the three levels of residuals



a multilevel model. For the individual-level residual, boxplots of fitted values against residuals are given in figure 5.6. If the model fits well, these boxplots should be symmetrically distributed about zero and should have constant variance. The sloping upper and lower bounds of the whiskers on these boxplots is a result of the mental health score being bounded between zero and 100. The range of fitted values (from

Figure 5.6: Individual-level residuals: fitted values against residuals

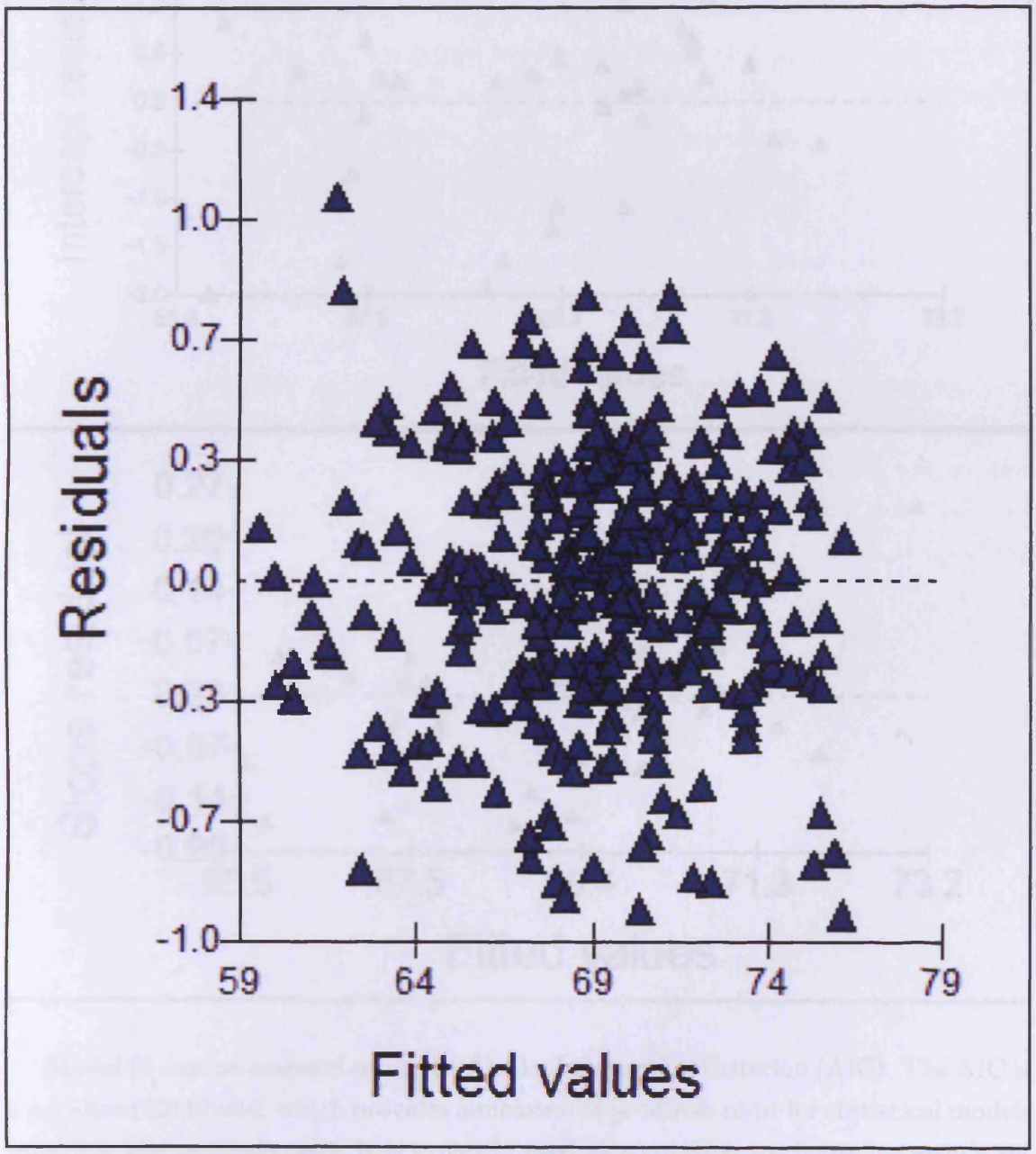


56 to 80), is much narrower than range of observed mental health scores, indicating that those with scores at either end of the possible mental health score range will be estimated poorly. The variables in the model at not good predictors of either very low or very high scores. This is unsurprising since the model is very simple and is unlikely to provide a very good fit for mental health. The skewed nature of the mental health score is apparent from the skewed nature of the boxplot whiskers, with the lower whiskers being larger than the upper ones. Since mental health is a very complicated to measure and predict and this model is extremely simple (with just gender, Townsend score and an interaction between the two fitted), this is unsurprisingly. The model fit is not sufficiently poor to be a worry.

The same diagnostic can be performed for the enumeration district level residuals, as in figure 5.7. Each point represents one of the 325 EDs in Caerphilly county

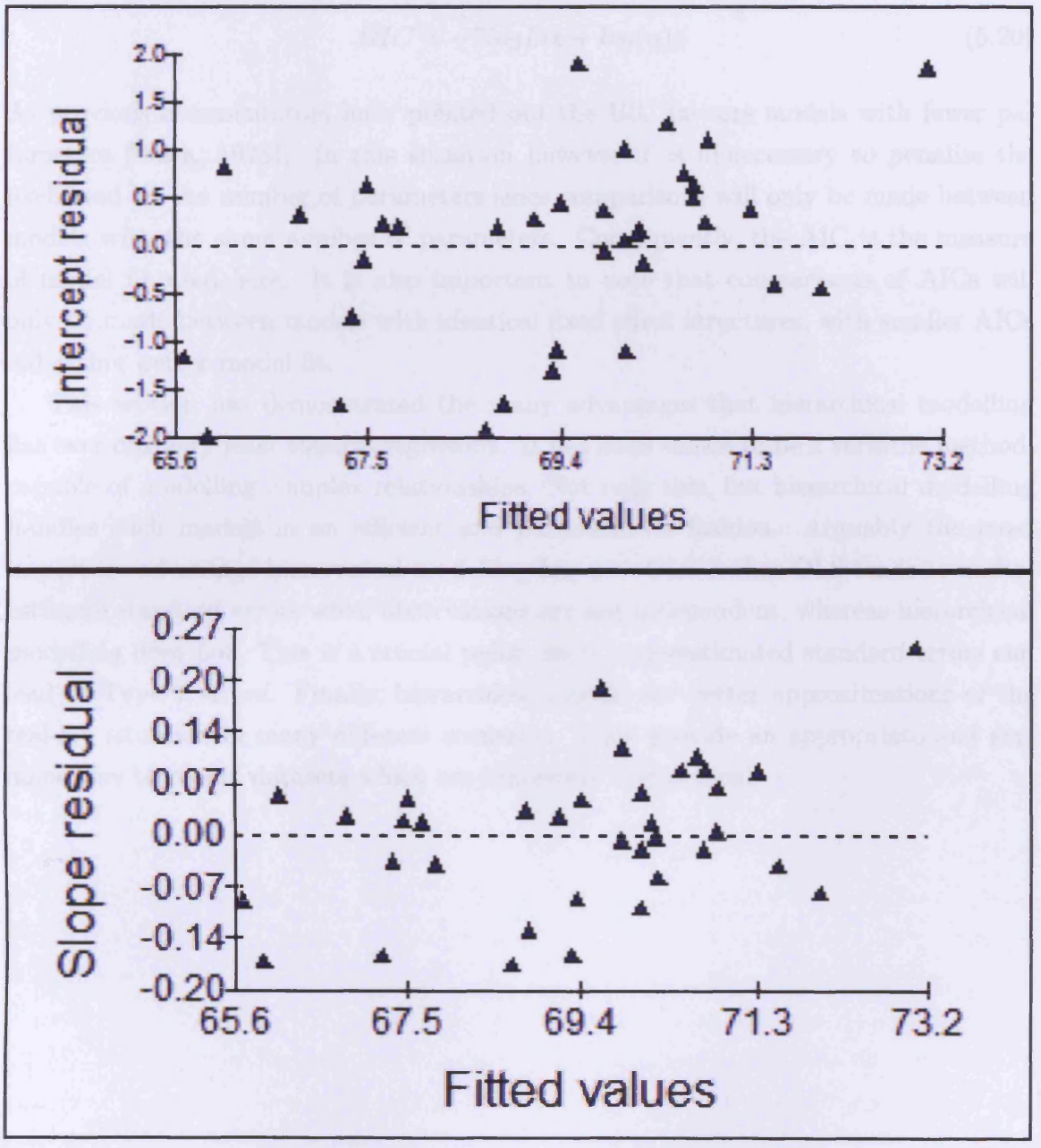
borough. This figure appears to show random scatter about the horizontal zero line, which indicates that this model satisfies the assumption of normally distributed errors, unrelated to the explanatory variables at this level.

Figure 5.7: Enumeration district level residuals: fitted values against residuals



Finally, the fitted values against the ward-level residuals are examined in figure 5.8. Since both the intercept and slope of the relationship between deprivation and mental health are allowed to vary randomly at this level, there are two types of residual to plot at this level. In both of the plots in figure 5.8 each point represents a ward in Caerphilly county borough. Both plots show random scatter about the horizontal line at zero and constant variance, providing no evidence for assumption violation.

Figure 5.8: Ward-level residuals: fitted values against residuals



Model fit can be assessed using the Akaike Information Criterion (AIC). The AIC is a penalised likelihood, which provides a measure of goodness-of-fit for statistical models (not only hierarchical ones). It is typically used as a criterion to choose between a set of models. Its form is given in equation 5.19, where p is the number of parameters in the model.

$$AIC = -2logLik + 2p \tag{5.19}$$

Another oft quoted measure of model fit is the Bayesian Information Criterion (BIC), also known as the Schwartz Information Criterion, given in equation 5.20. Here n is

the number of observations.

$$BIC = -2logLik + log(n)p \tag{5.20}$$

As previous commentators have pointed out the BIC favours models with fewer parameters (Sawa, 1978). In this situation however it is unnecessary to penalise the likelihood by the number of parameters since comparisons will only be made between models with the same number of parameters. Consequently, the AIC is the measure of model fit used here. It is also important to note that comparisons of AICs will only be made between models with identical fixed effect structures, with smaller AICs indicating better model fit.

This section has demonstrated the many advantages that hierarchical modelling has over ordinary least squares regression. It has been shown to be a versatile method, capable of modelling complex relationships. Not only this, but hierarchical modelling handles such models in an efficient and parsimonious fashion. Arguably the most important advantage hierarchical modelling has over OLS is that OLS tends to underestimate standard errors when observations are not independent, whereas hierarchical modelling does not. This is a crucial point, since underestimated standard errors can lead to Type 1 errors. Finally, hierarchical models are better approximations of the real-life situation in many different scenarios. They provide an appropriate and germane way to model datasets which are inherently hierarchical.

Chapter 6

Investigating the household level

6.1 Introduction

The choice of hierarchy is an extremely important consideration in any attempt to examine the contextual determinants of mental health (or any other outcome for that matter). There are many possible hierarchies that could be included in the multilevel model, however not all of them will be useful for modelling mental health. The main reason to include a level is if the measurements at that level are not expected to be independent, or to have some influence on one another. This is the situation that motivated the development of hierarchical analysis. Excluding such a level results in the variance structure of the model being incorrect, and can lead to underestimates of coefficient standard errors (Goldstein, 2003). One justification for including a given level is if there are variables that relate to risk factors for, or determinants of, the response variable at that level. In such a situation, excluding the level means that these variables will be assigned to the incorrect level. Another obvious reason to include a level is if it explicitly relates to the research question. This could be the case if the hypothesis of interest concerns the relative contributions of different contexts to the outcome of interest. Perhaps a less valid reason to include a level is if the information is freely available. Whatever the reasoning behind the inclusion of contextual levels, there is a separate methodological question regarding how statistically valid or advisable it is to include a given level. This chapter will investigate this question with specific reference to the problem of what will henceforth be called sparseness.

Sparseness refers to the situation where a given level, for whatever reason, has low numbers of sub-units in the hierarchy. An example might be an observational study which is interested in examining the differences in General Practitioner (GP) prescribing patterns across different areas within the same country. In such a situation, primary care organisation might be the highest level and GP the lowest level. Another level that could be considered is general practice, since GPs prescribing patterns within the

same practice might not be independent. If it were the case however, that a large number of general practices have only one GP in them (or perhaps if only one GP per practice agreed to be included in the study), this raises an interesting problem. For a large proportion of practices, the general practice-level is completely conflated with the individual-level. This is a potential problem for hierarchical modelling since it is impossible to ascertain the relative contributions of each of these levels to the variability in the response.

Perhaps it is decided therefore to exclude the general practice from the hierarchy. This too could be a potential problem, since there may be excellent methodological reasons for including the general practice-level. It is unclear which course of action is better; to exclude a level that may be considered an important context on the causal pathway to the outcome of interest, or to include a level with worrying identifiability issues. This is quite a general problem and can occur in many different situations. Consider a study which takes repeated measurements of some outcome which is expected to vary over time (such as blood pressure) from a cohort of volunteers. Such studies may have the problem that some individuals volunteer for many measurements, but others volunteer for very few.

What seems obvious is that if there is only one observation in each unit (i.e. if a lower level is completely confounded with a higher level), then both of those levels should not be included. However, if at the other extreme each higher level contains many individual-level observations, then that level should be included. What should be done in situations that lie between the aforementioned scenarios is the main question. How many repeat observations should one have from each individual before it is worthwhile including individuals as a level (with measurements as the lowest level)? If there were two measurements from half of the cohort and only one from the rest, are the variance component estimates for the individual level reliable? What is the effect of including sparse levels on model fit, or coefficient estimation, or the precision of variance components? What level of sparseness should be tolerated before a given level is excluded from an analysis? The answers to these questions are not clear and will be investigated in this chapter.

As has already been mentioned, the problem is a general methodological one for hierarchical modelling; however the question is motivated by the CHSNS dataset. The household level is an option for inclusion in the hierarchy. There are a number of reasons why this might be a useful thing to do. Firstly, there are a number of important socioeconomic variables that are undeniably measured not at the individual level, but at the household level, e.g. gross household income, tenure and council tax band. Furthermore, it seems reasonable to suggest that individuals living under the same roof would, in general, be similar in terms of diet, affluence, social class as well as (by definition) living conditions. Such commonality of exposure is exactly what hi-

erarchical modelling is designed to deal with, as described in chapter 5. Finally, there is a growing body of literature which indicates that the household level is a useful level to include in multilevel analyses of the contextual determinants of mental health (Chandola et al., 2003, 2005; Propper et al., 2005; Weich et al., 2003b, 2006, 2005), which will be examined and critiqued in section 6.2.

It seems that the household context is an important one to consider. However, some datasets may not be well equipped to model the household level due to suffering from sparseness. The CHSNS dataset is one such dataset. The 10,653 complete questionnaires that comprise the CHSNS dataset represent 9,827 different households. This translates to just 15.2% of the individual's responses providing any information on the household level. Clearly, the CHSNS dataset has very few multiple response households; however the question remains, does it have enough?

The British Household Panel Survey (Taylor et al., 2005) is better equipped to investigate the impact of household as here the sampling unit was households. In the BHPS every member of a household containing a participant in the study is interviewed. Even adopting this approach does not completely avoid the problem of household sparseness, since a considerable proportion of the population reside alone. In fact 40% of the households in the BHPS are single response households, representing 22.2% of the individuals in the dataset. This means that 11,194 (73.8%) of the 14,669 individuals can be used to obtain information about the household level in the BHPS study wave 9. It seems more reasonable to include the household level when using this dataset, but the impact of one fifth of the dataset belonging to single response households is unknown. Table 6.1 compares the two datasets.

As in any such situation, an obvious alternative is to exclude the sparse level from the analysis. Some work has been done on the consequences of ignoring an intermediate level in a hierarchical model (Moerbeek, 2004). It was discovered that the exclusion of such a level resulted in the variation that would have been attributable to that level being split between the higher and lower levels. This paper did not investigate the consequences of including or excluding a sparse level in a multilevel analysis.

This chapter will address the third stated objective of investigating the robustness of multilevel modelling techniques to sparse levels of data. This will be addressed under the following headings.

1. Evidence regarding the importance of including households in multilevel analyses of mental health from the literature will be reviewed.

2. A simulation study will be described, comprising four sections, each addressing a different situation.

3. The results from each simulation study section will be presented and interpreted.

Table 6.1: Breakdown of numbers of household responses

Study	No. of responses	1	2	3	4	5 or more
CHSNS	**Frequency**	9,035	761	28	3	0
	% (of households)	91.9	7.7	0.3	0.0	0.0
	% (of individuals)	84.8	14.3	0.8	0.1	0.0
BHPS	**Frequency**	3,475	4,194	806	270	50
	% (of households)	39.5	47.7	9.2	3.1	0.6
	% (of individuals)	22.2	53.7	15.5	6.9	1.7

4. The discussion section will present the strengths and limitations of the study as well as highlighting the implications of the chapter for previously published studies

5. The findings of the chapter will be summarised and discussed and conclusions drawn.

6.2 Modelling the household level in studies of people, places and mental health

6.2.1 Commentary

Relevant papers were defined to be those that had a mental health outcome as the variable of interest, whose sample was from the general population, which employed hierarchical methods and included household in the hierarchy. A combination of database searching and expert knowledge was used to identify such papers.

Six papers were found which advocated the inclusion of households in the hierarchy for multilevel analyses of mental health, either explicitly or implicitly. Interestingly all six papers identified analysed the same dataset, that is the British Household Panel Survey (BHPS).

The first of these that will be examined is by Weich et al (2003b). The authors note that:

> *"Most previous studies have failed to take into account variability between households, resulting in overestimates of variance at higher levels."*

This paper compared the prevalence of common mental disorders (CMD) in urban and rural areas, as well as the association between socioeconomic deprivation and CMD. Data from the first wave of the British Household Panel Survey were used to investigate this association. Mental health was measured using the GHQ-12 (Goldberg & Williams, 1988) introduced in section 3.1. The GHQ-12 score was used to assess whether individuals were classified as a case or not and this dichotomous variable was used as the response in logistic hierarchical models. Socioeconomic deprivation was measured using three different proxies: the Carstairs index of socioeconomic deprivation (Morris & Carstairs, 1991), the Office of National Statistics classification of wards into 14 groups, (Wallace & Denham, 1996), and finally a measure of population density. The hierarchy employed comprised 8,978 individuals nested within 4,904 households, nested within 642 wards. Firstly a null model was fitted to the data. The variance attributable to the household level was assessed for significance using the Wald test, and found to be significant (variance 0.565, SE 0.077) (Wald statistic 53.84, p-value $<$ 0.001), but the ward-level variance was not (variance 0.035, SE 0.026) (Wald statistic 2.53, p-value 0.11). The standard estimate of the individual-level variance component of $\frac{\pi^2}{3}$ was used, indicating that the ICCs for the household- and ward-level variance components were 0.145 and 0.009 respectively. These tests were supplemented using MCMC methods. Using MCMC produced a household-level variance component of 0.794 (credible interval of 0.54-1.057), and a ward-level variance component of 0.032 (credible interval of 0.001-0.098). Controlling for individual, household and area-level covariates had little effect on the household-level variance, while the ward-level variance was further reduced. The authors conclude by saying:

> *"Our results are consistent with previous research suggesting that features of households (or areas) may be most salient for those who are not in work, and who spend the most time at home"*

This paper explicitly advocates the modelling of the household level.

The next paper of interest examined the association between self-rated health and four measures of social position, namely, occupational class, household social advantage, personal income and household income (Chandola et al., 2003). This study used the first and eighth waves of the BHPS. Only individuals with "excellent" or "good" self-rated health at wave one were included, in order to select out a healthy cohort. This resulted in a dataset of 10,264 individuals nested within 5,511 households.The

outcome variable was self-rated health measured at wave 8 using the GHQ-12 and dichotomised into those that rated their health between good and excellent, and those that rated their health fair to poor. Logistic models were fitted to this binomial response. The hierarchy employed was a two-level one with individuals nested within households. Age, sex and employment status were also included in the analysis. Five models were fitted: the first four included one of the four measures of social position while the last model included all four. For all five models the variance at the household level was deemed significantly different from zero since the variance estimate was over twice its standard error (the variance ranged between 0.76 and 0.85, while the standard error ranged between 0.17 and 0.18). These same five models were fitted using longitudinal weights at both individual and household levels (again in order to make the sample more representative of the general population). None of the five models produced "significant" household-level variance in these models. The authors acknowledge the problem of sparseness in their explanation of this non-significance of the household-level variance component by commenting:

> "*This may be attributable to the comparatively greater proportion of single person economically inactive households, which reduces the likelihood of distinct household-level effects separate from individual-level effects* "

The authors conclude by recommending further investigation of the similarities in health between household members, clearly indicating that they believe there is a contextual effect of household on individual mental health.

A paper published in 2005 (Weich et al., 2005) was explicitly concerned with estimating the variance contributions of the individual-, household- and electoral ward-levels in hierarchical models investigating the onset and maintenance of the common mental disorders as measured by the GHQ-12. This was done using data from the first two waves of the BHPS. The onset cohort was defined to be those individuals classified as a non-case of CMD at wave one but classified as a case at wave two, while the maintenance cohort were those who met the case criteria for CMD at both waves. The onset cohort comprised 5,809 individuals nested within 3,679 households, nested within 615 wards, while the maintenance cohort comprised 1,850 individuals nested within 1,566 households, nested within 511 wards. The outcome variable was a binomial one (indicating whether the individual was an onset or maintenance case) and so logistic models were employed. The standard estimate of level-one variance of $\frac{\pi^2}{3}$ was used, as described in section 5.1. Three models were fitted for both cohorts: a three-level null model, a three-level model with individual- and household-level variables, and a three-level model with individual- and household-level variables and an area-level deprivation variable (Carstairs index). The ICC coefficients for the household level for the onset group ranged between 0.14 and 0.17, while for the maintenance

group it ranged between 0.12 and 0.33. The GHQ-12 was also modelled as a continuous variable, providing more reliable ICCs. These models consistently assigned about 12% of the variation in the response to the household-level.

Another paper published in 2005 (Propper et al., 2005) investigated the association between neighbourhood and common mental disorders (both levels of CMD and changes in CMD). Since the BHPS was again used, the GHQ-12 was the measure of mental health. This study used information from ten waves of the BHPS (1991-2000) providing information on 8,184 individuals nested within 4,341 households. Three-level hierarchies were fitted with individuals nested within households nested within "bespoke neighbourhoods". These bespoke neighbourhoods were created for each individual and can be thought of as containing the nearest 500 people to each individual's home address. Principal component analysis was used to create five components derived from 18 socio-economic and demographic variables extracted from 1991 census data. These components were calculated for the bespoke neighbourhoods and were called Disadvantage, Mobility, Age, Ethnicity and Urbanness. Tests of significance of variance components were carried out using the log- likelihood ratio test. Firstly, levels of CMD were investigated using a sample size of 8,184 individuals nested within 4,341 households. The response variable was GHQ-12 score. Five null models, each with one of the components included, were fitted. The household-level variance was deemed significant for all of these five models (p-value < 0.01 for all five models). When individual and household variables were included (age, gender, ethnicity, education, net household income, number of adults, number of children, house tenure type, and employment status of the head of the household) the significance remained (p-value < 0.01 for all five models). Household-level ICCs for all ten of these models ranged between 0.13 and 0.14. This procedure was followed again, but this time the response was 5-year change in GHQ-12 score. The sample size for this analysis was 7,047 individuals nested within 4,377 households. Here, three of the five models containing only one of the components produce significant household-level variances (for the Mobility, Age and Ethnicity components) (p-values < 0.01 for all three). This situation remains unchanged when individual and household-level variables are included. Household-level ICCs for these models remained at 0.15. The authors conclude by saying that the work:

> "*suggests that people and their households should be the focus of policy effort to alleviate the common mental disorders examined here.*"

A paper published in 2005 (Chandola et al., 2005) used wave 9 of the BHPS dataset to investigate both physical and mental heath. One of the study's goals was to investigate whether longitudinal analyses of area effects on health need to take account of clustering at the household level. As discussed in section 3.4.3, wave nine of the

BHPS included the SF-36 and this was used to assess both physical and mental health, using the PCS and MCS. The dataset used comprised 10,264 individuals nested within 5,511 households who had complete information at the first nine waves of the BHPS. Fourteen different models were fitted. Firstly, a single level null model was fitted using individuals. The second model was a two-level null model with individuals nested within households as recorded at the first wave of the BHPS. The third model was a cross classified two-level null model with individuals nested within the households they resided in over the course of the study. The above two models were repeated with households replaced by electoral wards. The sixth model was a three-level null model with individuals nested within wave one households nested within electoral wards. The seventh model was the same as the above, was allowed for multiple household membership. All seven of the above variance component models were repeated controlling for age, gender, marital status, employment and smoking status. ICCs along with credible intervals were calculated for the variance components. The household-level ICCs were very consistent for the four null models which included household, ranging only between 0.2 and 0.25. None of the credible intervals include zero (minimum lower 95% credible interval limit: 0.17, maximum upper 95% credible interval limit: 0.27). The authors take this as evidence that the "*household-level variance is statistically significant and different from zero*". Ward-level ICCs were also calculated and ranged between 0.02 and 0.06. For the four full models including household the ICC coefficients ranged between 0.09 and 0.15. The authors conclude that the household level has a contextual effect on people's mental health, indicating the importance of including households as a level. Moreover, they state that "*wherever sample designs select clusters as sampling units, such units should be taken account of in any subsequent analyses*".

The final paper discussed here used data from the first two waves of the BHPS to investigate rural/non-rural differences in the onset and maintenance episodes of common mental disorders. The hierarchy was composed of 7,659 individuals living in 4,338 households nested within 626 electoral wards. Variance components for the household level were not reported, however the authors support their inclusion of the household level by saying:

> "*Our estimates of standard errors for associations between area-level exposures were less prone to bias than those arising from studies in which individual- and household-level exposures were conflated*"

Table 6.2 summarises the household-level information included in each study. The consensus that these papers reach is that it is necessary to recognize that households are a useful and important level to model in studies of this type. The variance attributable to the household level is the largest in the two papers (Chandola et al., 2003, 2005) which model household level as the highest level. This is to be expected as

Table 6.2: Summary of papers which advocate modelling household as a level

Year	Study	No. of levels	Avg. per Household	Household ICC
2003	Weich et al	3	1.83	0.145
2003	Chandola et al	2	1.86	0.19-0.21
2005	Weich et al	3	1.77	0.12
2005	Chandola et al	2	1.86	0.09-0.29
2005	Propper et al			
	CMD prevalence	3	1.89	0.13-0.14
	5 year change in CMD	3	1.61	0.15
2006	Weich et al	3	1.77	Not reported

any contextual variation at excluded higher levels will be assigned to the highest level. The remaining papers consistently assign proportions of between 0.12 and 0.15 to the household level.

6.2.2 Critique

The results presented in the previous section indicate that the variance attributable to households in hierarchical models investigating mental health is likely to lead to ICC estimates of at least 0.1. This finding is fairly consistent across studies, which is perhaps unsurprising since they all use the same dataset (albeit different waves). The absence of other studies recommending the inclusion of households as a level is worrisome but this may be a result of the difficulty and expense involved in interviewing entire households. Apart from this worry however, there is a general problem shared by all of these papers regarding the methods used to assess the statistical significance of the household variance components. More pertinently, is it sensible or meaningful to discuss the statistical significance of variance components at all? Variances are, by definition, non-negative, meaning that the absolute minimum they can attain is zero. If the true value of a given level's variance component is zero, it would certainly be important to identify that. However, if the true value of a given level's variance component represents a tiny fraction of the total variance, it may be similarly important to reject that variance component as unimportant. The question therefore, is less about the probability that a given level's variance component is equal to zero, but more about how large a given level is likely to be. These methods varied from paper to paper and all have their weaknesses. This section will examine and critique the three main approaches to significance testing of variance components: standard significance testing, Wald statistics, and MCMC credible intervals.

Standard significance testing

Standard significance testing was used in the first paper by Chandola et al (2003)and involves constructing a 95% confidence interval around the variance component estimate using its standard error. If this interval includes zero, then the variance component is not deemed to be significantly different from zero. This is a crude way to test significance since standard 95% confidence intervals are not constrained to be positive, while variances are. This method also assumes that the variance component is distributed Normally which is impossible since it is strictly non-negative.

Wald Statistics

Wald statistics (Birkes, 1998) were used in two of the papers by Weich et al (2003b; 2005) and are a relatively crude way of calculating a p-value. It is described in equation 6.1, where $\hat{\theta}$ is the maximum likelihood estimate of θ (the variance component) and the null hypothesis is that $\hat{\theta} = \theta_0$. Under the null hypothesis, W is distributed as a chi-square with one degree of freedom.

$$W = \frac{(\hat{\theta} - \theta_0)^2}{var\hat{\theta}} \tag{6.1}$$

The Wald statistic has been criticised for having poor power in certain situations including binary logit models when the alternative is far from the null (Fears et al., 1996) and when the likelihood function is not well approximated by a quadratic (Pawitan, 2001). Another drawback is that Wald confidence intervals are always symmetric, which may not reflect reality, in particular if the confidence interval includes negative values for a strictly non-negative parameter (e.g. variance components). In situations with large samples however, it is less prone to such problems and is computationally very simple.

MCMC credible intervals

MCMC credible intervals were reported in three of the six papers summarised (Weich et al., 2003b, 2005; Chandola et al., 2005). This is a sophisticated way of assessing the variability of variance components and approaches the problem in a sensible way by providing a range of values where the variance component is likely to lie. It involves using MCMC methods to sample from the posterior distribution of the household-level variance component. The interpretation of a 95% credible interval in such a situation is not straightforward however, due to the fact that the absolute minimum value belonging to this distribution is zero. Even if there is no contextual effect of household the variance attributable to the household level is very likely to be slightly greater than zero due to chance. This is illustrated in the first paper by Weich et

al (2003b), where the p-value for the ward-level variance component was 0.11, yet the MCMC credible interval surrounding the variance component was 0.001-0.098, excluding zero. Similarly, the second paper by Weich et al (2005) reports an MCMC credible interval for a household variance component ranging from 0.03-0.97 (point estimate of 0.55). While this CI does not contain zero, it does range sufficiently low enough to call into question whether this variance component could not be due to chance.

Summary

In summary, none of the significance testing methods are completely satisfactory. This is due to the nature of high-level variances being non-negative and typically small. The absence of other studies advocating the household level, while perhaps a consequence of the difficulty of collecting datasets with enough multiple response households to model household, is still a concern. It would be useful to provide some guidelines to researchers about when it is appropriate to include the household level. In order to do this it is neccessary to investigate the effects of either including or excluding informative sparse levels, and including uninformative sparse levels. A simulation approach was undertaken as, described in section 6.3, to investigate these problems.

6.3 Simulation study investigating the effect of sparse levels on the results of multilevel modelling

6.3.1 The simulation hierarchies

Simulation studies are invaluable tools to assess the impact of assumption violation, particularly in fields where the "true" situation can never be ascertained. In this instance hierarchical datasets can be simulated with known variance structures and the average number of responses per household can be varied. Hierarchical models can be fitted to these datasets, and since the true values of coefficients and variance components are known, an objective assessment of the effect of single response households can be made. These datasets were constructed to satisfy all of the assumptions of hierarchical models. These basic assumptions are that the errors at each level are uncorrelated with other levels, and are distributed with mean zero (assuming that all the values of the explanatory variables are known). The level-one errors should follow a normal distribution with constant variance. Higher level errors should follow a multivariate normal distribution with constant covariance matrix.

There were two steps to this simulation process. Firstly, the hierarchy was specified. This provided information about which group each individual belonged to at

Table 6.3: Breakdown of numbers of household responses

	No. of responses	1	2	3	4	5 or more
CHSNS	**Frequency**	9,035	761	28	3	0
Poisson+1 ($\lambda = 0.084054$)	**Expected Frequency**	9,035	759	32	1	0

each level. To imitate the CHSNS study, a three-level hierarchy was chosen, with simulated ward-sized groups at the highest level, then simulated households, and finally individuals.

This was constructed as follows. The number of individuals per household was simulated using a Poisson distribution. This distribution was chosen for three reasons. Firstly, the Poisson distribution, being a non-negative discrete distribution, is a reasonable choice to model the number of individuals per household. Secondly, the numbers of responses per household in the CHSNS dataset closely followed a Poisson distribution shifted by 1, as demonstrated in table 6.3. If the number of household responses follows a distribution called X, then $X - 1$ is distributed as a Poisson variable with mean 0.084054. The value 0.084054 comes from the average number of respondents per household being 1.084054. Thirdly, the Poisson distribution parameter, λ, has an intuitive interpretation in this context, since $\lambda + 1$ can be thought of as the average number of respondents per household, μ. This was varied between 1.05 and 6. The number of households simulated was varied with μ so that the expected value of the total number of individuals was 10,000. Finally, each household was randomly allocated to one of 30 wards (chosen as a convenient number to imitate the 36 1991 electoral wards). This resulted in a hierarchy with 30 wards at the highest level, variable numbers of households at the middle level and approximately 10,000 individuals at the lowest level. Low values of μ correspond with low numbers of responses per household, which correspond to large numbers of single response households. Conversely, high values of μ correspond to low numbers of single response households.

The second step involves specifying the variance structure of the hierarchy. Area, household, and individual random effects were drawn from normal distributions. Four different approaches were taken, called A, B, C and D and are summarised in table 6.4. For all four approaches 30 observations were drawn from $N(0, 0.5)$ (to represent ward random effects). In scenario A, the household-level variance was set to be equivalent in size to the individual level. While such a situation is not the norm in area-effects literature, this scenario investigates the effect of sparseness when the sparse level has a large variance component. In this scenario, household random ef-

Table 6.4: Summary of simulation scenarios

	Variance components			
Scenario	**Area**	**Household**	**Individual**	**Sample size**
A	10	10	0.5	10,000
B	20	1.5	0.5	10,000
C	20	0	0.5	10,000
D	20	1.5	0.5	1,000

fects were drawn from $N(0, 10)$, and individual random effects were also drawn from $N(0, 10)$. Secondly, the household-level variance component was set to be much smaller than the individual-level variance component. This is much more similar to the observed variance component sizes for the household level from the literature. Here the household-level random effects were drawn from $N(0, 1.5)$, while the individual-level random effects were drawn from $N(0, 20)$. This is scenario matches the CHSNS dataset in terms of sample size and variance component magnitudes. This will be referred to as scenario B. The third simulation was the same as the second except the household variance component was set to zero. This simulation will be referred to as scenario C. This simulation investigates the effect of including a level that is unrelated to the outcome. Since variance components are constrained to be non-negative, the minimum value they can attain is zero. However, even an uninformative level in a hierarchy is likely to have some proportion of the total variance (inappropriately) attributed to it, merely by chance. To date, no one has investigated how large this proportion is likely to be. The results of this work can provide an alternative method of assessing variance component magnitudes. Estimating the variance component for an uninformative level provides a baseline against which other variance components can be compared. Finally, another simulation will employ the same variance structure as scenario B (imitating the CHSNS dataset), but with a smaller average sample size of 1,000. This scenario investigates the effect of a sparse level combined with a small sample size. This will be called scenario D.

6.3.2 The simulation models fitted

Once the dataset itself was simulated, six models were fitted to it. Firstly, a three-level null model was fitted to the data, as in equation 6.2.

$$MH_{ijk} = \alpha_0 + \upsilon_k + \tau_j + \varepsilon_i \tag{6.2}$$

The null model is the simplest of models. The outcome variable is referenced as MH indicating that this outcome could be a mental health measure. Here υ_k references

the k^{th} area error, τ_j references the j^{th} household error, and ε_i references the i^{th} individual error. The only parameters estimated from this were the mean of the response and the variance components. This model was used to assess the precision of variance component estimation as it relates to different sparseness conditions. Here the estimation of the variance components under various sparseness conditions was investigated. Secondly, a null model with the household level excluded was fitted to investigate the effect of excluding an intermediate level, as in equation 6.3.

$$MH_{ik} = \alpha_0 + \upsilon_k + \varepsilon_i \tag{6.3}$$

There are no j subscripts in this model since the household level is excluded. Thirdly, a three-level model with an area-level explanatory variable was fitted, so that the impact on coefficient estimation could be examined, as in equation 6.4.

$$MH_{ijk} = \alpha_0 + \beta_1 \times X_k + \upsilon_k + \tau_j + \varepsilon_i \tag{6.4}$$

Here X_k references the value of the fixed effect for area k, while β_1 references the coefficient of the fixed effect. In this simulation, the β_1 coefficient is set to be five. Area level effects (X_ks) were drawn from a N(0,1) distribution. The fourth model was the same as the third model but with the area-level variable replaced with a household-level variable, as in equation 6.5.

$$MH_{ijk} = \alpha_0 + \beta_2 \times X_j + \upsilon_k + \tau_j + \varepsilon_i \tag{6.5}$$

Here β_2 references the coefficient for a household level variable (the different subscript is used to distinguish it from β_1 above). Again, the X_js are drawn from a N(0,1) distribution.

The fifth model again includes the area-level fixed effect, but with the household level excluded. This investigates the effect of excluding an intermediate level on higher-level fixed effect estimation. The model is the same as equation 6.3 but with the area-level fixed effect included, as given in equation 6.6.

$$MH_{ik} = \alpha_0 + \beta_3 \times X_k + \upsilon_k + \varepsilon_i \tag{6.6}$$

The sixth and final model, replaces the area-level fixed effect with the household-level fixed effect. Since household is not included in this model, the household-level fixed effect is included as an individual level fixed effect, as in equation 6.7.

$$MH_{ik} = \alpha_0 + \beta_4 \times X_i + \upsilon_k + \varepsilon_i \tag{6.7}$$

Only results regarding fixed effect estimation will be presented for the last four models, whereas results for the first two will be more comprehensive. These simulations will provide evidence regarding the effect of both including and excluding a sparse level of information on the results of a multilevel analysis.

6.3.3 The technical details of the simulation procedure

As mentioned earlier, the average number of individuals per household was varied between 1.05 and 6 in increments of 0.05. For each average number of individuals per household 200 hierarchies were simulated. This resulted in 20,000 hierarchies being simulated for each of the four simulations (A, B ,C and D). Creating, storing, managing and analysing such a large number of hierarchies was no trivial task and obviously required heavily automated procedures. Simulating such a large number of hierarchies was only possible through the use of Condor (Litzkow et al., 1988) a Cardiff University wide parallel computing network. The basic principle of the Condor system is that at any one time there are a large number of computers across Cardiff university that are not being used to their full potential. Condor can send tasks to these computers which run in the background without noticeably slowing down the computer for the owner. A number of schools have signed up to this system as well as many of the open access computer laboratories. The Cardiff Condor computer pool is the second largest pool of its kind in Europe.

Before any tasks could be sent through Condor, the R computing environment needed to be distributed to the computers in the pool. The open-access, non-licensed nature of R facilitated this process greatly. Since Condor harnesses the power of ordinary desktop computers University wide, and not a supercomputer, then each task (or job) sent through it has to be small enough to run in a small amount of time (approximately one hour). If the owner of the computer being used by the Condor pool logs out or turns off the computer in the middle of a job, that job is resubmitted and needs to begin again from the start. Therefore large jobs that would take days to run are very inefficient, and need to be split up into a large number of smaller jobs. Small, self-contained tasks, incorporating various functions and programs necessary for model fitting were written in text files. Thousands of such text files were written (automated in R), and submitted to Condor. At any one time, up to 500 computers were running programs for this study. The text files specified the hierarchy for a given simulation, and then fitted a hierarchical model to it. Instructions were provided on which parameters of interest needed to be recorded and returned to the submission computer. For each hierarchy simulated, the hierarchy itself was returned as well as various model parameters in the form of an R workspace. Each of these workspaces was collated into one large file for analysis. For this simulation, the time taken to

create a simulated hierarchy, apply the six models of interest to that hierarchy, and extract the required information from that model took, on average, six minutes for simulations A to C. Simulation D was quicker being based on a smaller total sample size, but still took on average 2-3 minutes for each. This equates to over nine months of computing time if this were to be performed on a single machine. Utilising Condor split this computing time between many machines reducing the total time to less than two weeks (jobs were fed into the system manually, meaning that if Condor finished a set of simulations overnight, it would lie idle until more were added). A benefit of this drastically reduced computing time was that if a mistake was made in any part of the algorithm, or even if extra information was required in order to further investigate the results produced by preliminary analysis, it was much easier to fix the mistake, or extract the extra information.

The information was returned to the submit machine in batches. These batches contained the results from a number of different hierarchies, typically ten. Information on both the selected model information and the simulated hierarchy itself (for the purposes of manually checking that the results were correct) were returned. For each scenario then, the information of interest was contained in 10,000 folders. Obviously, it would have been too time consuming to manually collate all of this information. This process was automated in R. An extraction algorithm was created which checked each of the 10,000 folders for information (some jobs were rejected by Condor resulting in no information in a given folder). If a given folder contained information, that information was extracted and collated into a large dataset, which was output. Without this automated information extraction algorithm, Condor would have been of little benefit. The next section examines the results of this process.

6.4 Results

6.4.1 Scenario A

Three-Level Null Model

Firstly the three-level null model is examined for the case where the household and individual levels contribute equal amounts of variation to the outcome. The variance contributions from the individual, household and area levels are 10, 10 and 0.5, respectively. The relationship between the variance components produced and sparseness is displayed in figure 6.1. Each sparseness level (from 1.05 to 6 in steps of 0.05) is represented by a boxplot. The true values for the variation attributable to each level are depicted by solid horizontal lines in each plot. The x-axis represents the average number of individuals per household. As this value is increased the number of single response households decreases. Figure 6.1 shows how the estimates of the variance

components (for each of the three levels) change as the average number of responses per household changes. The ward-level variance component is unbiasedly estimated with constant variance for all levels of sparseness. The household and individual-level variance components (plotted on the same scale for comparison) are unbiasedly estimated for all levels of sparseness. There is evidence that the household-level variance component is more variable than the individual-level variance component. The individual-level component is estimated less precisely when the number of individuals per household is small.

This increased variance component variability at high sparseness levels has consequences for the ICC coefficients. Figure 6.2 shows the ICC coefficients for each of the three levels in the hierarchy. Horizontal lines indicate the true ICC coefficients for each level. The ward-level ICC appears to be estimated without bias, and with constant variance for all sparseness levels. For both the household and individual levels the ICCs are unbiased, however they are more variable when the average number of individuals per household is small. This is confirmed by the summary information in table 6.6, with reported variances three times greater when the average number of individuals per household is between 1.05 and 2, than when the average is greater than 2.

Model fit is examined using the AIC (introduced in section 5.2). The relationship between the average number of individuals per household and the resultant three-level null model AIC is displayed in figure 6.3. This figure shows that when the average number of individuals per household is low, the model fit is worse.

Since the total number of individuals was not set to be 10,000, but varied around 10,000, it is useful to examine the impact of these different sample sizes on the estimated variance components. This is plotted in figure 6.4. The total number of individuals is centred around 10,000, but extends as low as 9,687 and as high as 10,346. Such relatively small differences in sample size would not be expected to have a large impact on the variance component estimation, and indeed that is the case with the variance components for all three levels exhibiting nothing more than random variation. The individual-level variance component is much less variable than the household level (note that the household and individual-level variance components are plotted on the same axes).

Two-level Null Model

The model which excludes the household level is examined in figure 6.5. The ward-level variance component is not hugely affected by the exclusion of the household level, except when the average number of individuals per household is large (when the average is between 5-05 and 6 this variance component is 0.66 on average. There is evidence that the ward-level variance component is estimated more accurately with

Figure 6.1: Relationship between the variance components and the average number of individuals per household for three-level null model in scenario A

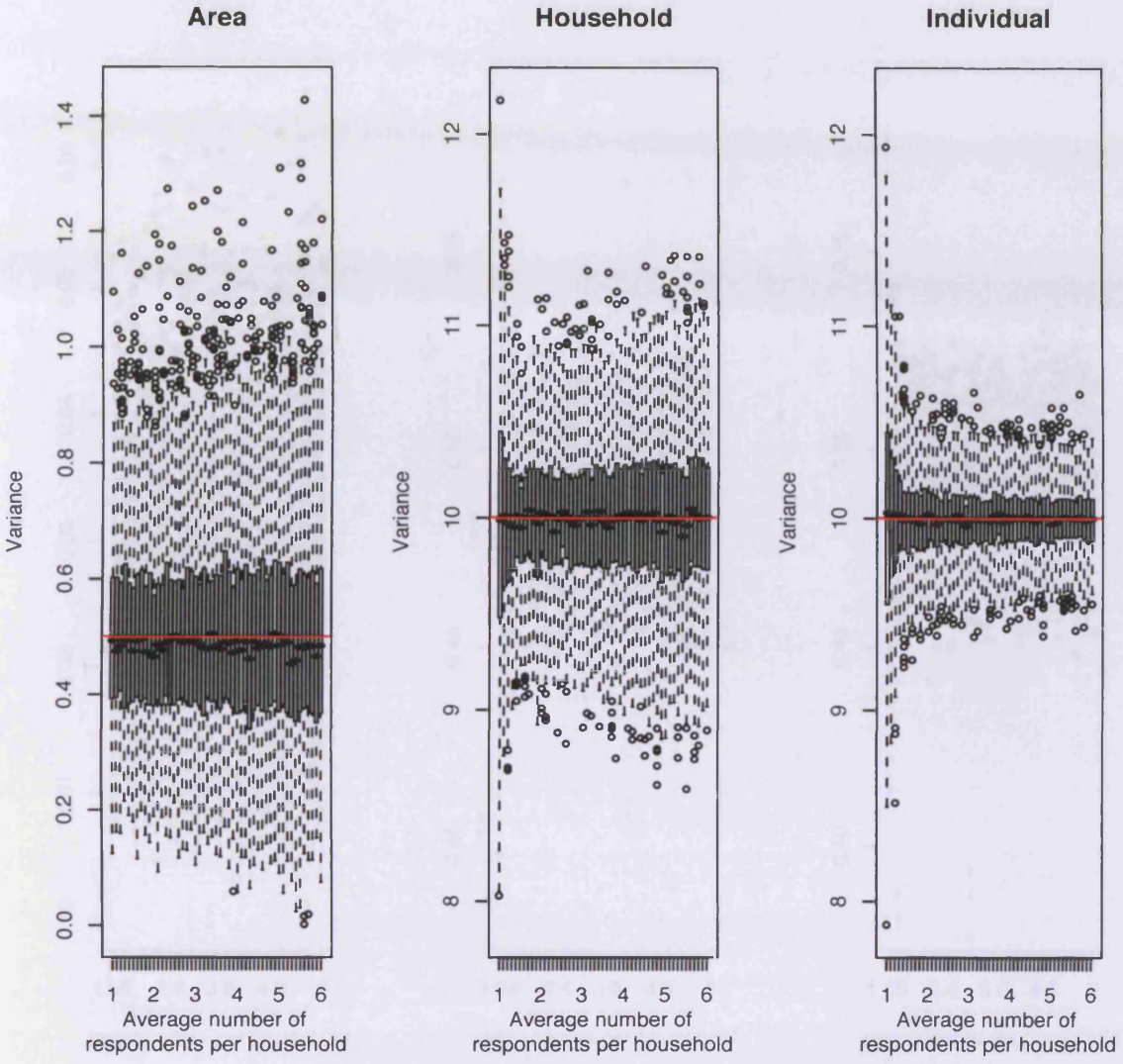


Table 6.5: Summary information for figure 6.1

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area mean	0.50	0.50	0.50	0.50	0.50
Area variance	0.02	0.03	0.03	0.03	0.03
Household mean	9.99	10.00	10.00	10.00	10.00
Household variance	0.17	0.11	0.13	0.14	0.16
Individual mean	10.00	10.00	10.00	10.00	10.00
Indivdual variance	0.11	0.03	0.03	0.03	0.02

Figure 6.2: Relationship between the ICC coefficients and the average number of individuals per household for three-level null model in scenario A

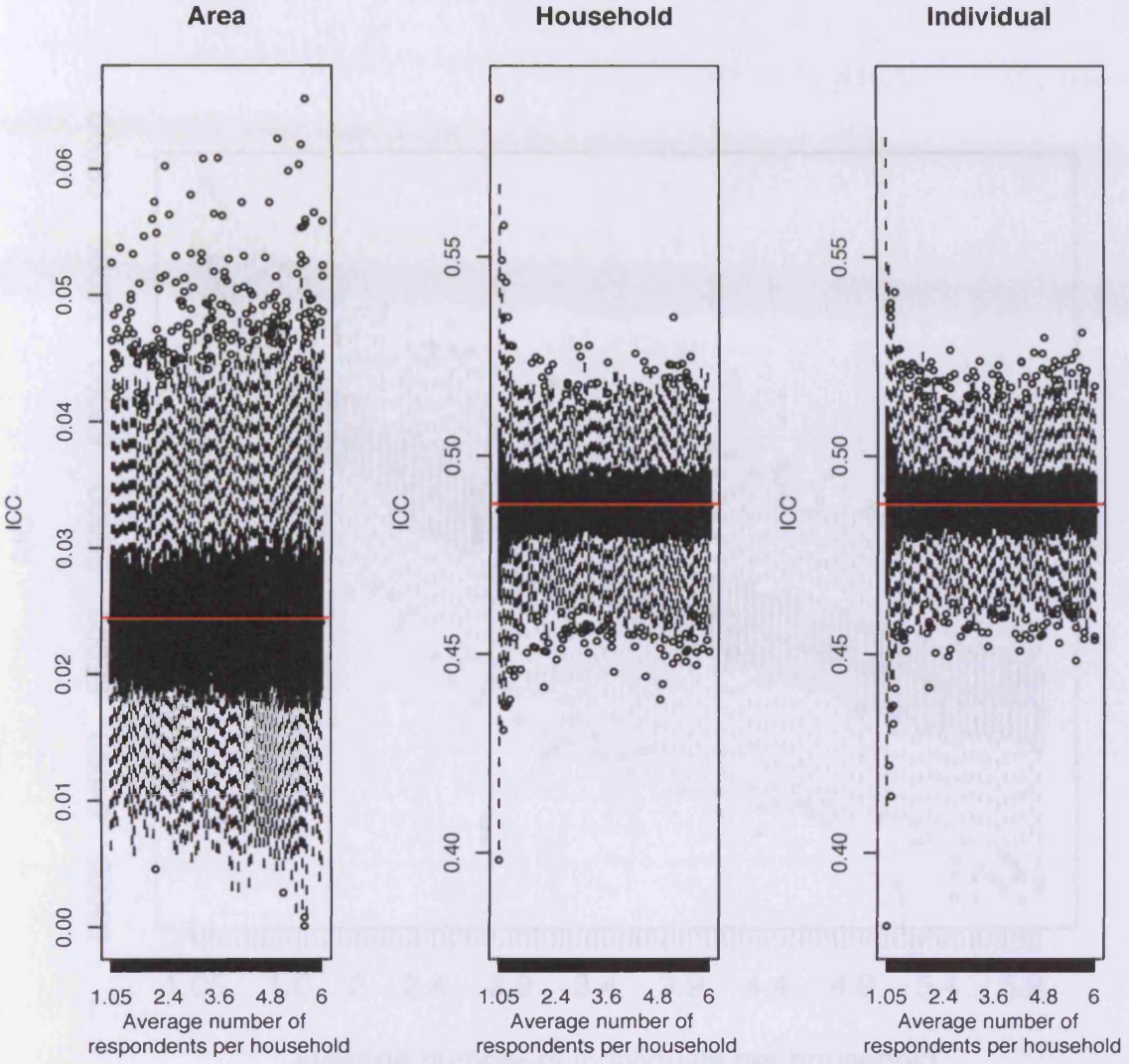


Table 6.6: Summary information for figure 6.2

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area mean	0.0244	0.0245	0.0244	0.0245	0.0242
Area variance	0.0001	0.0001	0.0001	0.0001	0.0001
Household mean	0.4874	0.4876	0.4878	0.4876	0.4877
Household variance	0.0003	0.0001	0.0001	0.0001	0.0001
Individual mean	0.4882	0.4879	0.4878	0.4878	0.4881
Individual variance	0.0003	0.0001	0.0001	0.0001	0.0001

Figure 6.3: Relationship between the average number of individuals per household and the model fit for the three-level null model in scenario A

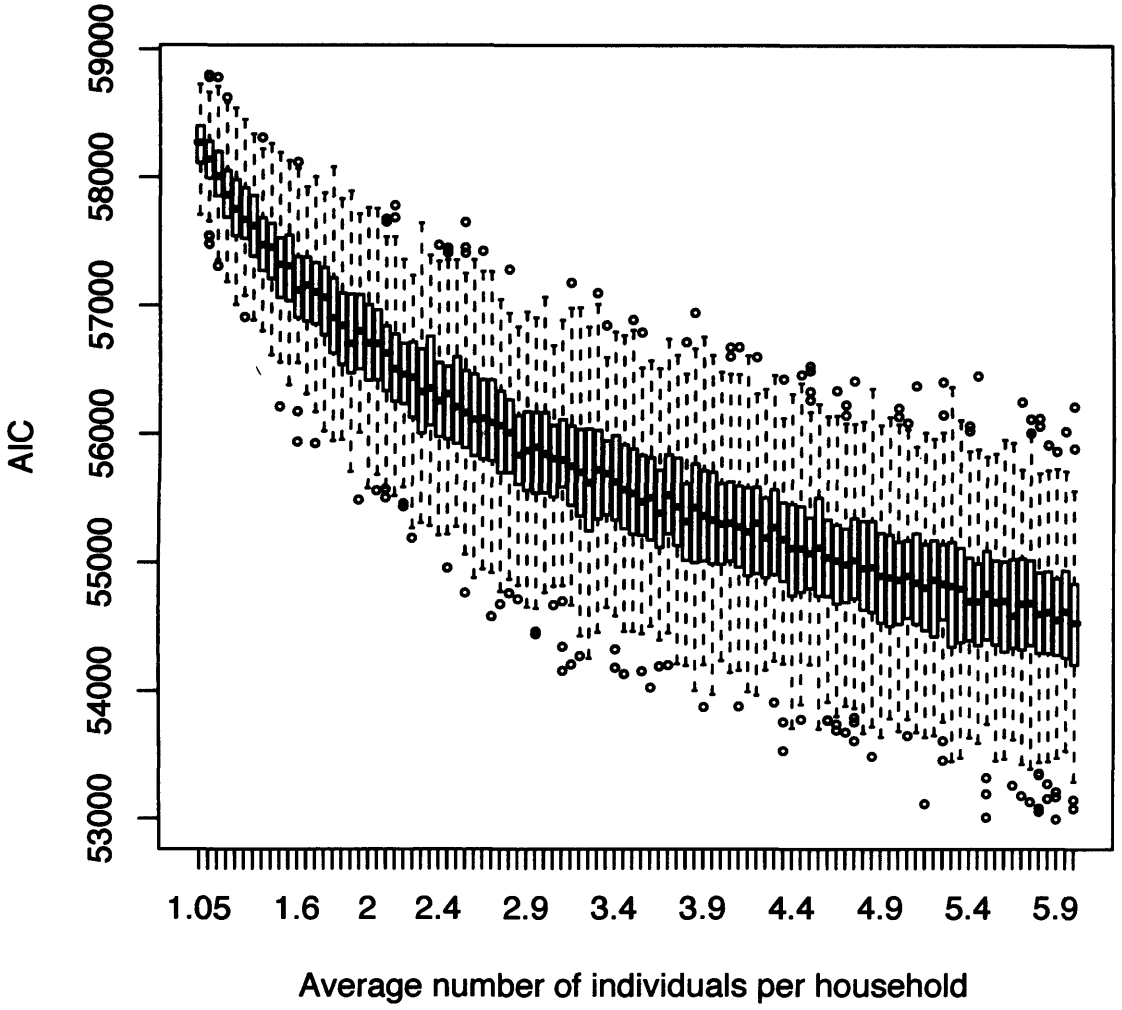


Table 6.7: Summary information for figure 6.3

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
AIC mean	57,511	56,197	55,543	55,086	54,714
AIC variance	116,745	191,990	220,983	233,724	249,972

Figure 6.4: Relationship between the variance components and the total number of individuals for three-level null model in scenario A

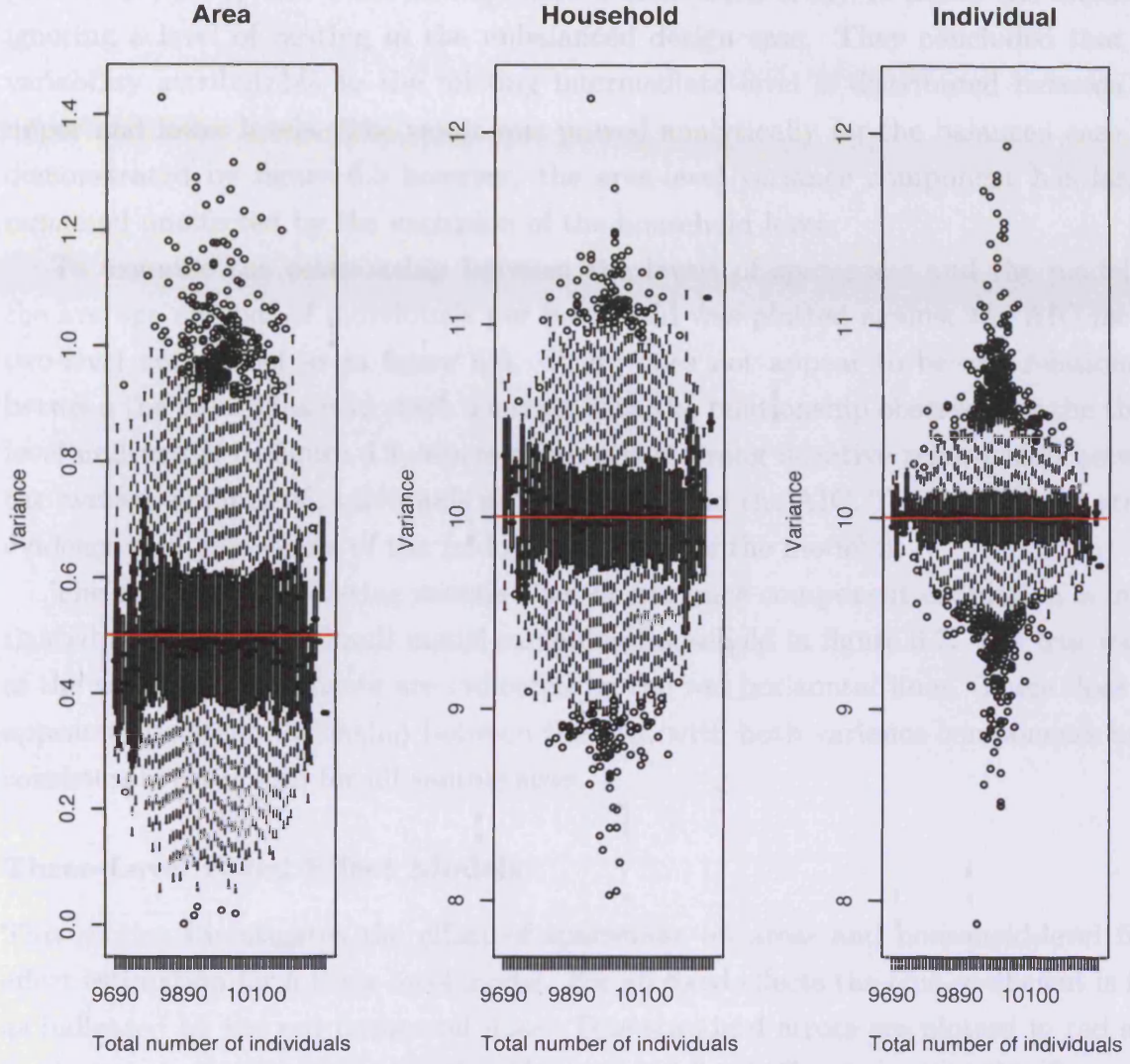


Table 6.8: Summary information for figure 6.4

Rounded number of individuals	9700	9800	9900	10000	10100	10200	10300
Area mean	0.46	0.52	0.50	0.50	0.50	0.50	0.49
Area variance	0.04	0.03	0.03	0.03	0.03	0.03	0.02
Household mean	10.01	10.02	9.99	10.00	10.01	9.99	10.07
Household variance	0.15	0.15	0.14	0.15	0.13	0.15	0.17
Individual mean	10.02	10.00	10.00	10.00	10.00	9.99	9.93
Individual variance	0.02	0.03	0.03	0.06	0.03	0.03	0.02

high levels of sparseness. The variability attributed to the missing level is attributed to the individual level in this situation. This corresponds only partially with a different study which investigated the consequences of ignoring a level in a multilevel analysis (Moerbeek, 2004). Here the authors used a simulation study to assess the impact of ignoring a level of nesting in the unbalanced design case. They concluded that the variability attributable to the missing intermediate level is distributed between the upper and lower levels. The result was proved analytically for the balanced case. As demonstrated by figure 6.5 however, the area-level variance component has largely remained unaffected by the exclusion of the household level.

To examine the relationship between the levels of sparseness and the model fit, the average number of individuals per household was plotted against the AIC for the two-level null model as in figure 6.6. There does not appear to be any relationship between the two. This is in stark contrast with the relationship observed for the three-level null model in figure 6.3, where there was a strong negative relationship between the average number of individuals per household and the AIC. This constitutes strong evidence that sparseness of the middle level reduces the model fit.

The effect of the differing sample sizes on variance component estimation is investigated for the two-level null model excluding household in figure 6.7. The true values of the variance components are indicated by the red horizontal lines. There does not appear to be any relationship between the two, with both variance components being consistently estimated for all sample sizes.

Three-Level Fixed Effect Models

This section investigates the effect of sparseness on area- and household-level fixed effect estimation for a three-level model. For all fixed effects the true coefficient is five, as indicated by the red horizontal lines. True standard errors are plotted in red also, on the standard error plots. Firstly, the area-level fixed effect is examined in figure 6.8. The fixed effect itself is estimated without bias for all levels of sparseness as can be seen here (and in table 6.12). When the average number of individuals per household is low there appears to be a slight underestimate of the standard error of this fixed effect, and when the average number per household is large, an overestimate. Neither is a large effect however.

Next, the household-level fixed effect is examined. Figure 6.9 shows the effect of sparsity on household-level fixed estimation for scenario A. Figure 6.9.(a) shows that the household-level fixed effect is unbiasedly estimated, although there is evidence that larger numbers of individuals increases the standard error of those estimates. This is confirmed in figure 6.9.(b) and table 6.13. When the sparseness is quite extreme (less than 1.5 respondents per household) the standard errors for the household-level fixed

Figure 6.5: Relationship between the variance components and the average number of individuals per household for the two-level null model, excluding household in scenario A

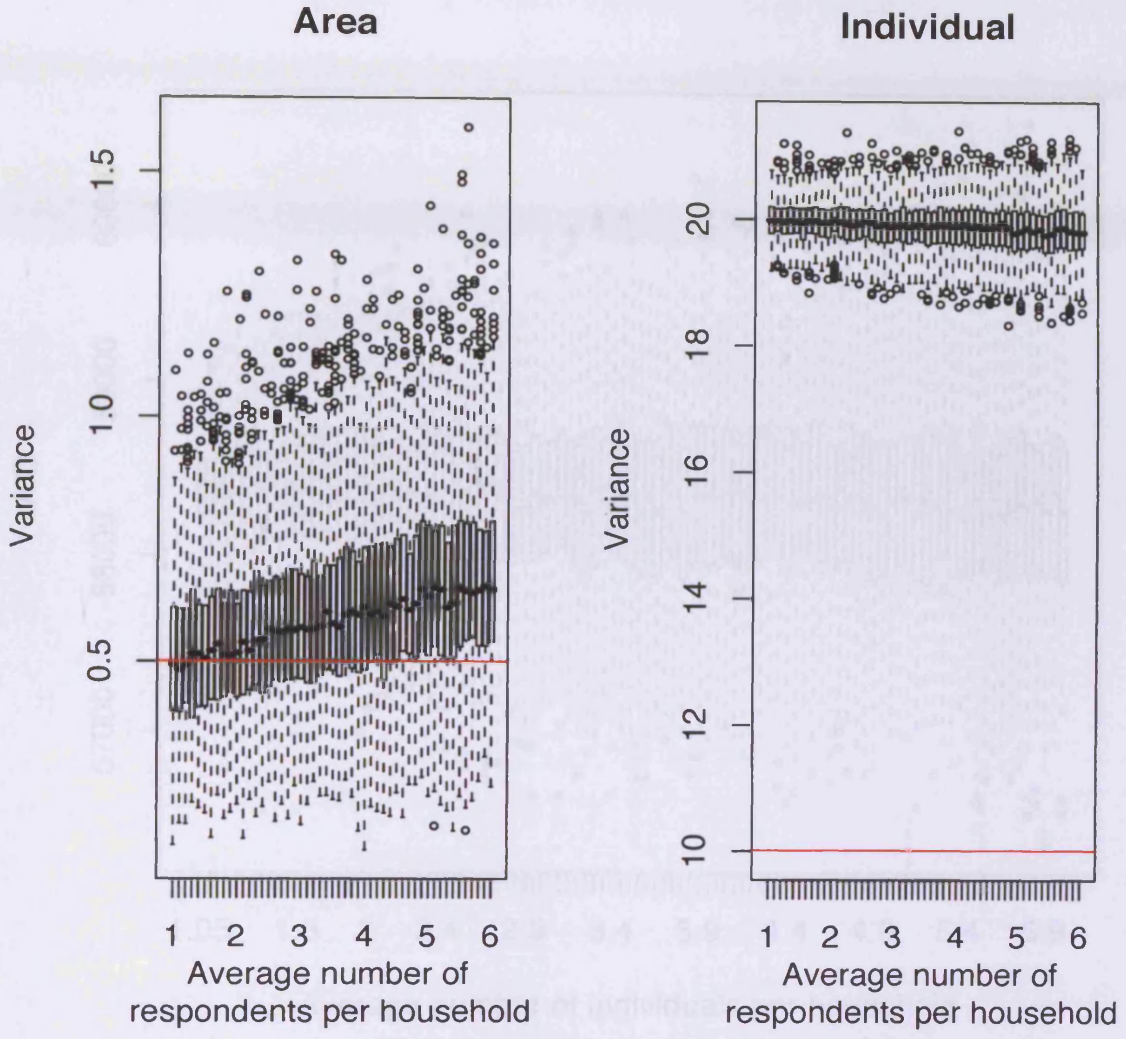


Table 6.9: Summary information for figure 6.5

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area mean	0.52	0.57	0.60	0.64	0.66
Area variance	0.02	0.03	0.03	0.03	0.04
Individual mean	19.97	19.93	19.90	19.87	19.84
Individual variance	0.11	0.03	0.03	0.03	0.02

Figure 6.6: Relationship between the average number of individuals per household and model fit for the two-level null model, excluding household in scenario A

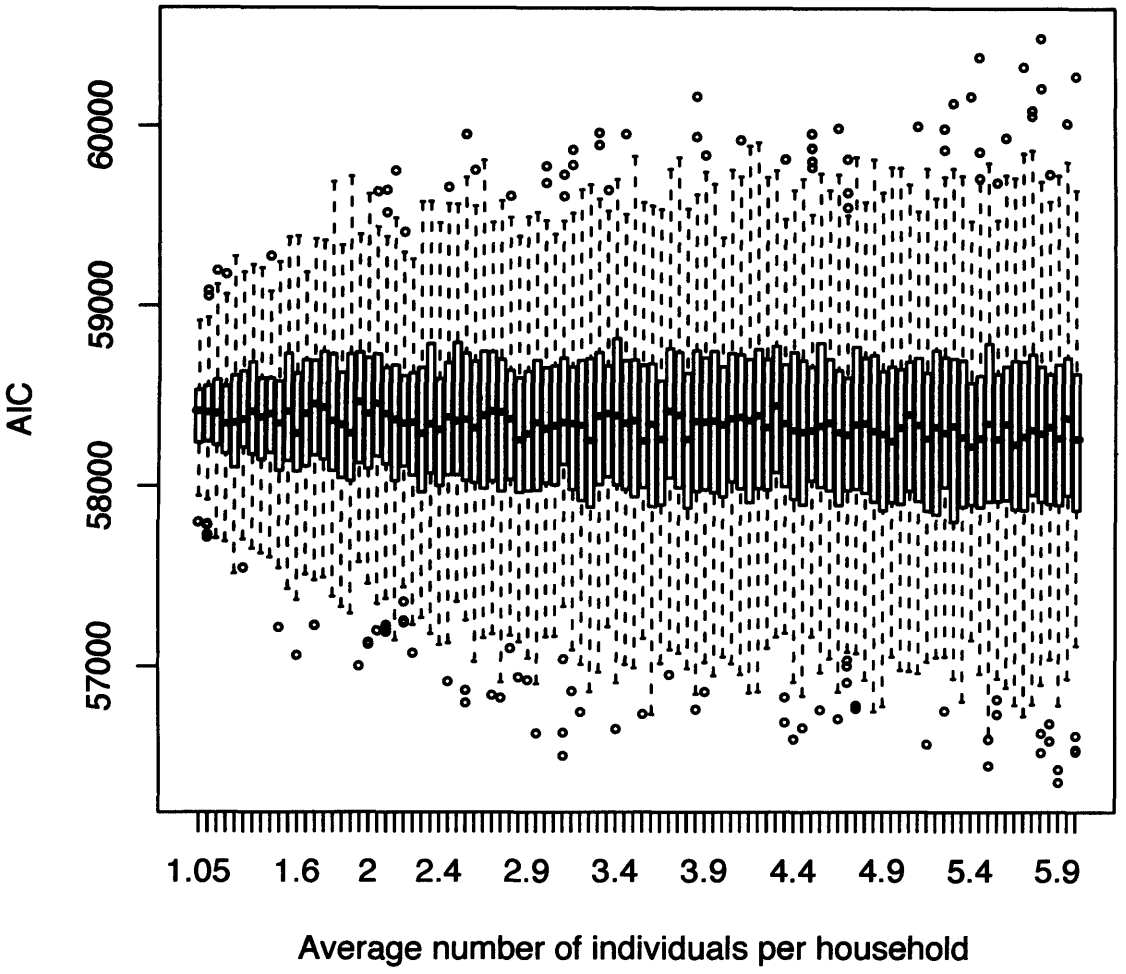


Table 6.10: Summary information for figure 6.6

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
AIC mean	58391	58356	58342	58333	58300
AIC variance	120075	231543	276095	299278	324429

Figure 6.7: Relationship between the variance components and total number of individuals for the two-level model, excluding household in scenario A

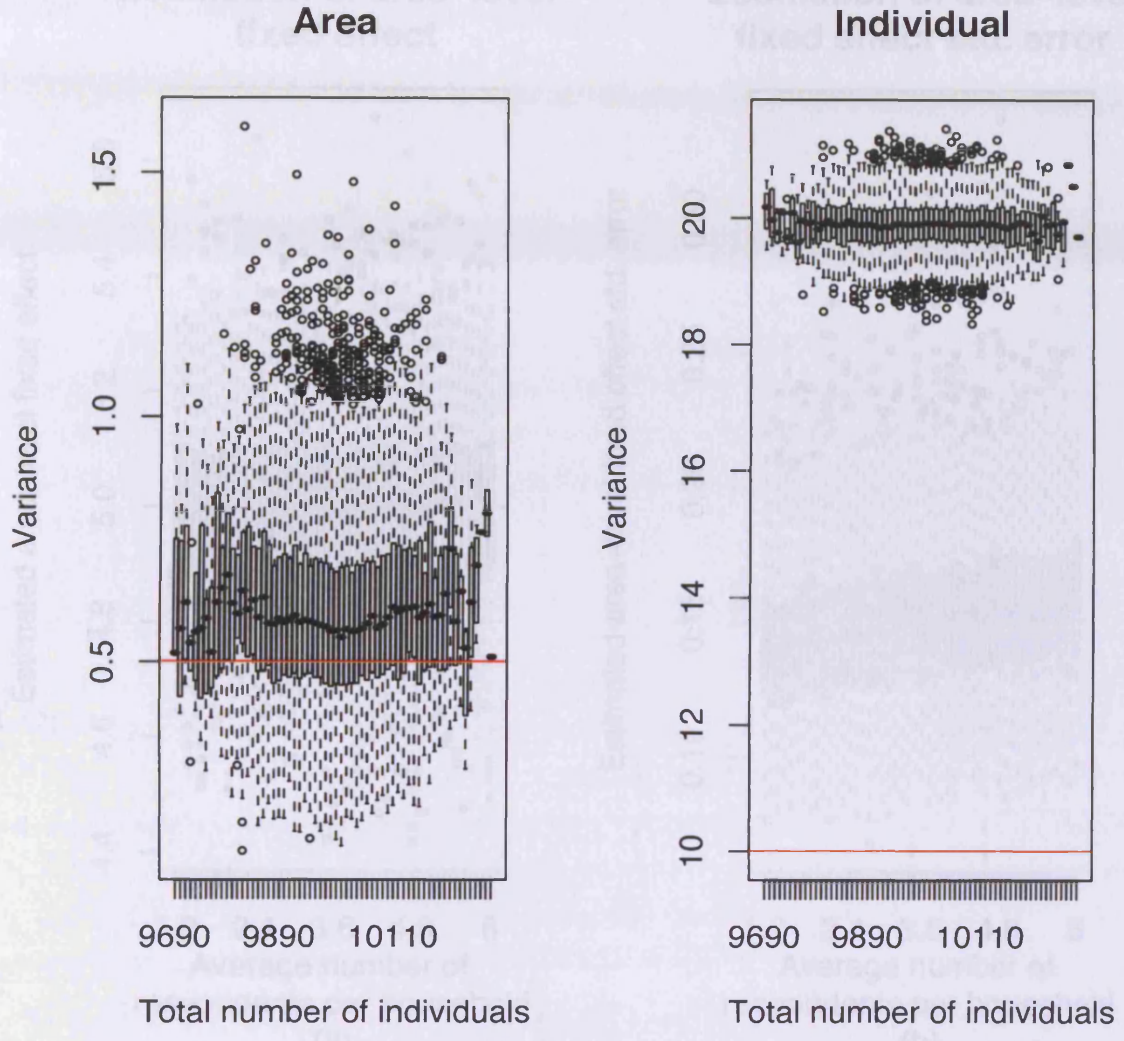


Table 6.11: Summary information for figure 6.7

Rounded number of individuals	9700	9800	9900	10000	10100	10200	10300
Area mean	0.61	0.64	0.60	0.59	0.60	0.62	0.64
Area variance	0.05	0.04	0.03	0.03	0.03	0.03	0.02
Individual mean	19.89	19.90	19.89	19.91	19.90	19.87	19.88
Individual variance	0.15	0.17	0.15	0.14	0.14	0.17	0.18

Figure 6.8: Relationship between the sparseness and area-level fixed effect estimation for the three-level model in scenario A

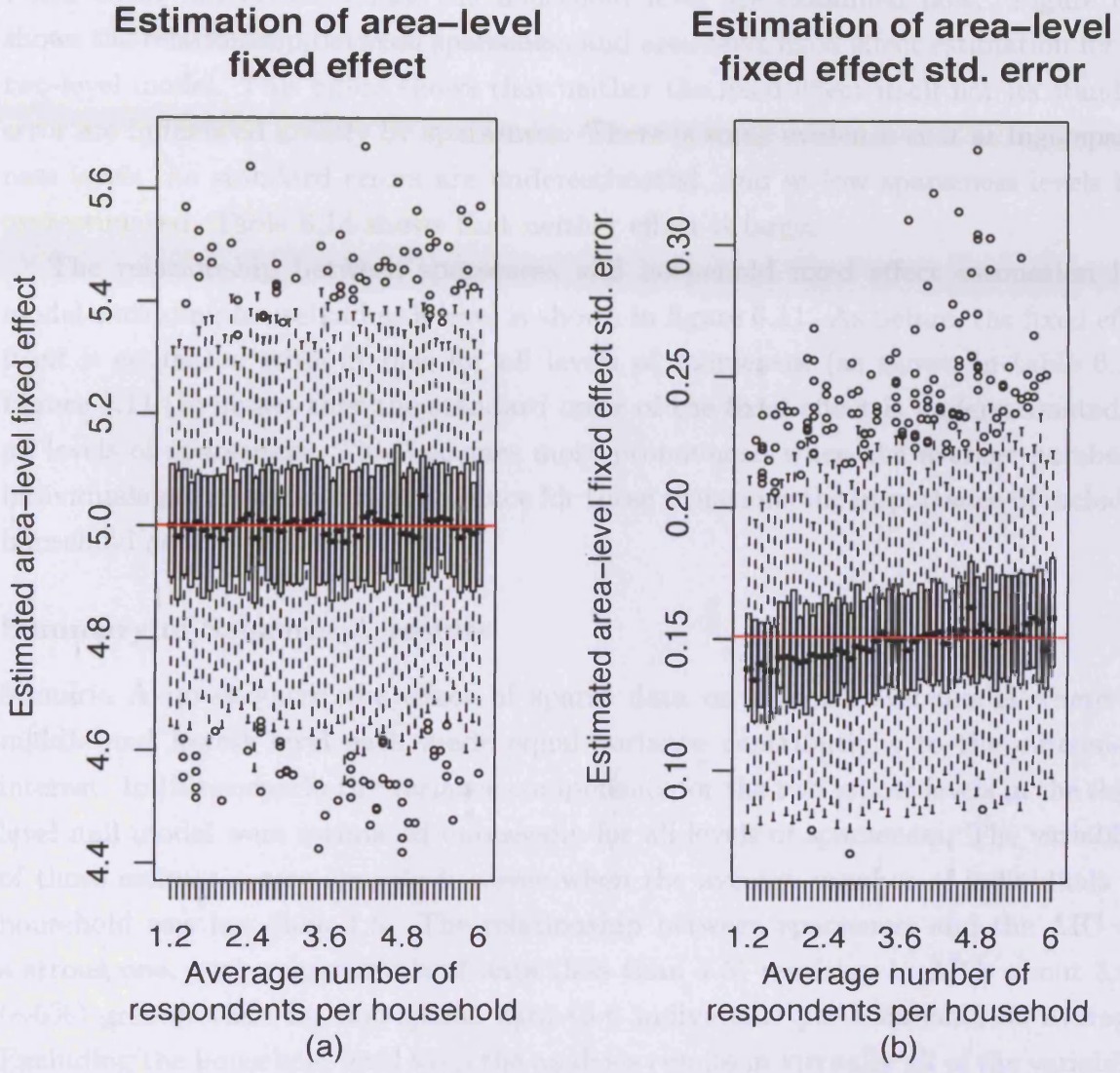


Table 6.12: Summary information for figure 6.8

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area fixed effect mean	4.99	5.00	5.00	5.00	5.00
Area fixed effect var.	0.02	0.02	0.02	0.03	0.02
Area fixed effect std. error mean	0.143	0.147	0.151	0.156	0.157
Area fixed effect std. error variance	0.001	0.001	0.001	0.001	0.001

effect are underestimated.

Two-Level Fixed Effect Models

Fixed effect models excluding the household level are examined now. Figure 6.10 shows the relationship between sparseness and area-level fixed effect estimation for the two-level model. This figure shows that neither the fixed effect itself nor its standard error are influenced greatly by sparseness. There is some evidence that at high sparseness levels the standard errors are underestimated, and at low sparseness levels it is overestimated. Table 6.14 shows that neither effect is large.

The relationship between sparseness and household fixed effect estimation in a model excluding household as a level is shown in figure 6.11. As before, the fixed effect itself is estimated without bias for all levels of sparseness (as shown in table 6.15). Figure 6.11.(b) shows that the standard error of the fixed effect is underestimated for all levels of sparseness. This becomes more pronounced when the average number of individuals per household is large, since for these situations the importance of including household as a level is large.

Summary of Scenario A results

Scenario A investigated the effect of sparse data on a three-level model where the middle and lowest level each made equal variance contributions to the outcome of interest. In this scenario the variance components for the two lowest levels in the three-level null model were estimated unbiasedly for all levels of sparseness. The variability of those estimates rose sharply however when the average number of individuals per household was less than 1.5. The relationship between sparseness and the AIC was a strong one, with sparse levels of data (less than 1.5) resulting in AICs about 3,000 ($\sim$5%) greater than for non-sparse data (5-6 individuals per household on average). Excluding the household level from the analysis results in virtually all of the variability at that level being attributed to the lower level. The higher level is estimated without bias when the number of individuals per household is low. This makes sense since in this situation it is impossible to distinguish household effects from individual effects and so the household variability is erroneously attributed to the individual level. As the number of individuals per household increases, so does the ward-level variance component. The magnitude of this overestimation is not large however. Excluding the household level results in the relationship between the average number of individuals per household and the AIC disappearing. Sparseness does not introduce bias into area- or household-level fixed effect estimates for either three-level models including household, or two-level models excluding household. Area-level fixed effect standard errors were underestimated when the sparseness was extreme (less than 1.5 individuals

Figure 6.9: Relationship between the sparseness and household-level fixed effect estimation for the three-level model in scenario A

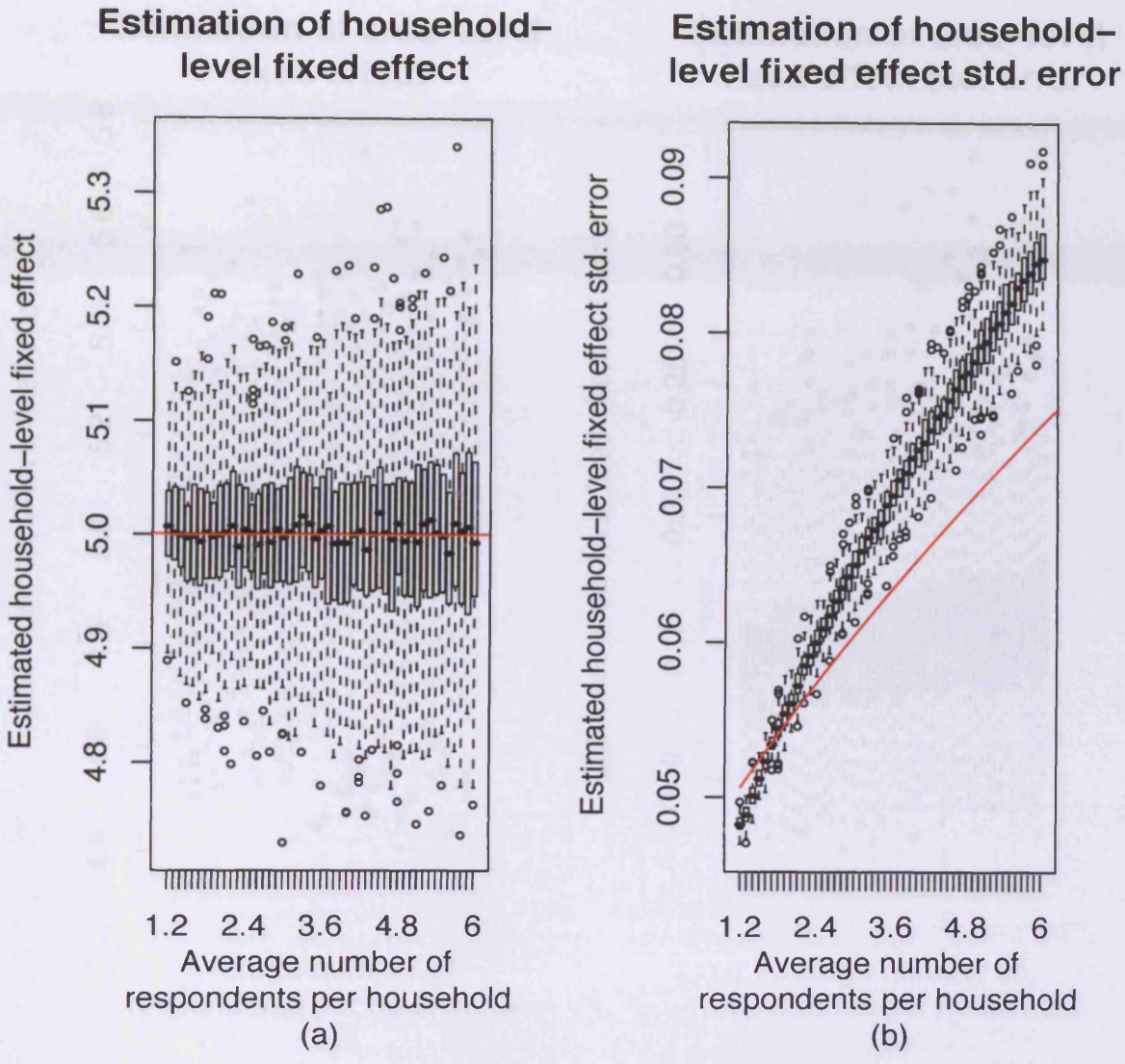


Table 6.13: Summary information for figure 6.9

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Household fixed effect mean	5.002	5.000	5.001	5.001	5.003
Household fixed effect var.	0.003	0.004	0.005	0.006	0.007
Household fixed effect std. error mean	0.053	0.061	0.069	0.076	0.082
Household fixed effect std. error variance	0.000	0.000	0.000	0.000	0.000

Figure 6.10: Relationship between the sparseness and area-level fixed effect estimation for a two-level model excluding household in scenario A

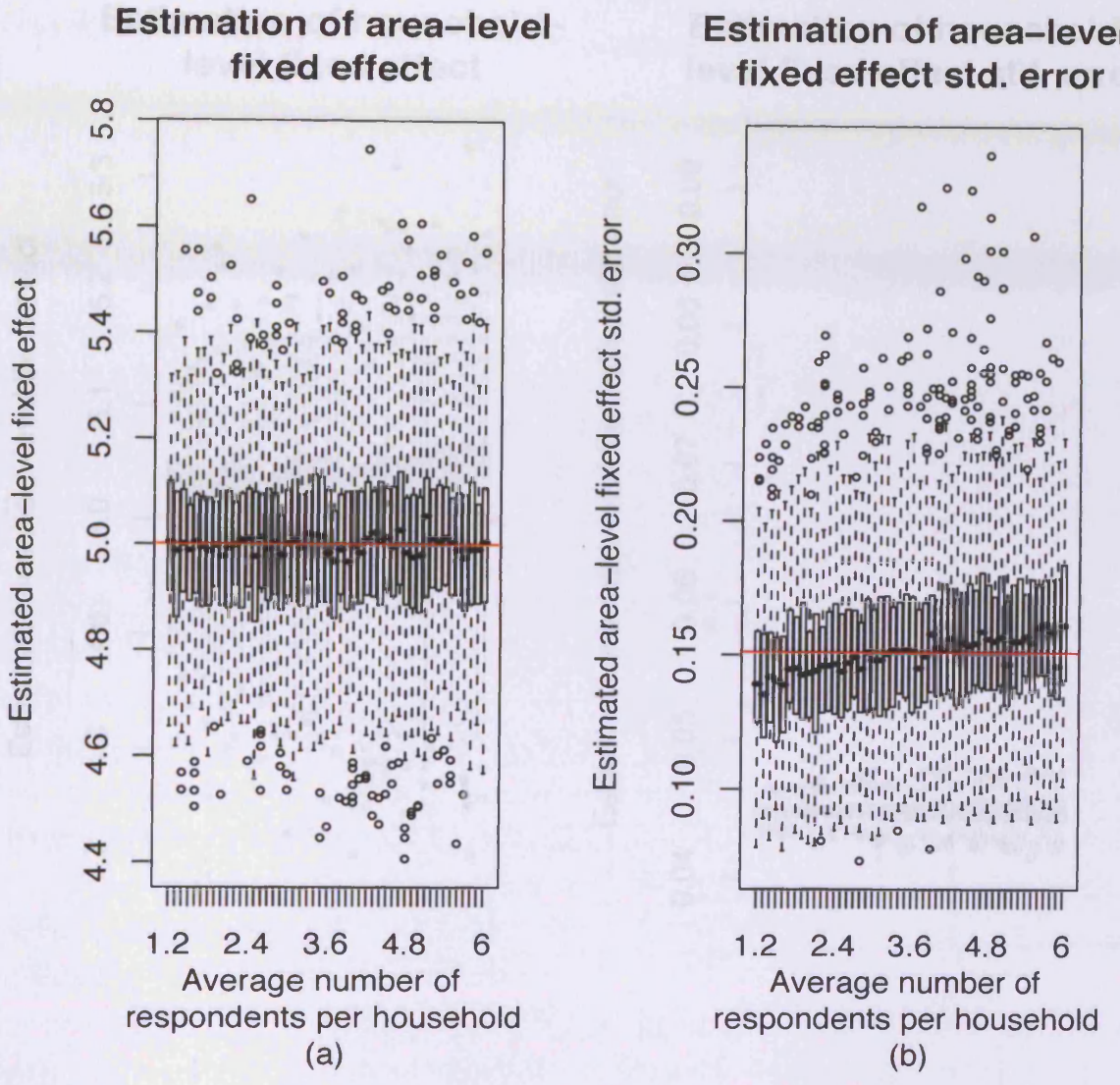


Table 6.14: Summary information for figure 6.10

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area fixed effect mean	4.994	5.001	4.999	4.999	5.003
Area fixed effect var.	0.022	0.022	0.025	0.026	0.025
Area fixed effect std. err. mean	0.144	0.148	0.153	0.158	0.159
Area fixed effect std. err. var.	0.001	0.001	0.001	0.001	0.001

Figure 6.11: Relationship between the sparseness and household fixed effect estimation for a two-level model excluding household in scenario A

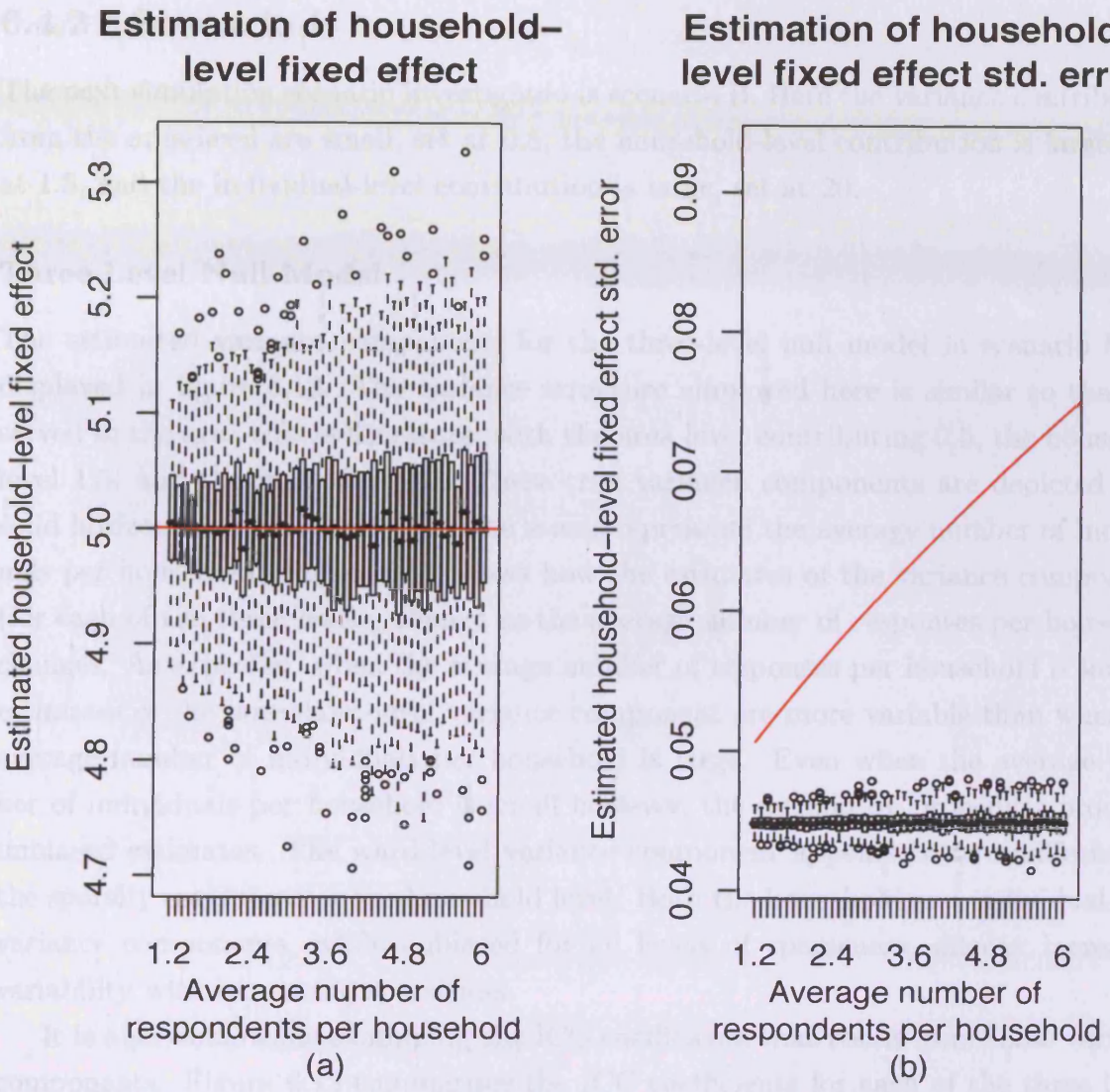


Table 6.15: Summary information for figure 6.11

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Household fixed effect mean	5.0020	5.0010	5.0010	5.0010	5.003
Household fixed effect var.	0.0030	0.0040	0.0050	0.0060	0.007
Household fixed effect std. error mean	0.0448	0.0449	0.0449	0.0449	0.045
Household fixed effect std. error var.	0.0000	0.0000	0.0000	0.0000	0.000

per household) for both three and two level models. Household-level fixed effects were underestimated at extreme sparseness levels in the three level model, and always underestimated in the two level model.

6.4.2 Scenario B

The next simulation scenario investigated is scenario B. Here the variance contribution from the area-level are small, set at 0.5, the household-level contribution is larger, set at 1.5, and the individual-level contribution is large, set at 20.

Three-Level Null Model

The estimated variance components for the three-level null model in scenario B are displayed in figure 6.12. The variance structure employed here is similar to that observed in the area effects literature, with the area level contributing 0.5, the household level 1.5, and the individual 20. These true variance components are depicted by a solid horizontal line in each plot. The x-axis represents the average number of individuals per household. Figure 6.12 shows how the estimates of the variance components (for each of the three levels) change as the average number of responses per household changes. As expected, when the average number of responses per household is low the estimates of the household-level variance component are more variable than when the average number of individuals per household is large. Even when the average number of individuals per household is small however, the estimation procedure produces unbiased estimates. The ward-level variance component appears to be unaffected by the sparsity conditions at the household level. Both the household and individual-level variance components, while unbiased for all levels of sparseness, display increasing variability with increasing sparseness.

It is also worthwhile examining the ICC coefficients that result from these variance components. Figure 6.13 summarises the ICC coefficients for each of the three levels of data. The household level ICC varies much more when the sparseness is high. The true ICC is 6.8% as indicated by the horizontal line. When the average number of individuals per household is less than 1.5, a quarter of the household-level ICCs are overestimated by at least 20% (and a quarter are underestimated by at least 20%).

To investigate the impact of sparse data on model fit the average number of individuals per household is plotted against the resultant three-level null model AIC in figure 6.14. Higher values of the AIC indicate poor model fit. Figure 6.14 shows that sparseness has only a small effect on model fit. This is quite different from the previous simulation where there was a strong relationship between these two (figure 6.3). This indicates that the effect of sparseness on model fit depends on the relative variance contribution of the sparse level. In this situation, where the household contribution is

Figure 6.12: Relationship between the variance components and the average number of individuals per household for three-level null model in scenario B

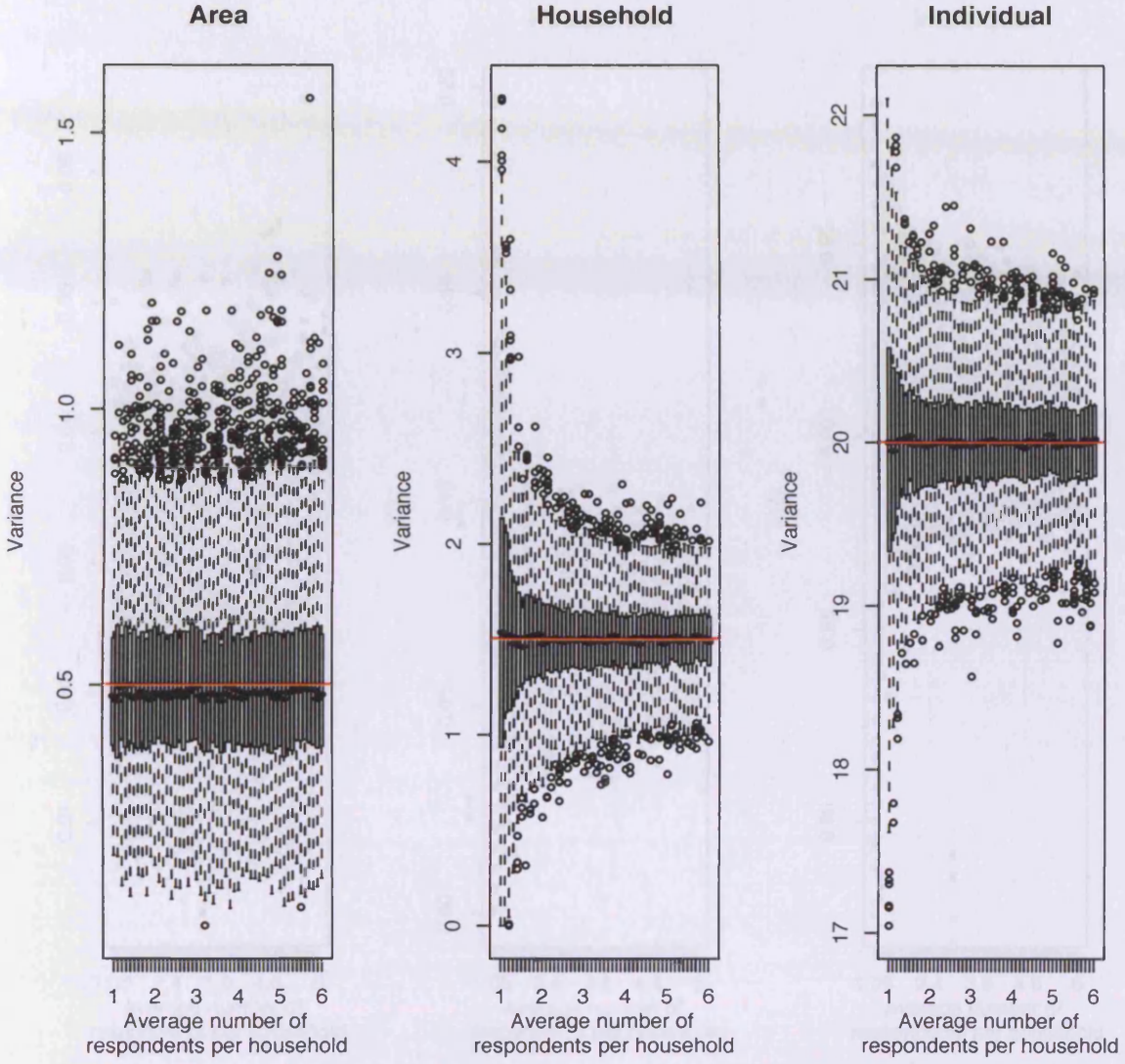


Table 6.16: Summary information for figure 6.12

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area mean	0.50	0.50	0.50	0.50	0.50
Area variance	0.02	0.02	0.02	0.02	0.02
Household mean	1.51	1.50	1.50	1.50	1.50
Household variance	0.20	0.06	0.04	0.04	0.03
Individual mean	19.99	19.99	20.00	20.00	20.00
Individual variance	0.26	0.12	0.11	0.10	0.10

Figure 6.13: Relationship between the ICC coefficients and the average number of individuals per household for three-level null model in scenario B

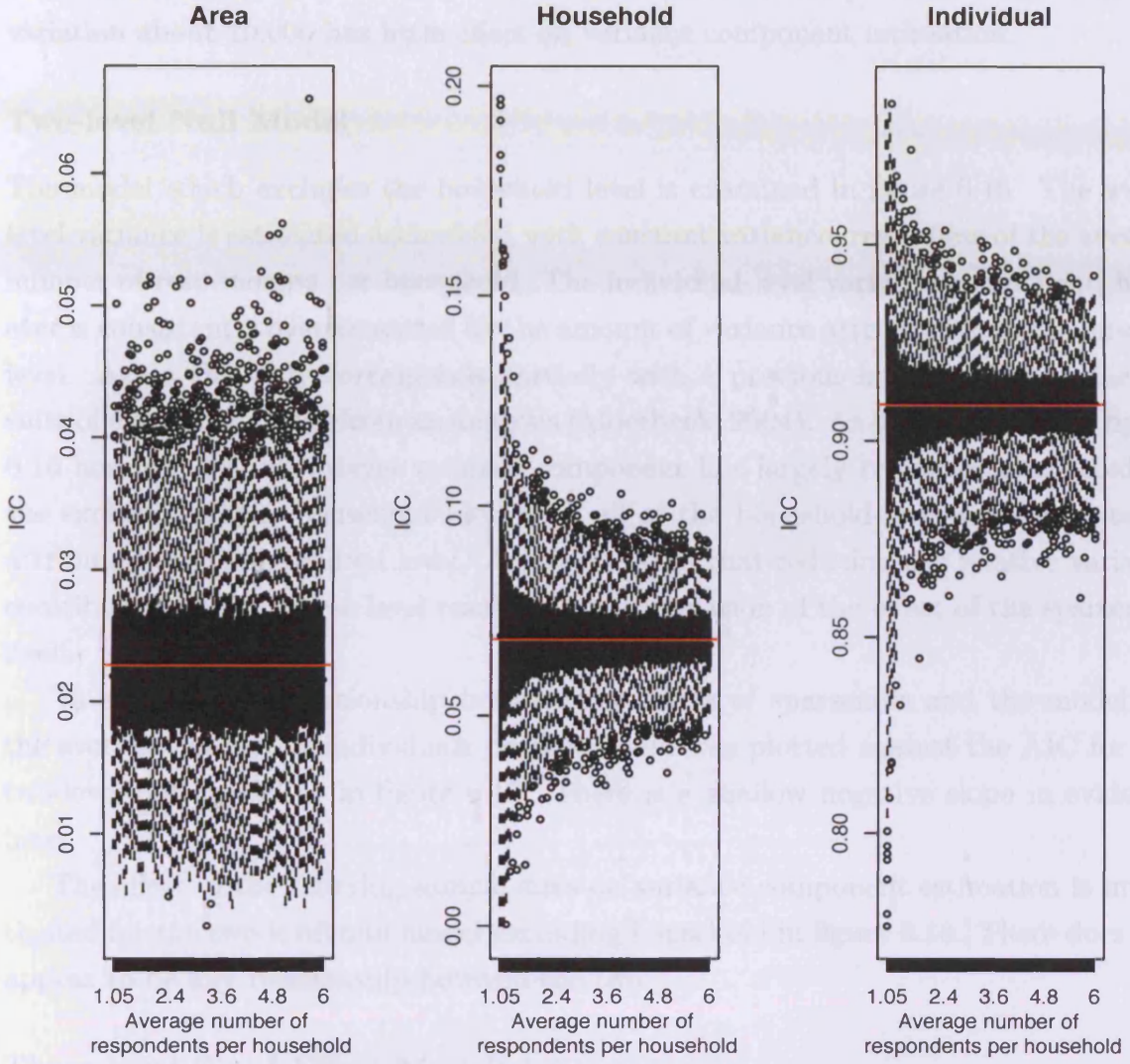


Table 6.17: Summary information for figure 6.13

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area mean	0.0227	0.0227	0.0226	0.0226	0.0227
Area var.	0.0000	0.0000	0.0000	0.0000	0.0000
Household mean	0.0685	0.0683	0.0681	0.0681	0.0683
Household var.	0.0004	0.0001	0.0001	0.0001	0.0001
Individual mean	0.9088	0.9090	0.9093	0.9092	0.9090
Individual var.	0.0004	0.0001	0.0001	0.0001	0.0001

small (6.8%), the effect of sparseness on model fit is also small.

The relationship between variance component estimation and the total number of individuals is plotted in figure 6.15. The total number of individuals is centred about 10,000, but ranges between 9,632 and 10,363. Similar to the previous simulation, this variation about 10,000 has little effect on variance component estimation.

Two-level Null Model

The model which excludes the household level is examined in figure 6.16. The ward-level variance is estimated accurately, with constant variance, regardless of the average number of respondents per household. The individual-level variance component, however is consistently overestimated by the amount of variance attributable to the missing level. Again this only corresponds partially with a previous investigation on the results of excluding a level from an analysis (Moerbeek, 2004). As demonstrated by figure 6.16 however, the ward-level variance component has largely remained unaffected by the exclusion of the household level with all of the household-level variability being attributed to the individual level. Again, it seems that reducing the relative variance contribution of the sparse level results in an attenuation of the effect of the sparseness itself.

To examine the relationship between the levels of sparseness and the model fit, the average number of individuals per household was plotted against the AIC for the two-level null model as in figure 6.17. There is a shallow negative slope in evidence here.

The effect of the differing sample sizes on variance component estimation is investigated for the two-level null model excluding household in figure 6.18. There does not appear to be any relationship between the two.

Three-level Fixed Effect Models

As described earlier, simulated fixed effects were created at both the area and household levels. These were included in separate three-level models. The true coefficient for both models was five. Firstly, the area-level fixed effect estimation is examined. Figure 6.19 shows the relationship between sparseness and the estimates of both the fixed effect and its standard error. Increasing sparseness does not appear to have any effect on the estimation of an area-level fixed effect, either in terms of bias (figure 6.19.(a)) or in terms of precision (figure 6.19.(b)). This is confirmed in table 6.23.

Next the household-level fixed effect is examined. Figure 6.20 shows the effect of sparseness on household-level (or more generally, an intermediate-level) variable estimation. Again, the fixed effect coefficient itself is unbiasedly estimated for all levels of sparseness. The standard error associated with the household-level fixed effect however

Figure 6.14: Relationship between the average number of individuals per household and the model fit for the three-level null model in scenario B

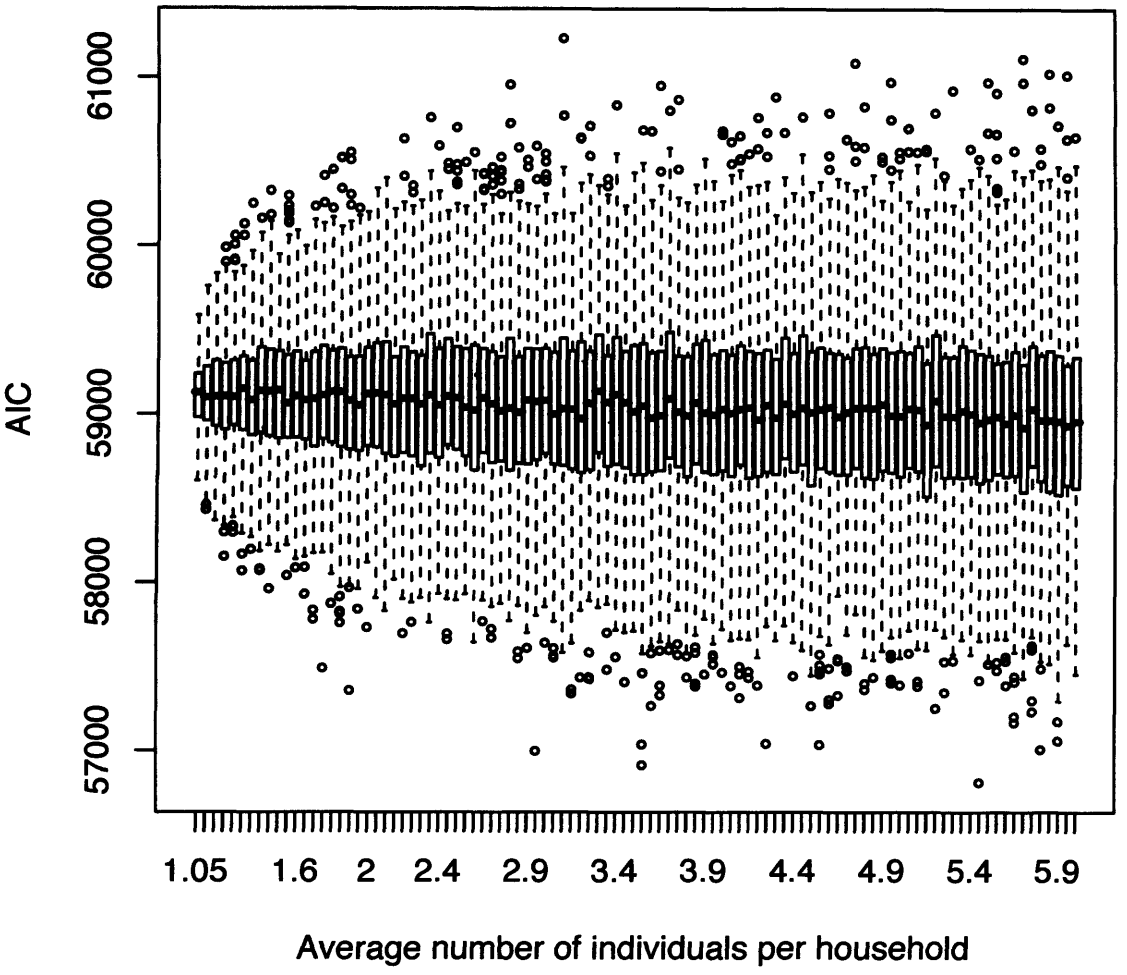


Table 6.18: Summary information for figure 6.14

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
AIC mean	59107	59071	59039	59019	58986
AIC variance	131472	233323	273426	290335	301396

Figure 6.15: Relationship between the variance components and the total number of individuals for three-level null model in scenario B

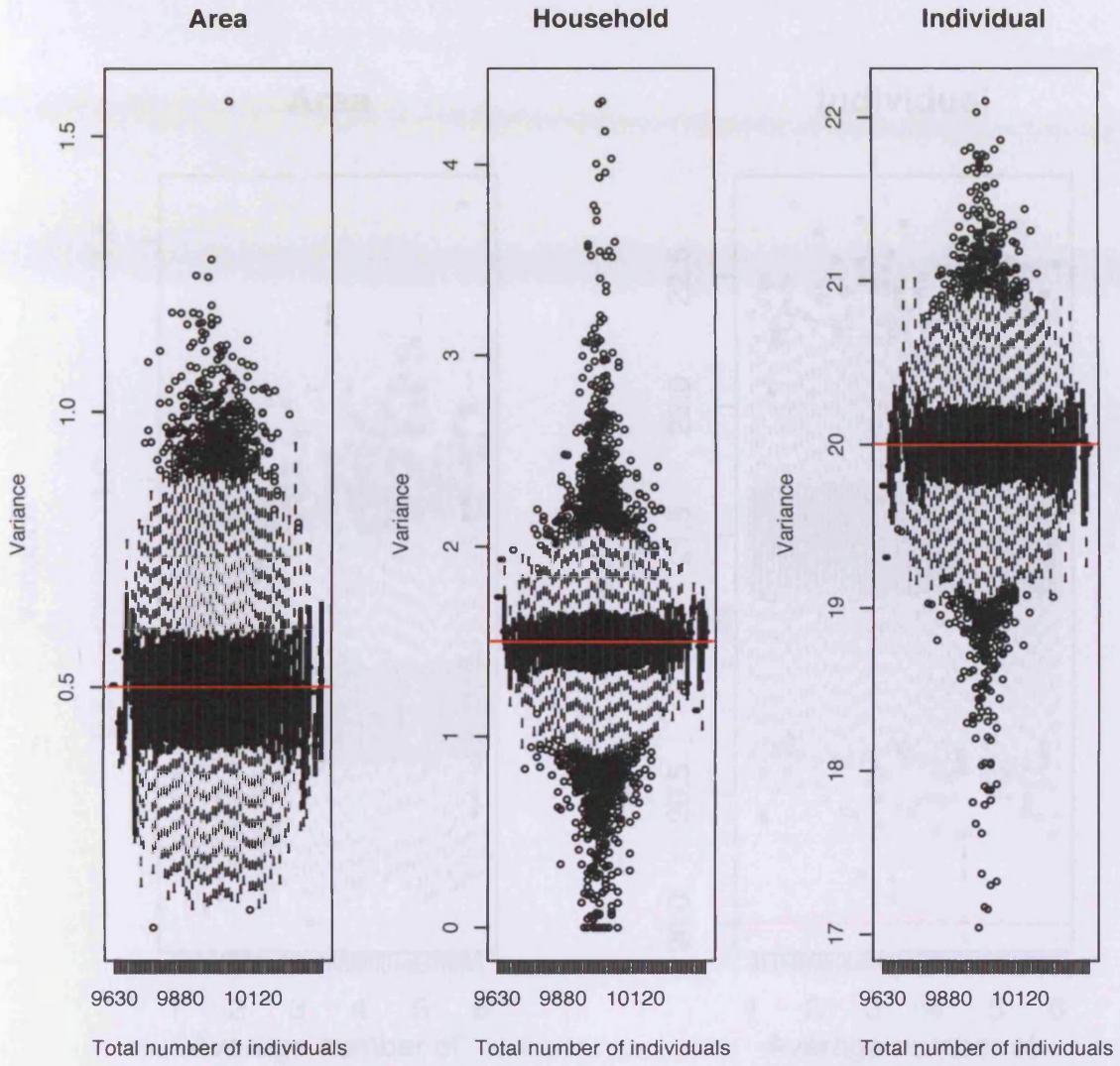


Table 6.19: Summary information for figure 6.15

Rounded number of individuals	9600	9700	9800	9900	10000	10100	10200	10300	10400
Area mean	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.50	0.51
Area variance	0.00	0.02	0.02	0.02	0.02	0.02	0.02	0.03	0.04
Household mean	1.73	1.47	1.50	1.50	1.51	1.50	1.50	1.51	1.55
Household variance	0.00	0.04	0.04	0.05	0.09	0.05	0.04	0.04	0.01
Individual mean	19.75	20.03	19.99	20.00	19.99	20.00	19.99	19.96	19.99
Individual variance	0.00	0.11	0.11	0.12	0.15	0.11	0.12	0.14	0.00

Figure 6.16: Relationship between the variance components and the average number of individuals per household for the two-level model, excluding household in scenario B

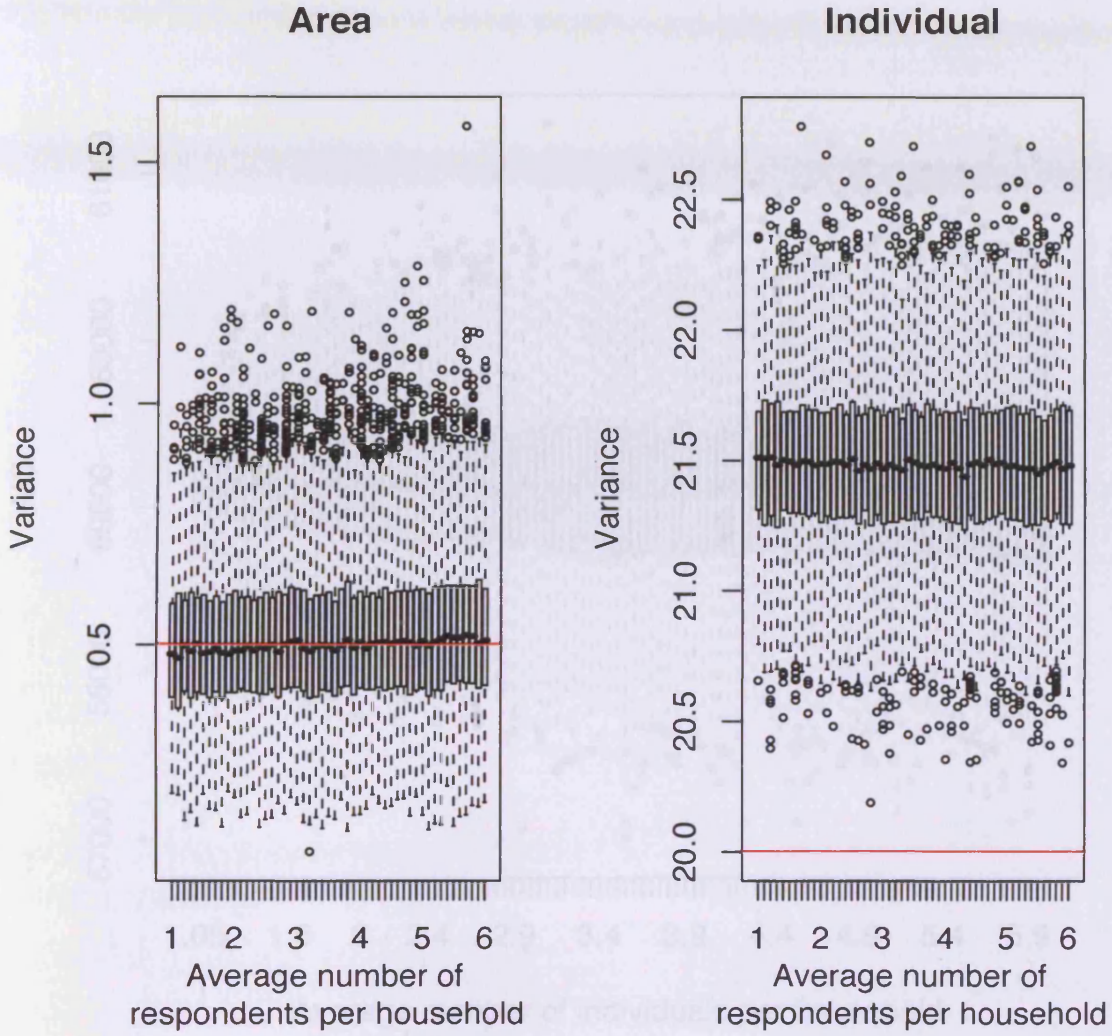


Table 6.20: Summary information for figure 6.16

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area mean	0.50	0.51	0.51	0.52	0.53
Area variance	0.02	0.02	0.02	0.02	0.02
Individual mean	21.50	21.48	21.49	21.48	21.48
Individual variance	0.26	0.12	0.11	0.10	0.10

Figure 6.17: Relationship between the average number of individuals per household and model fit for the two-level null model, excluding household in scenario B

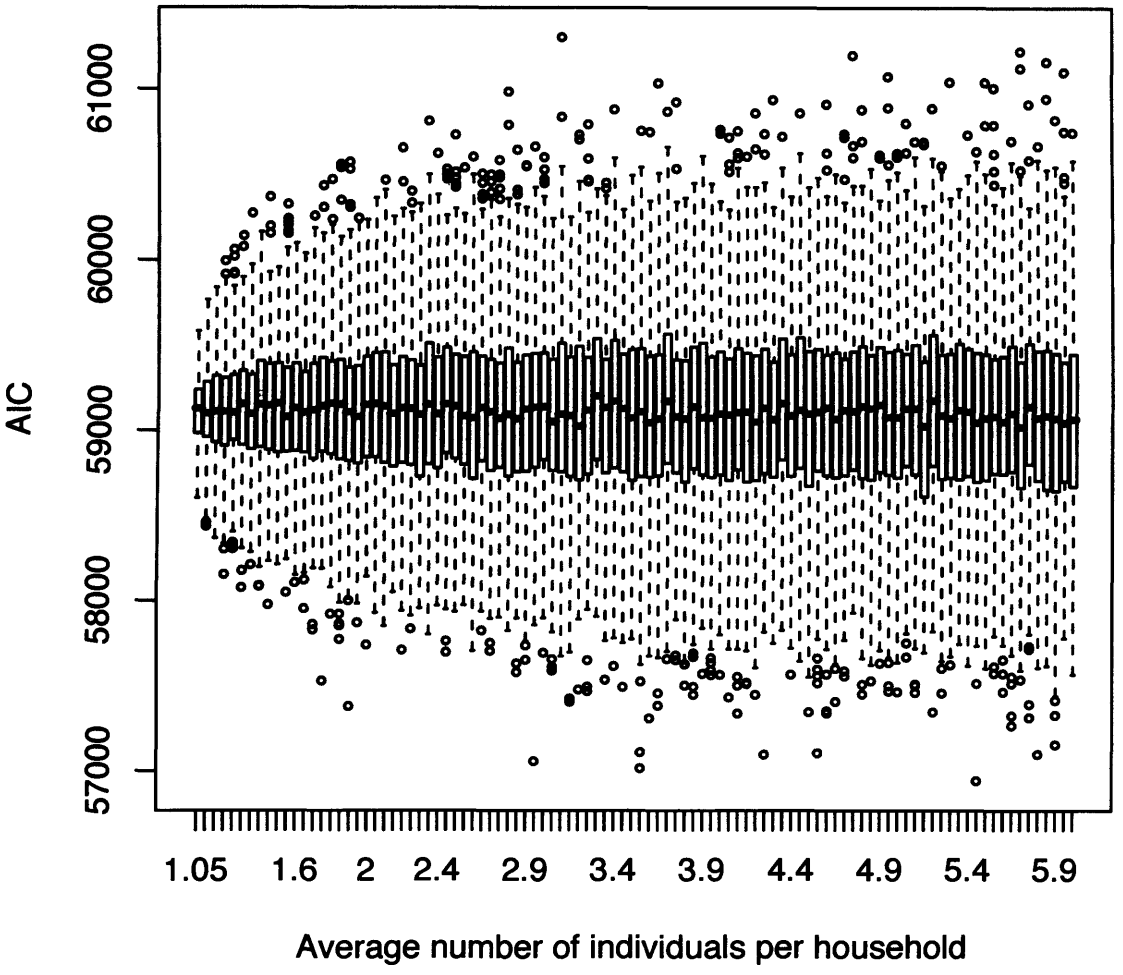


Table 6.21: Summary information for figure 6.17

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
AIC mean	59125	59118	59106	59105	59091
AIC variance	130086	233486	275059	292414	304360

Figure 6.18: Relationship between the variance components and total number of individuals for the two-level model, excluding household in scenario B

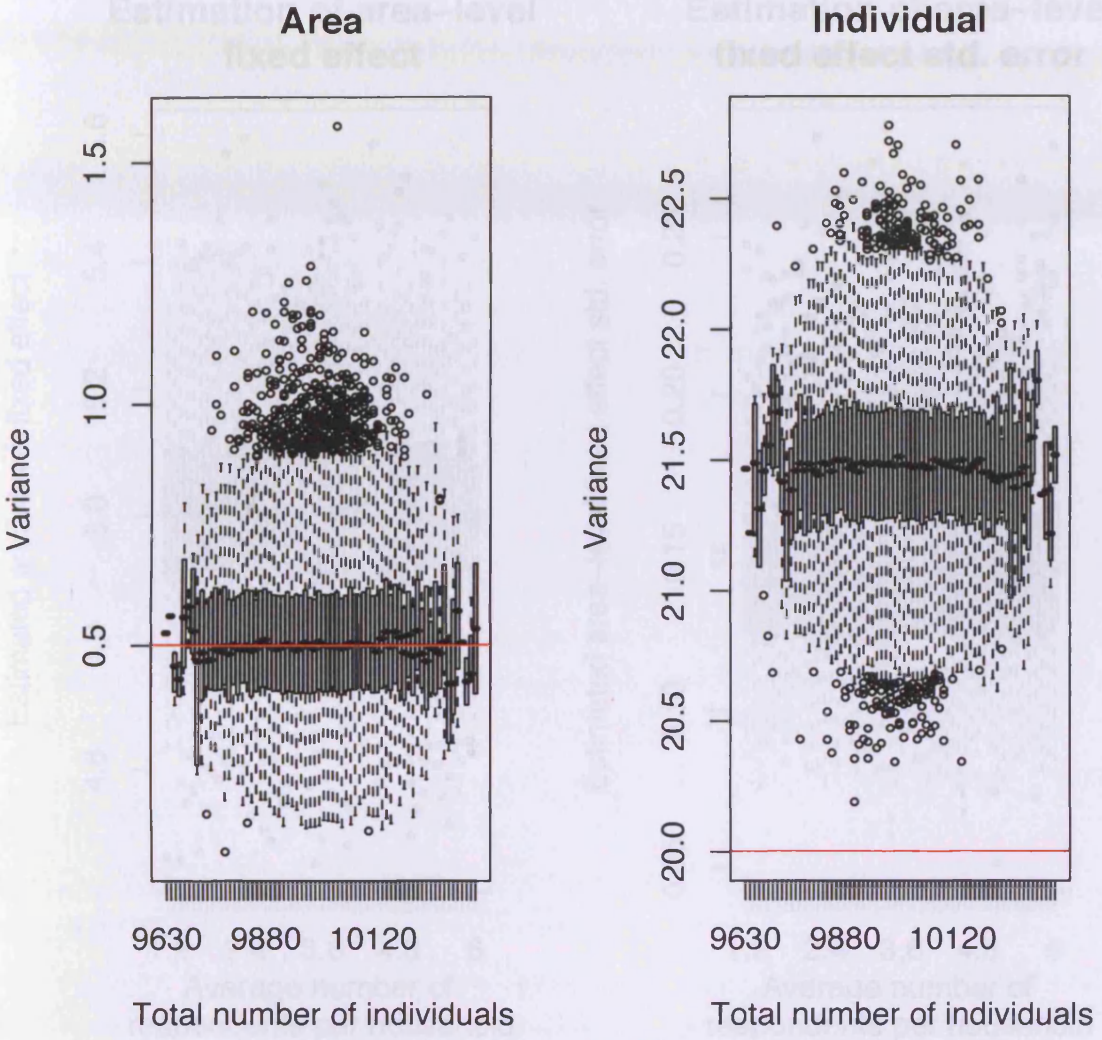


Table 6.22: Summary information for figure 6.18

Rounded number of individuals	9600	9700	9800	9900	10000	10100	10200	10300	10400
Area mean	0.53	0.52	0.51	0.51	0.51	0.52	0.52	0.51	0.53
Area variance		0.02	0.02	0.02	0.02	0.02	0.02	0.02	0.04
Individual mean	21.47	21.48	21.47	21.49	21.48	21.48	21.47	21.45	21.52
Individual variance		0.09	0.09	0.09	0.09	0.09	0.10	0.13	0.02

Figure 6.19: Relationship between the area-level fixed effect estimation and average number of individuals for the three-level model, in scenario B

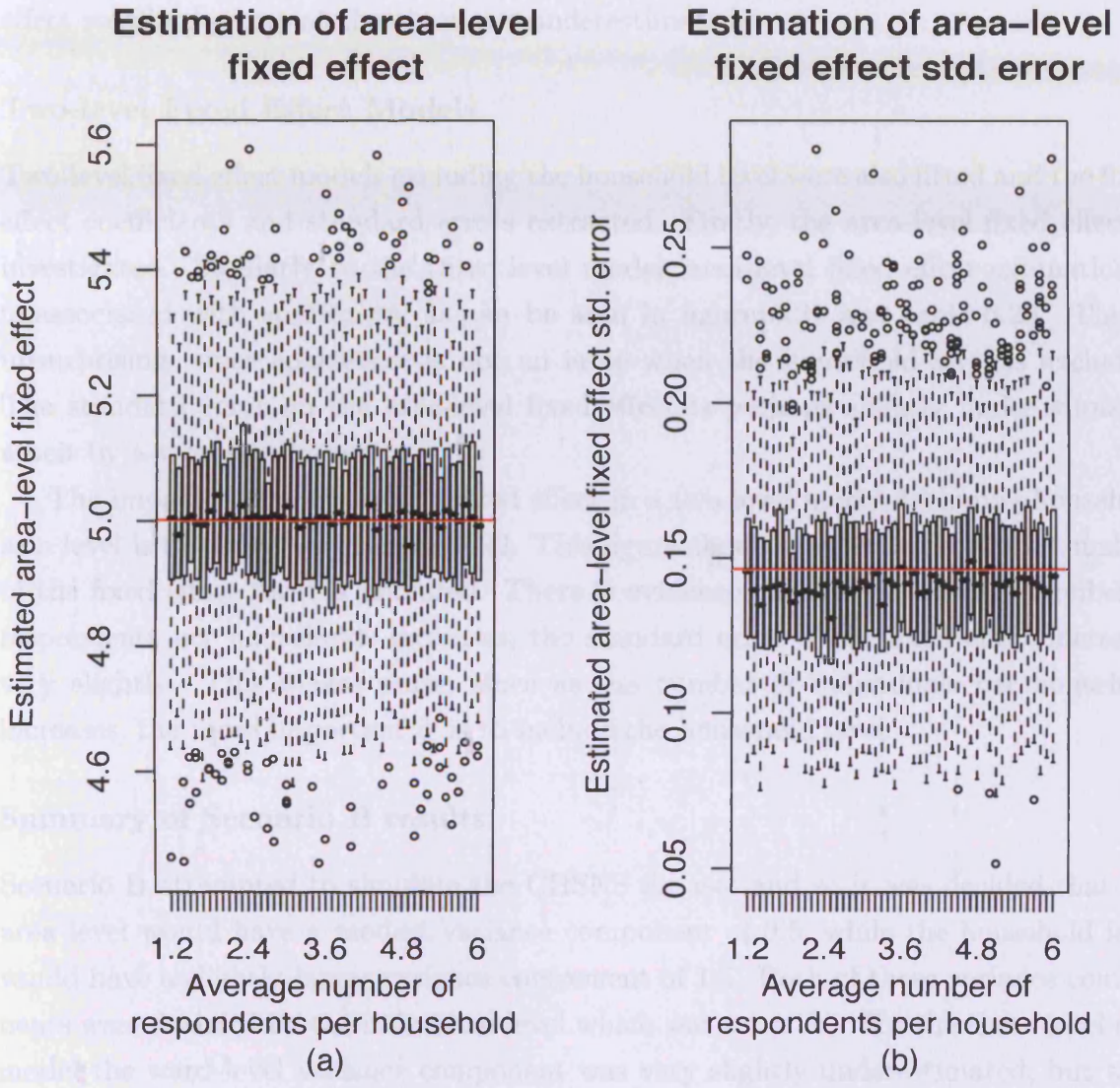


Table 6.23: Summary information for figure 6.19

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area fixed effect mean	5.00	5.01	5.00	5.01	5.01
Area fixed effect var.	0.02	0.02	0.02	0.02	0.02
Area fixed effect std. err. mean	0.1421	0.1428	0.1448	0.1454	0.1450
Area fixed effect std. err. var.	0.0007	0.0008	0.0008	0.0007	0.0008

is affected by sparseness. When the average number of individuals per household is between 1.05 and 2, the average standard error is 0.048. When the average number of respondents per household is high (between 5.05 and 6), the average standard error is over 10% larger at 0.054. This indicates that when the household level is sparse, fixed effect standard errors at that level are underestimated.

Two-level Fixed Effect Models

Two-level fixed effect models excluding the household level were also fitted and the fixed effect coefficients and standard errors extracted. Firstly, the area-level fixed effect is investigated. Similarly to the three level model, area-level fixed effect estimation is unassociated with sparseness, as can be seen in figure 6.21 and table 6.25. This is unsurprising, since sparseness is not an issue when the household level is excluded. The standard error on the area-level fixed effect is perhaps slightly underestimated albeit by a very small amount.

The impact on the household fixed effect in a two-level model excluding household as a level is investigated in figure 6.22. This figure shows that as before the estimation of the fixed effect itself is unbiased. There is evidence that as the average number of respondents per households increases, the standard error of the fixed effect increases very slightly. This makes sense, since as the number of individuals per household increases, the more important it is to include the household level.

Summary of Scenario B results

Scenario B attempted to simulate the CHSNS dataset and so it was decided that the area level would have a modest variance component of 0.5, while the household level would have a slightly larger variance component of 1.5. Both of these variance components were dwarfed by the individual level which was set at 20. For the three-level null model the ward-level variance component was very slightly underestimated, but with constant variability for all sparseness levels. Both the household and individual levels were unbiasedly estimated, however both demonstrated increasing variability with decreasing number of individuals per household. When the household level was excluded all of the variability attributable to that level was instead assigned to the individual level. Both the area- and household-level fixed effects are unbiasedly estimated for all sparseness levels. Similarly to scenario A however, the standard errors around those estimates are not so robust. For the area-level fixed effect the standard error is very slightly underestimated in both the three- and two-level models. The underestimation is quite small however and is unlikely to be a large problem. Household-level fixed effect standard errors for the three-level model however are underestimated even when the average number of individuals is relatively large (less than three per household). The

Figure 6.20: Relationship between the household-level fixed effect estimation and average number of individuals for the three-level model, in scenario B

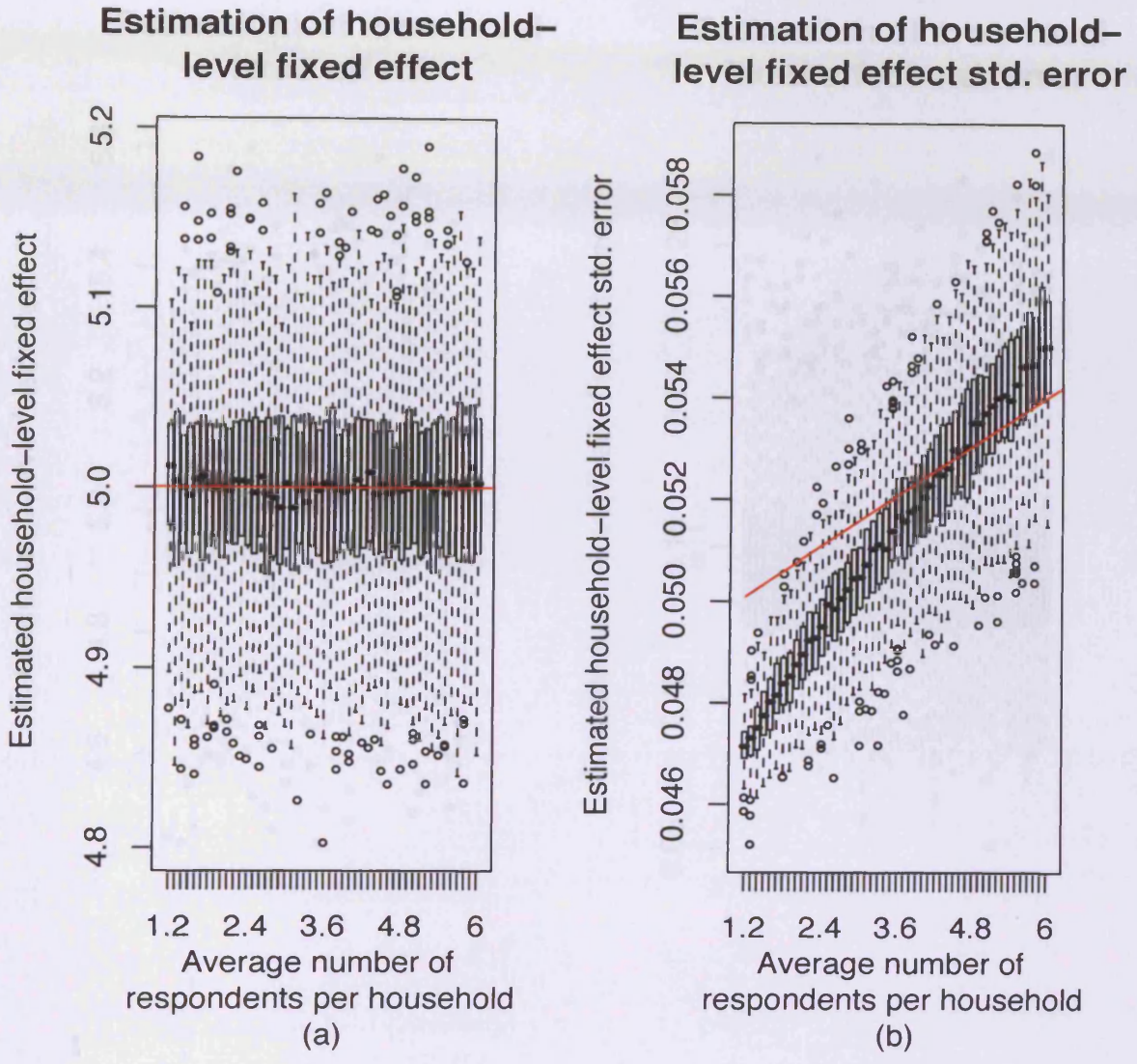


Table 6.24: Summary information for figure 6.20

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Household fixed effect mean	5.00	5.00	5.00	5.00	5.00
Household fixed effect var.	0.00	0.00	0.00	0.00	0.00
Household fixed effect std. err. mean	0.048	0.050	0.051	0.053	0.054
Household fixed effect std. err. var.	0.000	0.000	0.000	0.000	0.000

Figure 6.21: Relationship between the area-level fixed effect estimation and average number of individuals for the two-level model excluding household, in scenario B

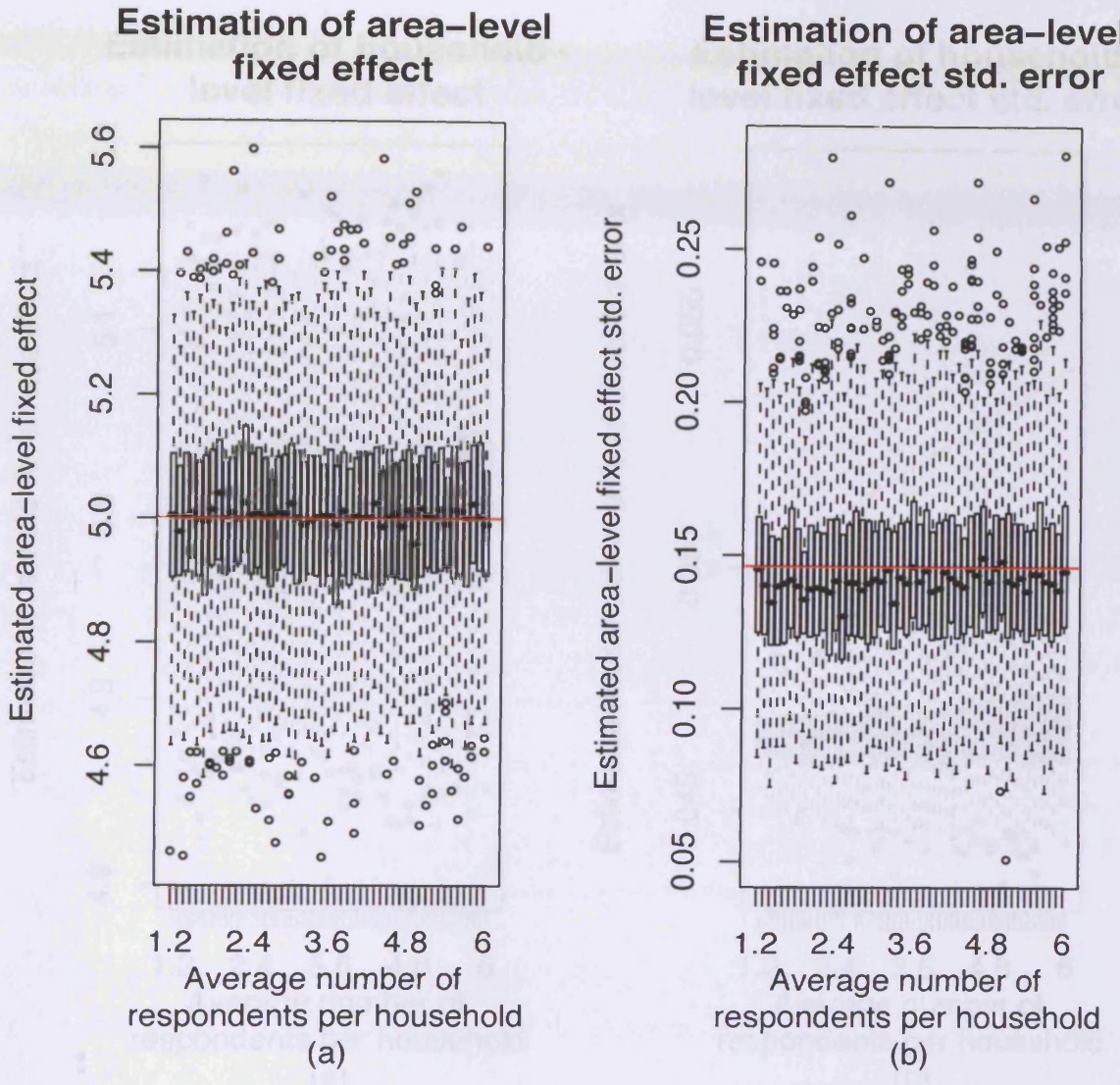


Table 6.25: Summary information for figure 6.21

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area fixed effect mean	5.00	5.01	5.00	5.01	5.01
Area fixed effect var.	0.02	0.02	0.02	0.02	0.02
Area fixed effect std. err. mean	0.142	0.143	0.145	0.145	0.145
Area fixed effect std. err. var.	0.001	0.001	0.001	0.001	0.001

Figure 6.22: Relationship between the household-level fixed effect estimation and average number of individuals for the two-level model excluding household, in scenario B

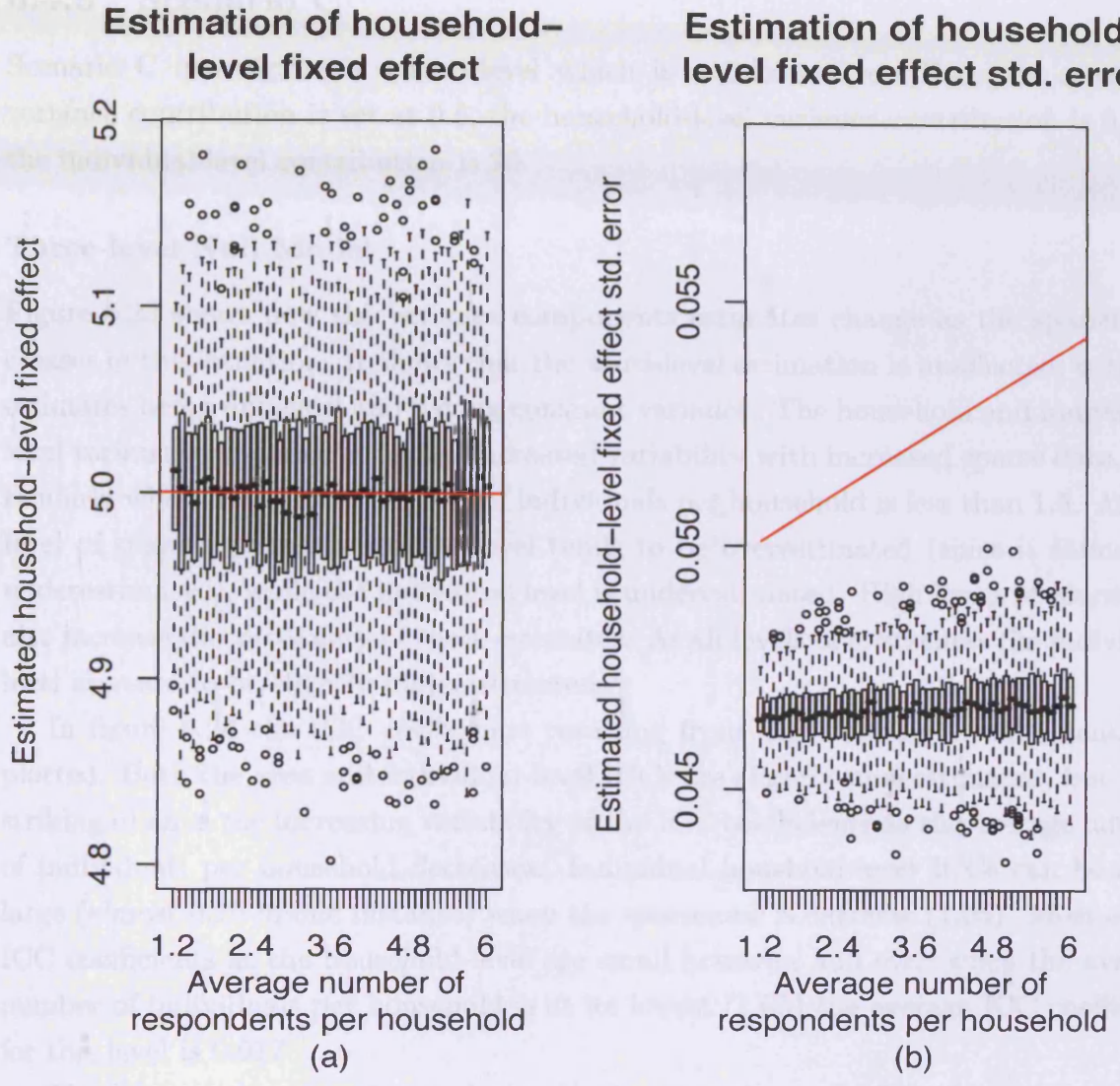


Table 6.26: Summary information for figure 6.22

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Household fixed effect mean	5.001	5.001	4.999	4.999	5.000
Household fixed effect var.	0.002	0.003	0.003	0.003	0.003
Household fixed effect std. err. mean	0.047	0.047	0.047	0.047	0.047
Household fixed effect std. err. var.	0.000	0.000	0.000	0.000	0.000

two-level model performs even worse however producing underestimated household-level standard errors for all levels of sparseness, getting progressively worse for larger numbers of individuals per household.

6.4.3 Scenario C

Scenario C investigates a sparse level which is uninformative. Here the area-level variance contribution is set at 0.5, the household-level variance contribution is 0, and the individual-level contribution is 20.

Three-level Null Model

Figure 6.23 shows how the variance components estimates change as the sparsity increases in this situation. It shows that the ward-level estimation is unaffected, with the estimates being unbiased and having constant variance. The household and individual-level variance components display increased variability with increased sparse data, particularly when the average number of individuals per household is less than 1.5. At this level of sparseness the household level tends to be overestimated (since it cannot be underestimated), while the individual level is underestimated. High levels of sparseness also increase the variability of these estimates. At all levels of sparseness the individual level appears to be slightly underestimated.

In figure 6.24 the ICC coefficients resulting from these variance components are plotted. Both the area and individual-level ICCs are slightly understimated, but most striking of all is the increasing variability of the ICC coefficients as the average number of individuals per household decreases. Individual houshold-level ICCs can be quite large (almost 0.15 in one instance) when the sparseness is extreme (1.05). Most of the ICC coefficients at the household level are small however, and even when the average number of individuals per household is at its lowest (1.05) the average ICC coefficient for this level is 0.017.

There is no indication of any relationship between the model fit and average number of respondents per household as shown in figure 6.25. This is to be expected, because if the household effect is non-existent then the average number of individuals per household would not be expected to have an effect.

Figure 6.26 confirms that the small deviations in the total number of individuals from 10,000 have no appreciable effect on the variance components estimation. The small underestimation of the area and individual-level variance components is not associated with the total sample size.

Fixed effects are not investigated for this scenario because when the household-level has a variance contribution of zero, the household-level fixed effect and the area-level fixed effect would not be expected to be affected by sparseness.

Figure 6.23: Relationship between the variance components and the average number of individuals per household for three-level null model in scenario C

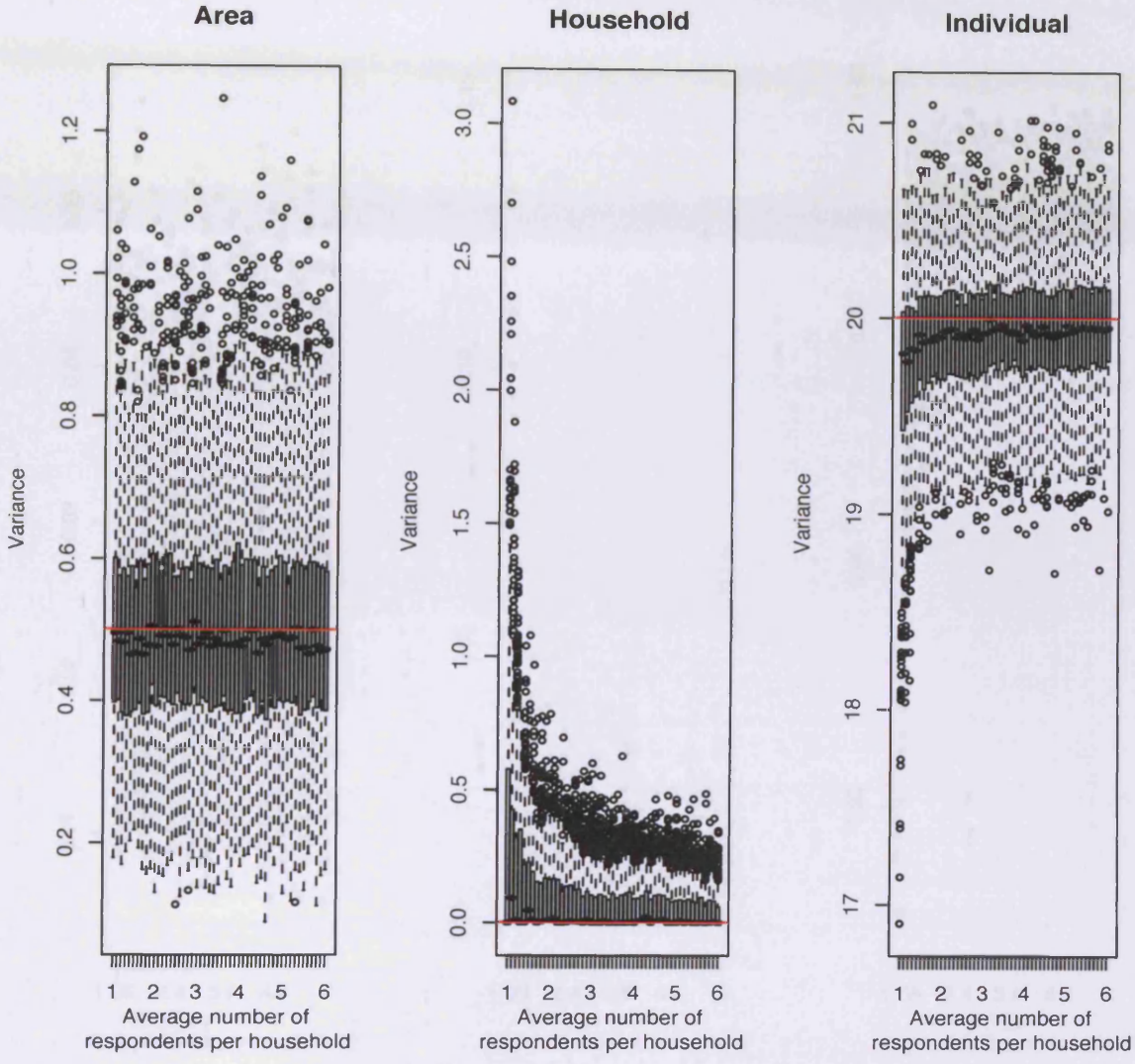


Table 6.27: Summary information for figure 6.23

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area mean	0.50	0.50	0.50	0.50	0.50
Area variance	0.02	0.02	0.02	0.02	0.02
Household mean	0.17	0.08	0.06	0.06	0.05
Household variance	0.08	0.01	0.01	0.01	0.01
Individual mean	19.83	19.92	19.94	19.94	19.95
Individual variance	0.16	0.09	0.08	0.09	0.08

Figure 6.24: Relationship between the ICC coefficients and the average number of individuals per household for three-level null model in scenario C

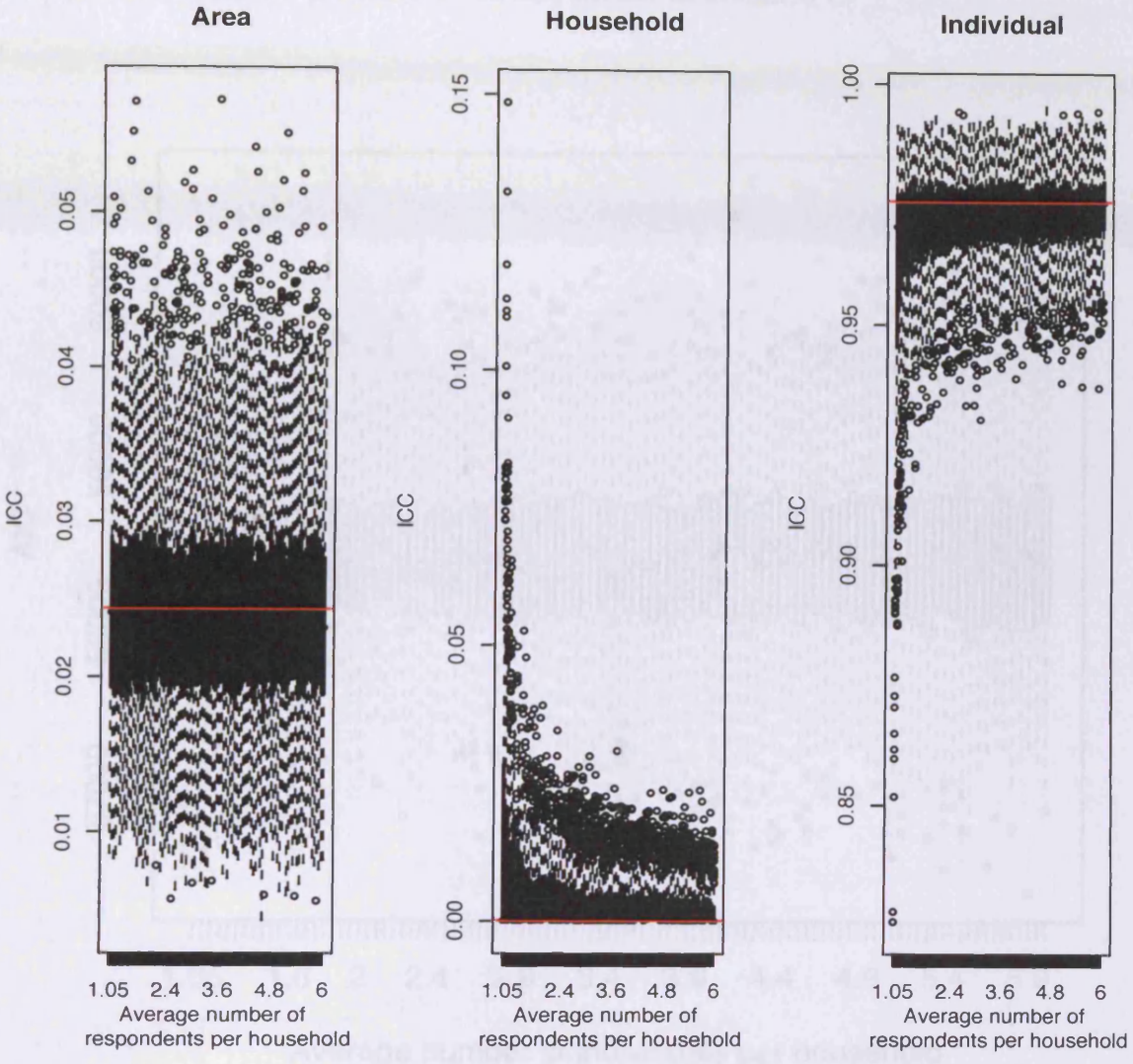


Table 6.28: Summary information for figure 6.24

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area mean	0.0242	0.0243	0.0242	0.0243	0.0242
Area variance	0.0001	0.0000	0.0000	0.0000	0.0000
Household mean	0.0082	0.0037	0.0029	0.0028	0.0024
Household variance	0.0002	0.0000	0.0000	0.0000	0.0000
Individual mean	0.9676	0.9720	0.9729	0.9729	0.9734
Individual variance	0.0003	0.0001	0.0001	0.0001	0.0001

Figure 6.25: Relationship between the average number of individuals per household and the model fit for the three-level null model in scenario C

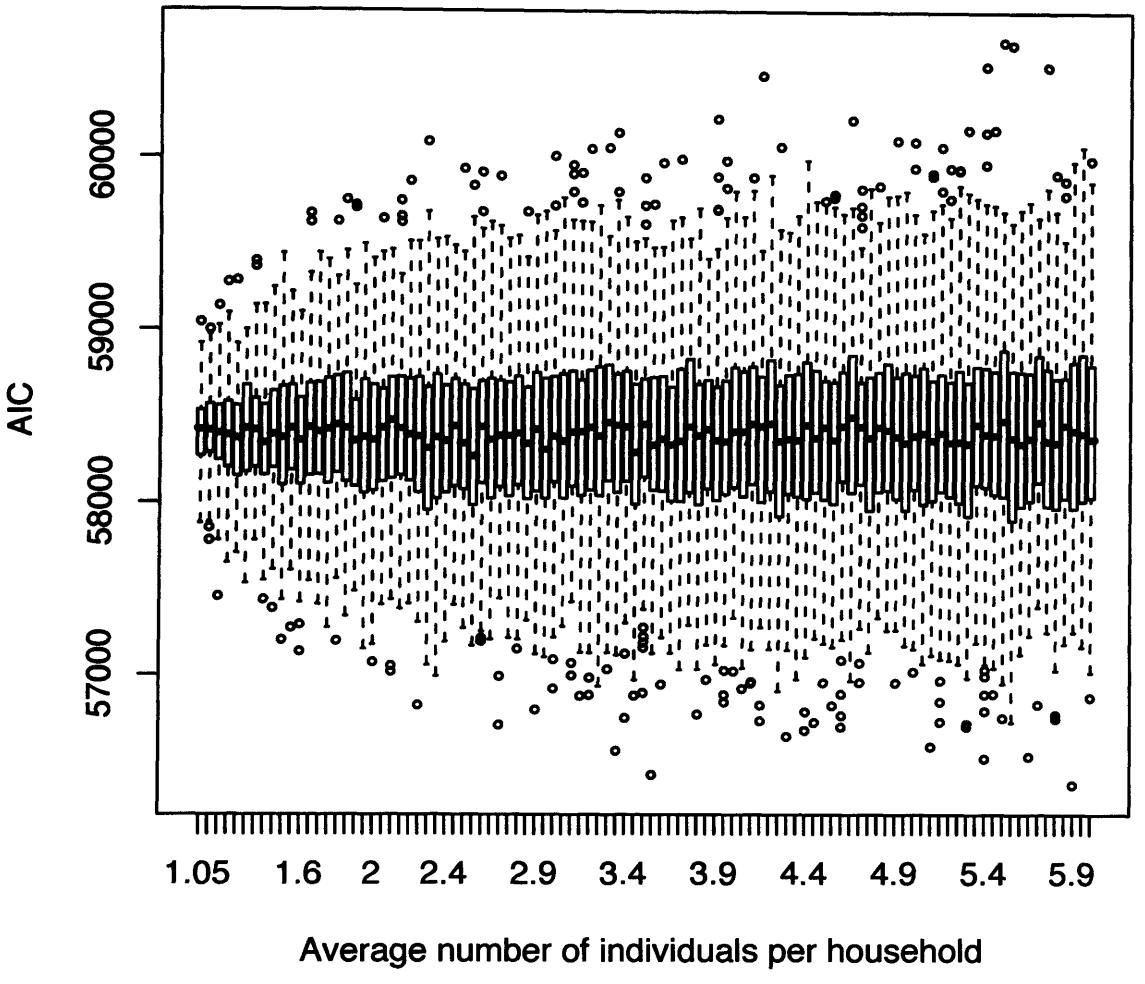


Table 6.29: Summary information for figure 6.25

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
AIC mean	58399	58389	58401	58409	58398
AIC variance	127431	218898	262144	280199	308395

Figure 6.26: Relationship between the variance components and the total number of individuals for three-level null model in scenario C

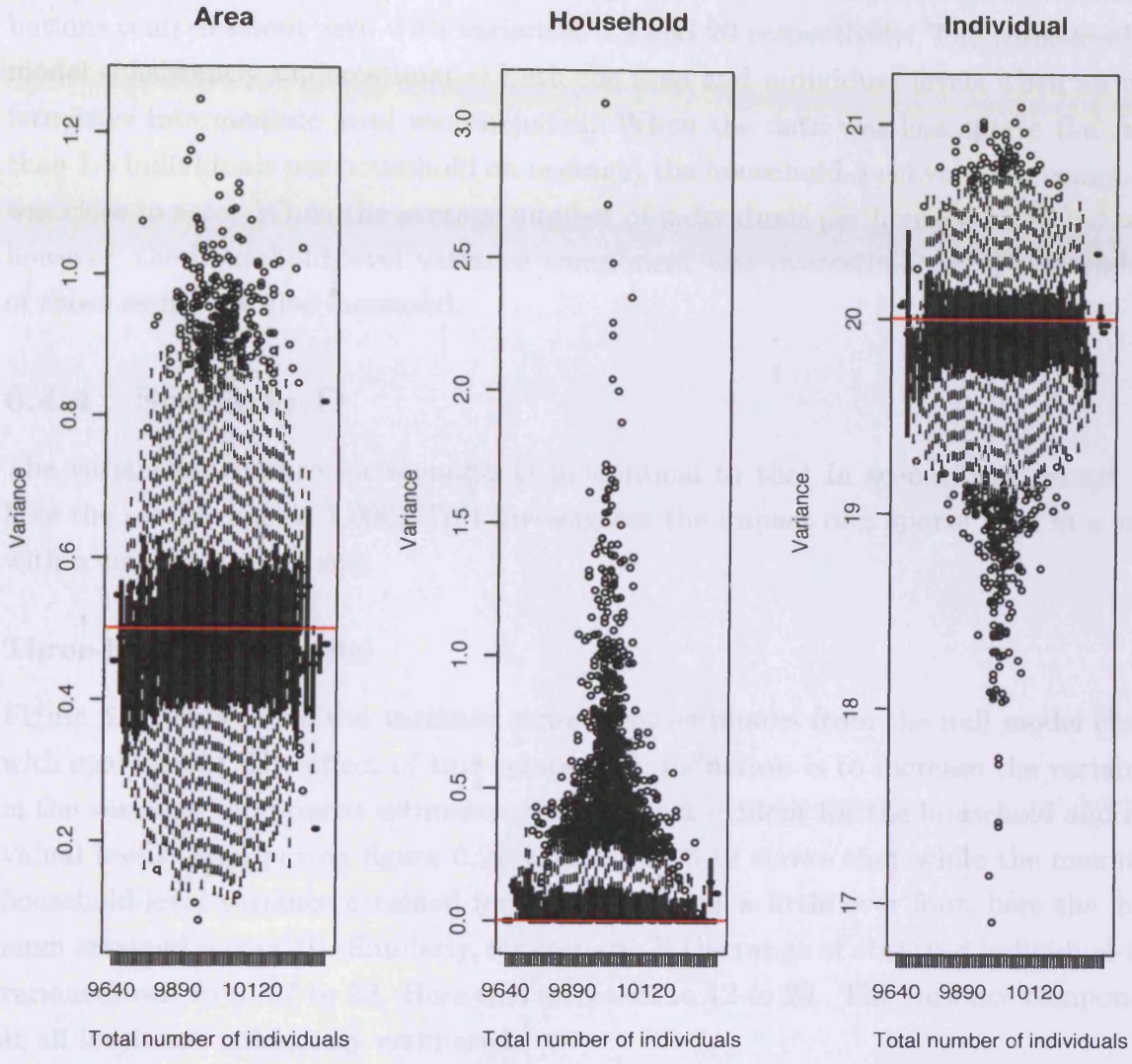


Table 6.30: Summary information for figure 6.26

Rounded number of individuals	9600	9700	9800	9900	10000	10100	10200	10300	10400
Area mean	0.45	0.45	0.50	0.50	0.50	0.50	0.50	0.51	0.56
Area variance		0.01	0.02	0.02	0.02	0.02	0.02	0.03	0.03
Household mean	0.00	0.09	0.06	0.07	0.09	0.07	0.06	0.04	0.10
Household variance		0.01	0.01	0.01	0.04	0.01	0.01	0.00	0.01
Individual mean	20.07	19.80	19.93	19.93	19.91	19.92	19.93	19.94	19.99
Individual variance		0.08	0.09	0.09	0.11	0.09	0.08	0.09	0.00

Summary of Scenario C results

Scenario C investigated the effect of including an uninformative level in the analysis. The area and individual-level variance components were simulated from normal distributions centred about zero with variances 0.5 and 20 respectively. The three-level null model consistently underestimated both the area and individual levels when an uninformative intermediate level was included. When the data was less sparse (i.e. more than 1.5 individuals per household on average) the household-level variance component was close to zero. When the average number of individuals per household fell below 1.5 however, the household-level variance component was overestimated. The variability of those estimates also increased.

6.4.4 Scenario D

The variance structure for scenario D is identical to that in scenario B, except that here the sample size is 1,000. This investigates the impact of a sparse level in a study with a smaller sample size.

Three-level Null Model

Figure 6.27 shows how the variance component estimates from the null model change with sparseness. The effect of this sample size reduction is to increase the variability in the variance component estimates. This is most evident for the household and individual levels. Comparing figure 6.27 with figure 6.12 shows that while the maximum household-level variance attained for scenario B was a little over four, here the maximum attained is over 10. Similarly, for scenario B the range of observed individual-level variances was from 17 to 22. Here this increases to 12 to 24. The variance components at all levels are unbiasedly estimated.

The relationship between the ICC coefficients for the each level and the sparseness is plotted in figure 6.28. This plot illustrates the large effect that sparseness can have on the ICC coefficient when the sample size is small. When the average number of individuals per household is 1.05 the lower and higher quartiles of the household ICC are 0 and 0.17. The true ICC for this level (as indicated by the horizontal line in figure 6.28 is 0.07.

The relationship between the sparseness of the household level and the model fit is a horizontal line as it is for scenario B (figure 6.14) and is not presented here. Similarly, the relationship between the sparseness and the total sample size is similar to that for scenario B (6.15), except that the total sample size is centred about 1,000 instead of 10,000 and the increased variability of the variance component estimates is also evident.

Figure 6.27: Relationship between the variance components and the average number of individuals per household for three-level null model in scenario D

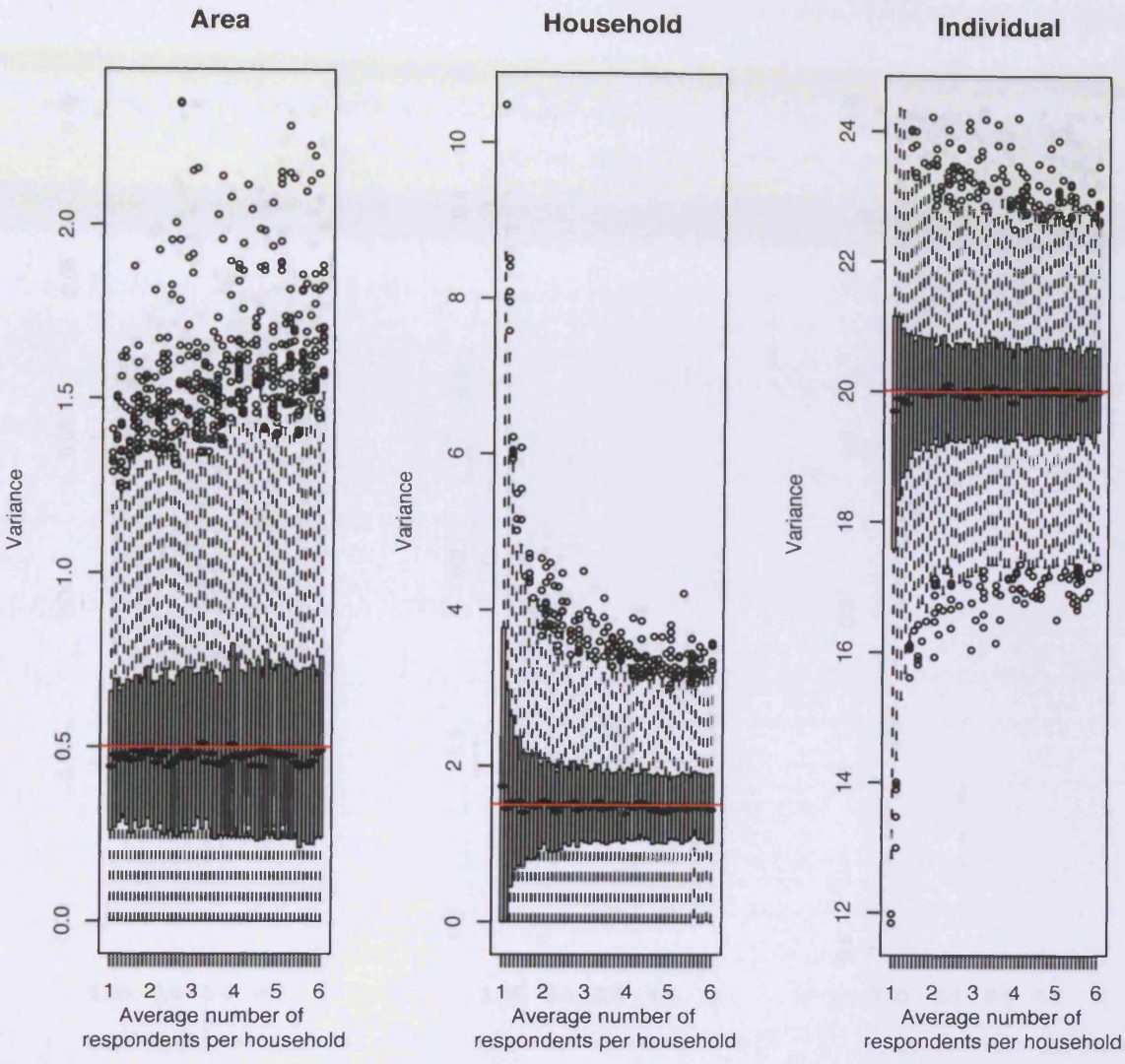


Table 6.31: Summary information for figure 6.27

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area mean	0.50	0.50	0.51	0.51	0.51
Area variance	0.09	0.11	0.11	0.13	0.13
Household mean	1.61	1.50	1.50	1.48	1.49
Household variance	1.46	0.59	0.46	0.39	0.38
Individual mean	19.88	19.99	20.01	20.01	20.00
Individual variance	2.13	1.19	1.09	1.01	0.98

Figure 6.28: Relationship between the ICC coefficients and the average number of individuals per household for three-level null model in scenario D

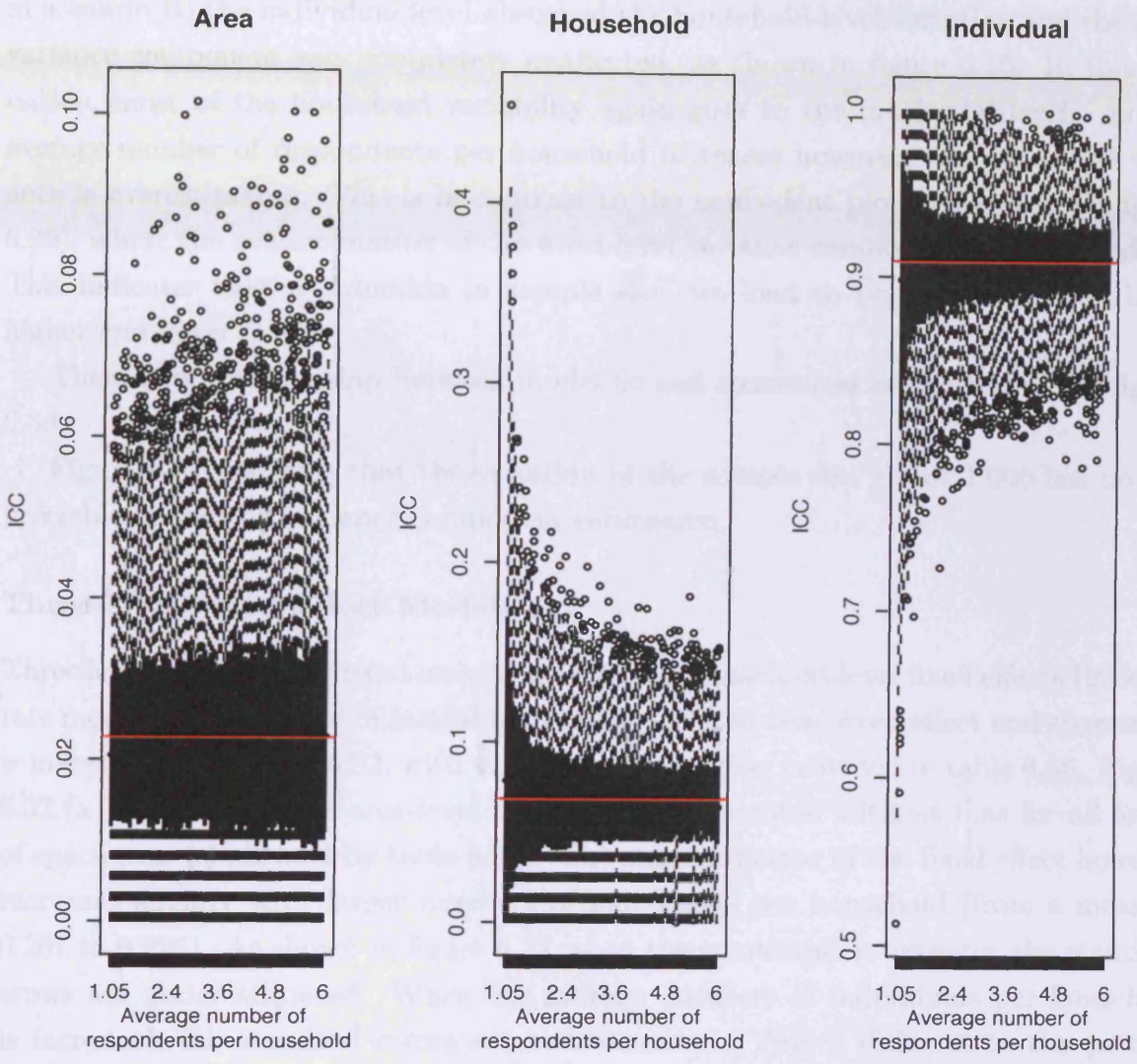


Table 6.32: Summary information for figure 6.28

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area mean	0.0227	0.0228	0.0230	0.0232	0.0230
Area variance	0.0002	0.0002	0.0002	0.0003	0.0003
Household mean	0.0733	0.0680	0.0679	0.0673	0.0676
Household variance	0.0030	0.0012	0.0009	0.0008	0.0007
Individual mean	0.9040	0.9092	0.9091	0.9095	0.9094
Individual variance	0.0030	0.0012	0.0010	0.0008	0.0008

Two-level Null Model

Excluding the household level has a slightly different impact on the variance components when the total sample size is small. When the total sample size was large (as in scenario B) the individual level absorbed the household-level variation and the area variance component was completely unaffected, as shown in figure 6.16. In this situation, most of the household variability again goes to the individual level. As the average number of respondents per household increases however, the ward-level variance is overestimated. This is in contrast to the equivalent plot in scenario B (figure 6.29), where the overestimation of the ward-level variance component was less evident. This indicates that a reduction in sample size can lead to poor estimation of both higher and lower levels.

There is no relationship between model fit and sparseness as can be seen in figure 6.30.

Figure 6.31 confirms that the variation of the sample size about 1,000 has no appreciable impact on variance component estimation.

Three-level Fixed Effect Models

Three level models were fitted including area- and household-level fixed effects (in separate models). Firstly, the relationship between the area-level fixed effect and sparseness is investigated in figure 6.32, with summary information provided in table 6.36. Figure 6.32.(a) shows that the area-level fixed effect is estimated without bias for all levels of sparseness (confirmed by table 6.36). The standard error of the fixed effect however increased slightly with larger numbers of individuals per household (from a mean of 0.207 to 0.226). As shown in figure 6.32 when the sparseness is extreme, the standard errors are underestimated. When the average numbers of individuals per household is increased, the standard errors are overestimated. This is different to the pattern observed in scenario B (which used the same variance structure but a larger sample size), indicating that this may just be a result of increased variability due to a small sample size.

Figure 6.33 shows the relationship between the household-level fixed effect estimation and sparseness for the three-level model. Again, the fixed effect itself is unbiasedly estimated for all levels of sparseness, and again the standard error of those fixed effects are positively related with increasing numbers of respondents per household. The average standard errors range from 0.153 (for average number of individuals per household between 1.05 and 2) and 0.176 (for average numbers of respondents 5.05 to 6) as shown in table 6.37 and are underestimated when the average per household is small.

Figure 6.29: Relationship between the variance components and the average number of individuals per household for the two-level model, excluding household in scenario D

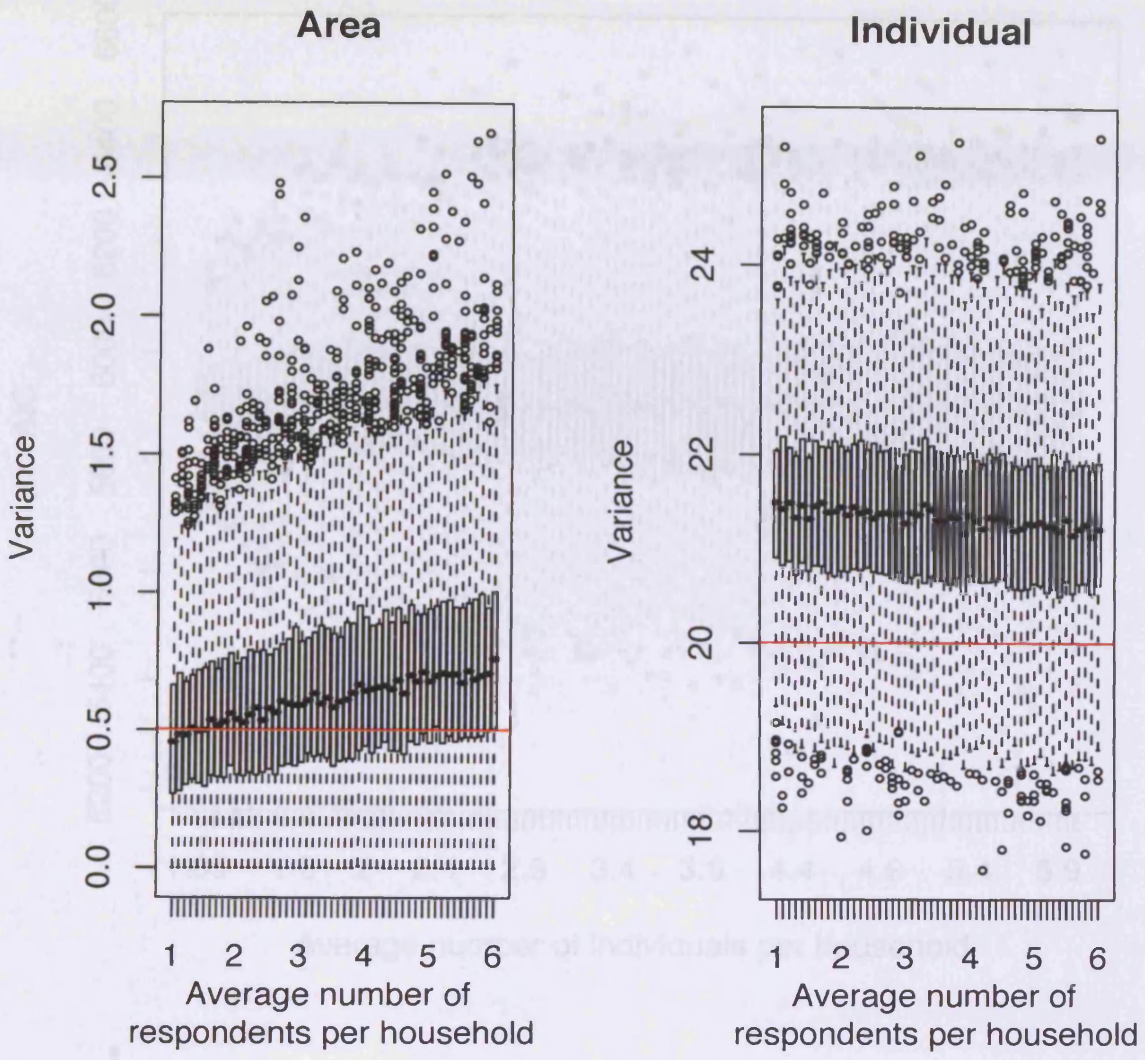
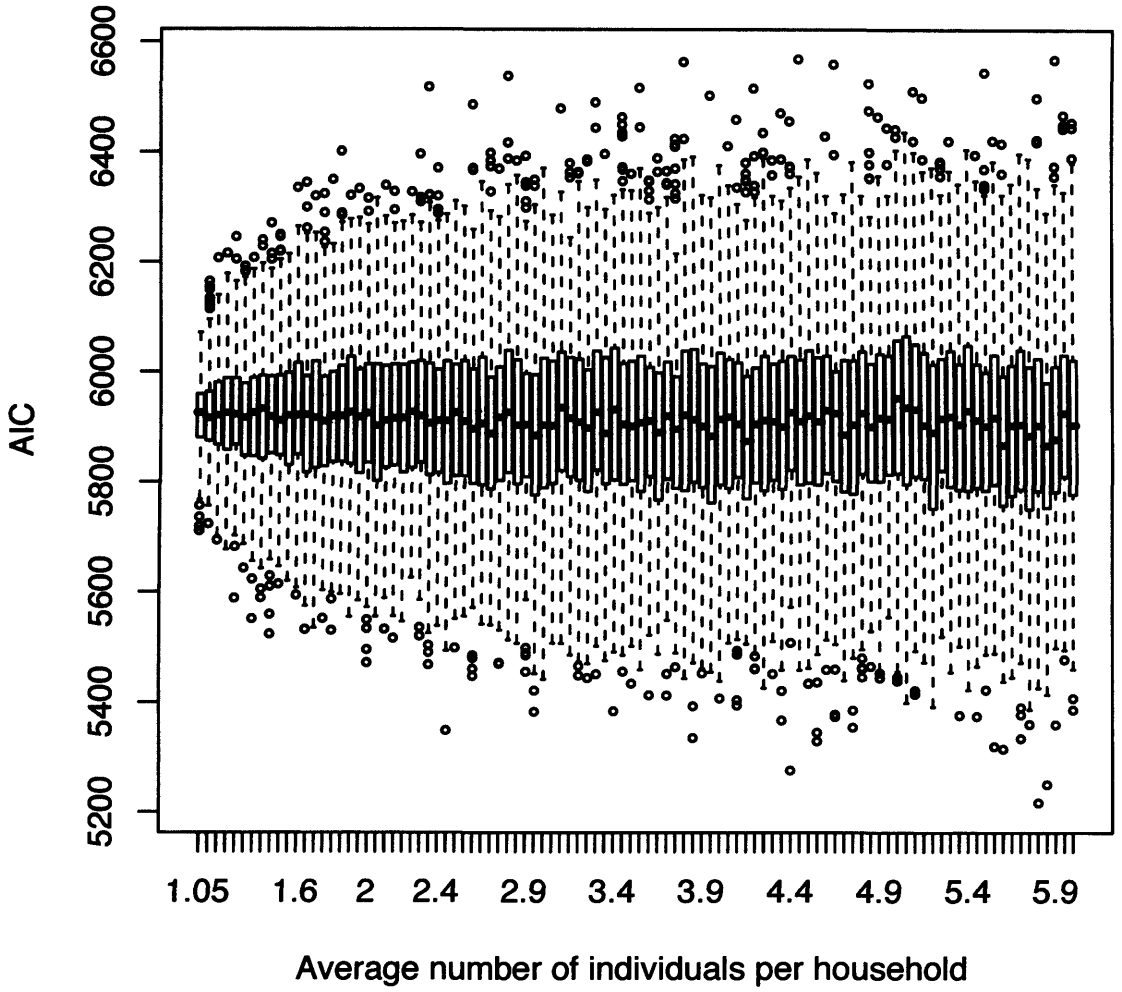


Table 6.33: Summary information for figure 6.29

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area mean	0.54	0.60	0.65	0.70	0.74
Area std. dev.	0.09	0.11	0.12	0.13	0.14
Individual mean	21.46	21.40	21.37	21.31	21.27
Individual variance	2.13	1.19	1.09	1.01	0.98

Figure 6.30: Relationship between the average number of individuals per household and model fit for the two-level null model, excluding household in scenario D



Two-level Fixed Effect Models

Here the effect of excluding the household level on fixed effect estimation is investigated. First the area-level fixed effect is examined. Figure 6.34 shows the relationship between area-level fixed effect estimation and sparseness for scenario D. As before, the fixed

Table 6.34: Summary information for figure 6.30

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
AIC mean	5921	5911	5911	5913	5906
AIC variance	13217	23330	27573	28949	31311

Figure 6.31: Relationship between the variance components and total number of individuals for the two-level model, excluding household in scenario D

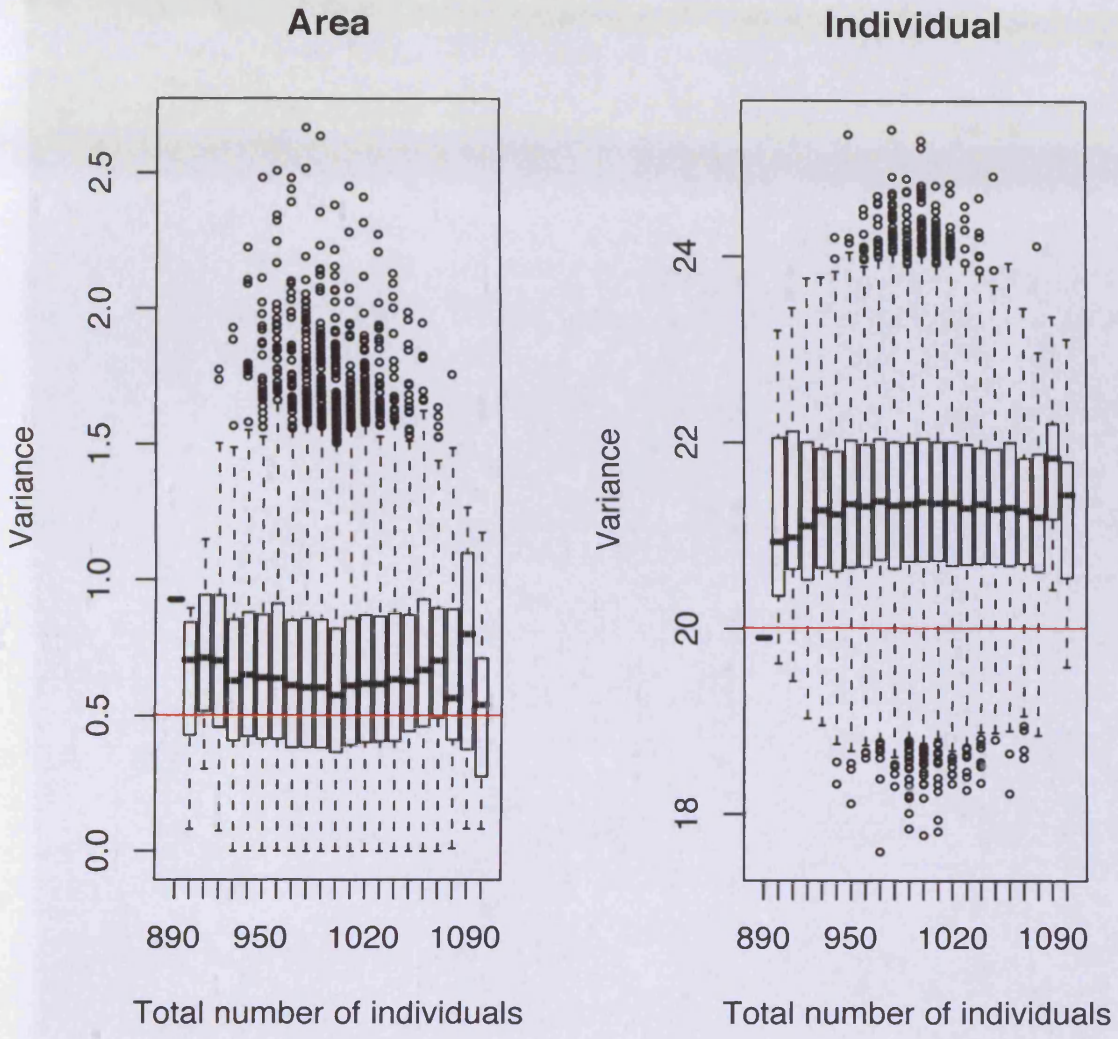


Table 6.35: Summary information for figure 6.31

Rounded number of individuals	900	1000	1100
Area mean	0.68	0.64	0.70
Area variance	0.13	0.13	0.12
Individual mean	21.32	21.36	21.30
Individual variance	1.03	0.96	0.95

Figure 6.32: Relationship between the area-level fixed effect estimation and the average number of individuals per household, for the three-level model in scenario D

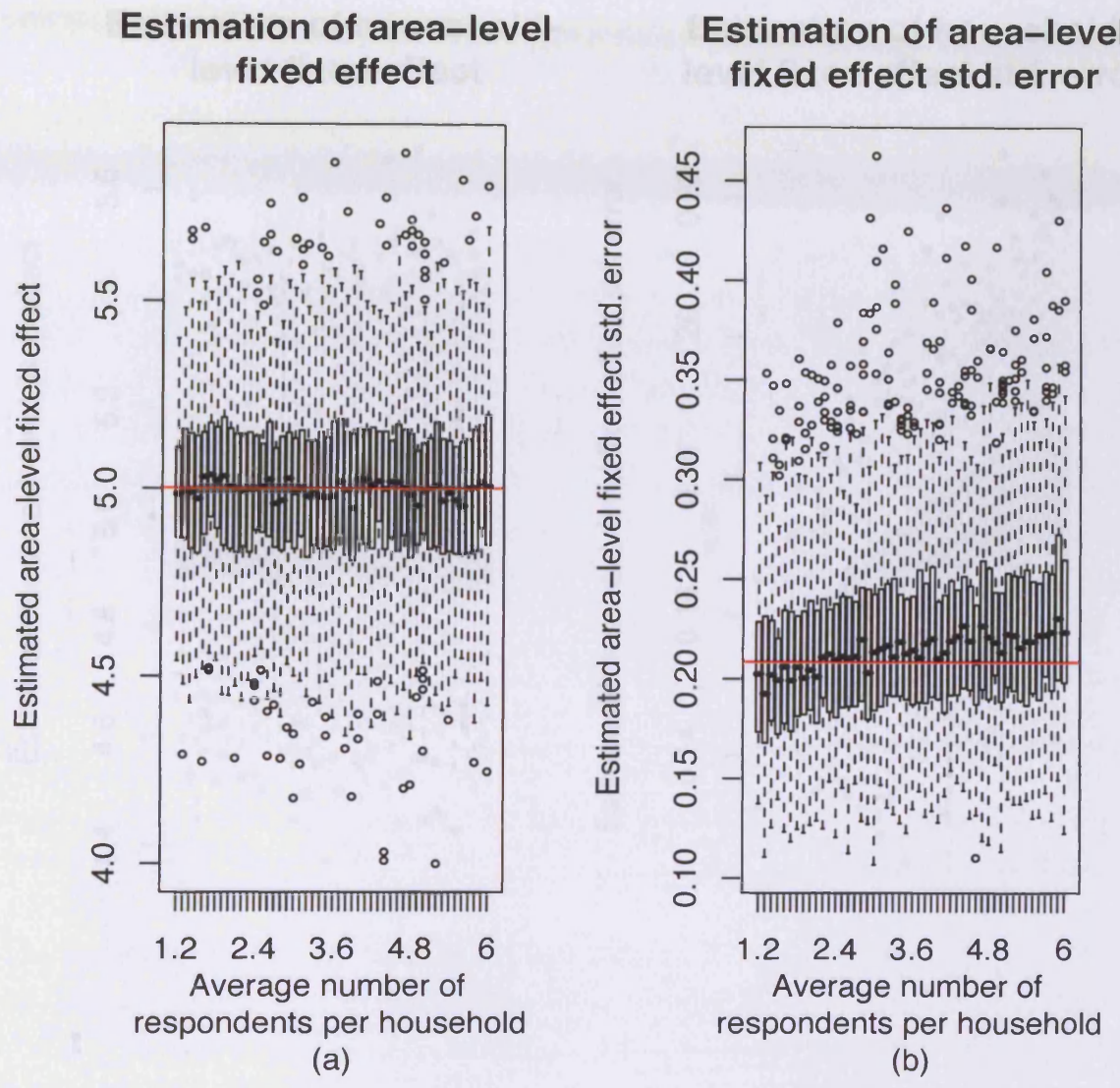


Table 6.36: Summary information for figure 6.32

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area fixed effect mean	5.01	5.00	4.99	5.00	5.00
Area fixed effect var.	0.04	0.05	0.05	0.05	0.05
Area fixed effect std. error mean	0.2074	0.2149	0.2192	0.2226	0.2261
Area fixed effect std. error var.	0.0016	0.0018	0.0018	0.0019	0.0020

Figure 6.33: Relationship between the household-level fixed effect estimation and the average number of individuals per household, for the three-level model in scenario D

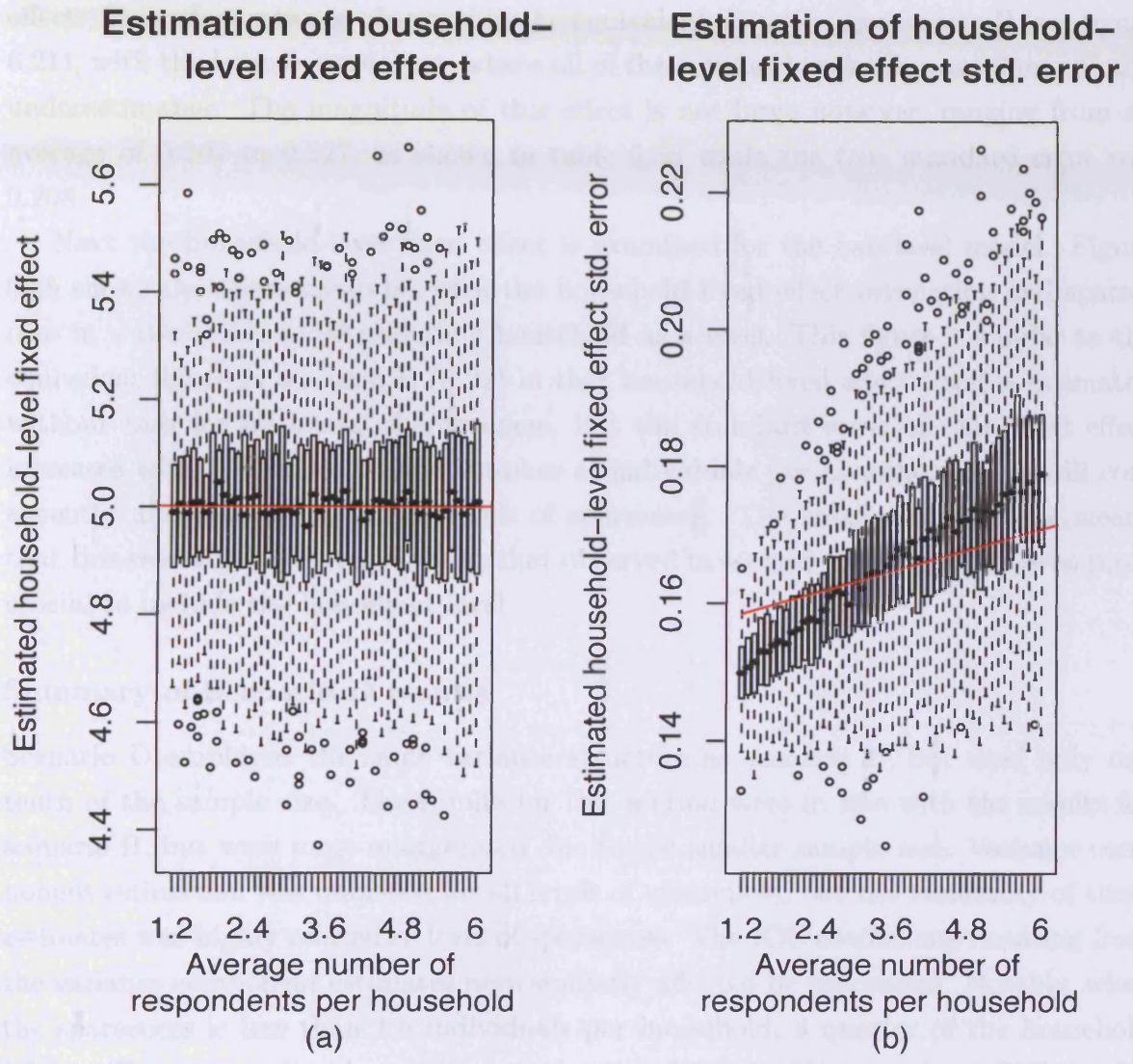


Table 6.37: Summary information for figure 6.33

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Household fixed effect mean	5.00	5.00	5.01	5.00	5.00
Household fixed effect var.	0.02	0.03	0.03	0.03	0.03
Household fixed effect std. error mean	0.1532	0.1593	0.1659	0.1715	0.1764
Household fixed effect std. error var.	0.0000	0.0001	0.0001	0.0002	0.0002

effect is unbiasedly estimated for all levels of sparseness. There is a small positive relationship between the area-level fixed effect standard error and the average number of individuals per household. This is perhaps surprising, since the exclusion of the household level would not be expected to have any impact on a higher level fixed effect. This effect was not observed in the equivalent situation in scenario B (see figure 6.21), with the larger sample size, where all of the standard errors appeared marginally underestimated. The magnitude of this effect is not large however, ranging from an average of 0.207 to 0.227, as shown in table 6.38 while the true standard error was 0.208.

Next the household-level fixed effect is examined for the two-level model. Figure 6.35 shows the relationship between the household fixed effect estimation and sparseness in a two-level model excluding household as a level. This figure is similar to the equivalent figure in scenario B (6.22) in that household fixed effect itself is estimated without bias for all levels of sparseness, but the standard error of that fixed effect increases with increasing average number of individuals per household. It is still consistently underestimated for all levels of sparseness. The smaller sample size means that this relationship is weaker than that observed in scenario B, where it is even more crucial to include the household level.

Summary of Scenario D results

Scenario D employed the same variance structure as scenario B, but used only one tenth of the sample size. The results for this section were in line with the results for scenario B, but were more exaggerated due to the smaller sample size. Variance component estimation was unbiased for all levels of sparseness, but the variability of those estimates was highly related to level of sparseness. The ICC coefficients resulting from the variance component estimates were similarly affected by sparseness. Notably, when the sparseness is less than 1.5 individuals per household, a quarter of the household ICC coefficients produced overestimate the true ICC (0.068) by at least 70% (and a quarter underestimate the true ICC by at least 66%). This level of precision calls into question studies reporting ICC coefficients for households from datasets with household sparseness. When the middle level was excluded the variability from that level was split between the higher and lower levels. This is in contrast to scenario B which had sufficient sample size to estimate the ward-level variability robustly. Fixed effect estimation for the area-level was unbiased and did not suffer greatly from underestimated standard errors for either the three- or two-level models. The household-level fixed effects standard errors however were underestimated at extreme sparseness for the three-level model and underestimated for all levels of sparseness in the two-level model.

Figure 6.34: Relationship between the area-level fixed effect estimation and the average number of individuals per household, for the two-level model in scenario D

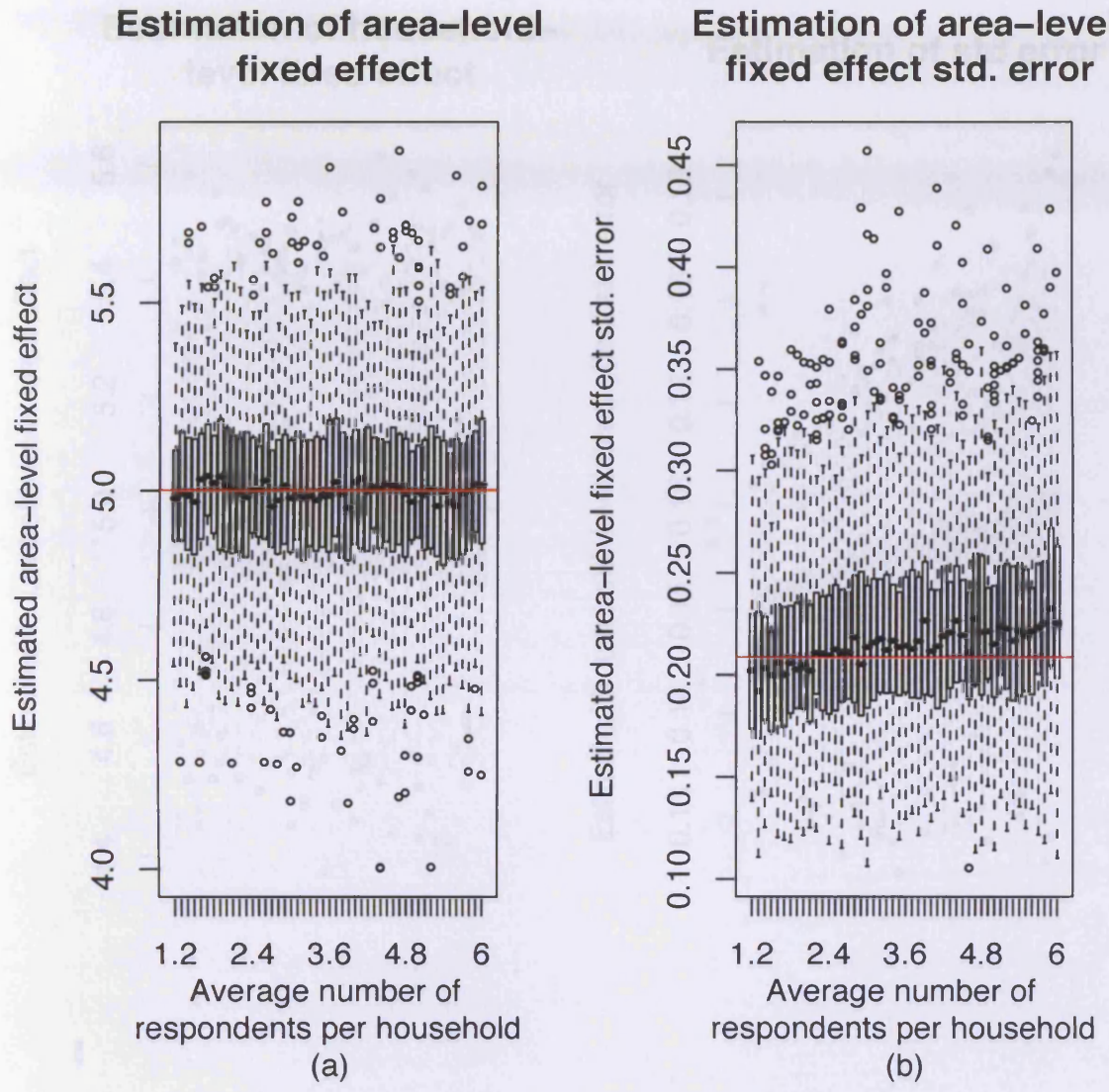


Table 6.38: Summary information for figure 6.34

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Area fixed effect mean	5.01	5.00	5.00	5.00	5.00
Area fixed effect var.	0.04	0.05	0.05	0.05	0.05
Area fixed effect std. error mean	0.207	0.215	0.219	0.223	0.227
Area fixed effect std. error var.	0.002	0.002	0.002	0.002	0.002

Figure 6.35: Relationship between the household-level fixed effect estimation and the average number of individuals per household, for the two-level model in scenario D

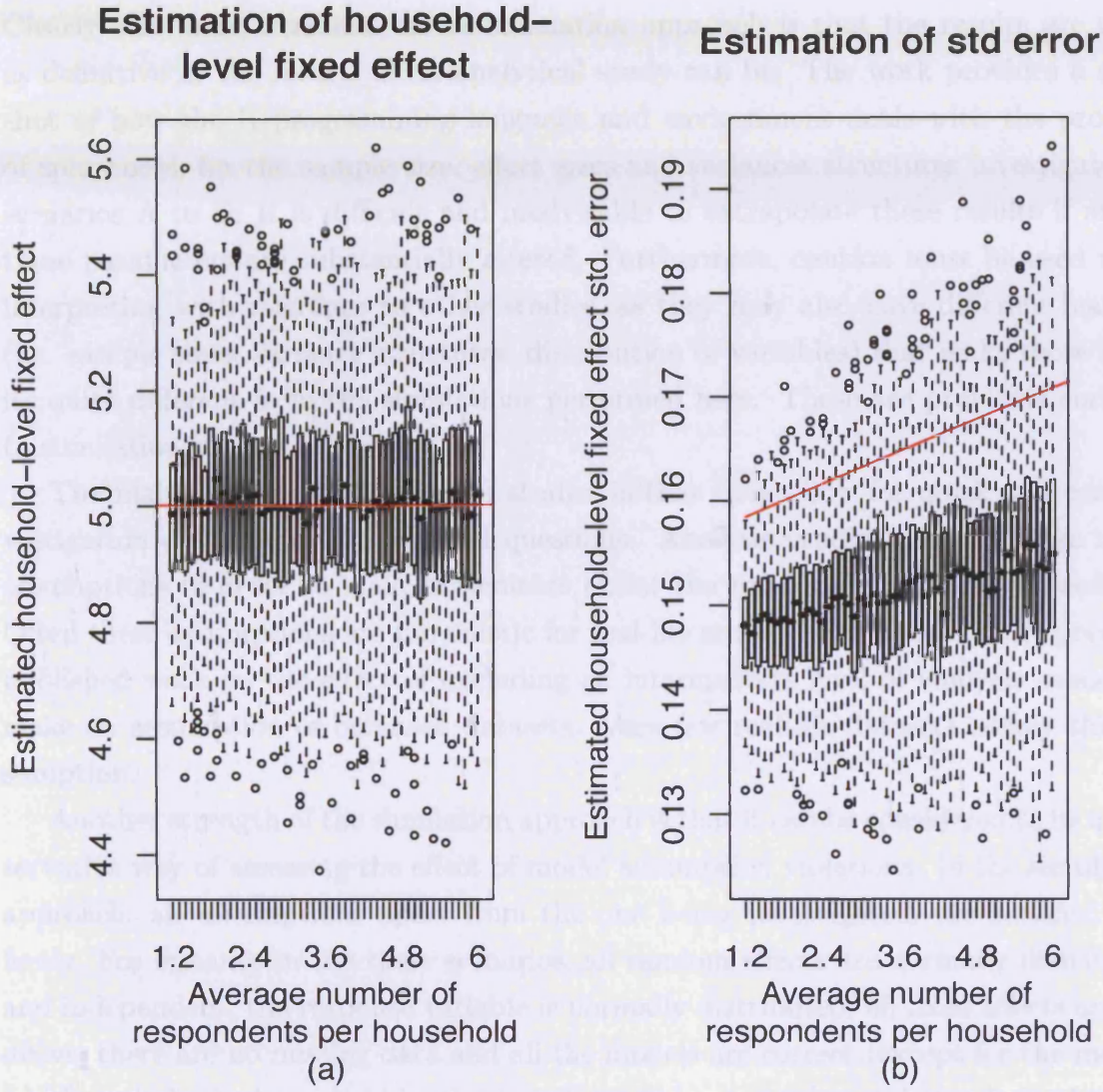


Table 6.39: Summary information for figure 6.35

Average per household	1.05-2	2.05-3	3.05-4	4.05-5	5.05-6
Household fixed effect mean	5.00	5.00	5.01	5.01	5.00
Household fixed effect var.	0.02	0.03	0.03	0.03	0.03
Household fixed effect std. error mean	0.1486	0.1498	0.1513	0.1528	0.1541
Household fixed effect std. error var.	0.0000	0.0000	0.0001	0.0001	0.0001

6.5 Discussion

6.5.1 Strengths and limitations of the simulation study

The simulation approach taken in this chapter has both strengths and weaknesses. Clearly, the main weakness of the simulation approach is that the results are never as definitive as the results of an analytical study can be. The work provides a snapshot of how the R programming language and environment deals with the problem of sparseness, for the sample size, effect sizes and variances structures investigated in scenarios A to D. It is difficult and inadvisable to extrapolate these results if any of these parameters are substantially altered. Furthermore, caution must be used when interpreting with reference to other studies, as they may also have different features (i.e. sample sizes, variance structures, distribution of variables) that make those studies quite different from the simulations performed here. These are problems endemic to simulation studies.

The main strength of simulation studies is that they allow for quick and easy investigation of intractable analytical questions. Analytic studies need to make many assumptions in order to make statements about the results of modelling procedures. Often these assumptions are unrealistic for real-life studies. For instance, the previous published work on the effect of excluding an intermediate level of analysis needed to make an assumption of balanced datasets. Very few real-life datasets satisfy this assumption.

Another strength of the simulation approach is that it can be considered to be a conservative way of assessing the effect of model assumption violations. In the simulation approach, all assumptions apart from the one being investigated, are satisfied perfectly. For instance in the these scenarios, all random effects are normally distributed and independent, the response variable is normally distributed, all fixed effects are additive, there are no missing data and all the models are correct (except for the models which exclude the household level, since this is the assumption violation being investigated). If the model fitting procedure in R performs poorly in such a situation, there is no reason to expect it to perform better in real-life datasets which may have skewed distributions or non-independent random effects or mis-specified models.

6.5.2 Implications for the results of previously published studies

In order to make statements about the implications this work has for previously published work it is important to recognise the limitations of the study listed in section 6.5.1. Assuming however, that the true variance structures and true variable distributions in the BHPS are similar to those modelled in this simulation study, there are a

number of points to make. Table 6.2 summarises the literature. The smallest sample size for any of these studies is 7,047 while the largest is 10,264. The household-level ICCs from the literature range between 0.09-0.29. The simulation scenario that most closely matches these parameters is scenario B. Here the sample size is around 10,000 and the household level ICC is 0.068. The average number of individuals per household from the literature ranges between 1.61 and 1.89. Those simulations from scenario B with an average number of individuals between (and including) 1.6 and 1.9 were extracted. ICC estimates at this level of sparseness were centred close to the true ICC values for all three levels, as shown in figure 6.36. Single hierarchy ICC estimates have a relatively large range however with the maximum household ICC reported being 0.113 and the minimum 0.015. For instance, the household-level ICCs range between 0.015 (underestimating the true ICC by almost 80%) and 0.113 (overestimating the true ICC by 66%). While these are extreme ICCs, it is still important to note that 25% of the simulated hierarchies underestimate the true household ICC by at least 13% while another 25% overestimate it by at least 13%.

Using the same subset of scenario B (those with average number of respondents per household between 1.6 and 1.9), fixed effect estimation is now examined. Firstly, the area-level fixed effect is investigated in figure 6.37. As can be seen, both the area-level fixed effect and its standard error are centred about the true mean.

The household-level fixed effect is now examined. The fixed effect is unbiasedly estimated, but its standard error is consistently underestimated. This implies that some of the significant relationships found by the literature may in fact be due to the effect of household-level sparseness.

6.6 Conclusion

This chapter has investigated the impact of sparseness at an intermediate level in a multilevel analysis in four different settings: firstly where the intermediate level contributes an equal variance component to the lowest level, the second where both the area and household levels are dominated by a large individual level, thirdly where the intermediate level is completely uninformative, and finally the second scenario was modelled with a smaller sample size. The results of each of these simulations is summarised in tables 6.43 and 6.44. Table 6.43 summarises the results when the sparseness is high (less than 1.5 per household on average), while table 6.44 summarises the results when the sparseness is low (more than 5.5 per household on average). These tables summarise the estimates from each simulation in terms of mean squared errors. These are calculated as in equation 6.8, where θ is the parameter to be estimated.

$$\mathrm{MSE}(\hat{\theta}) = \mathrm{E}\big((\hat{\theta} - \theta)^2\big) \tag{6.8}$$

Figure 6.36: ICC estimation for simulations with an average number of individuals between 1.6 and 1.9, for comparison with literature

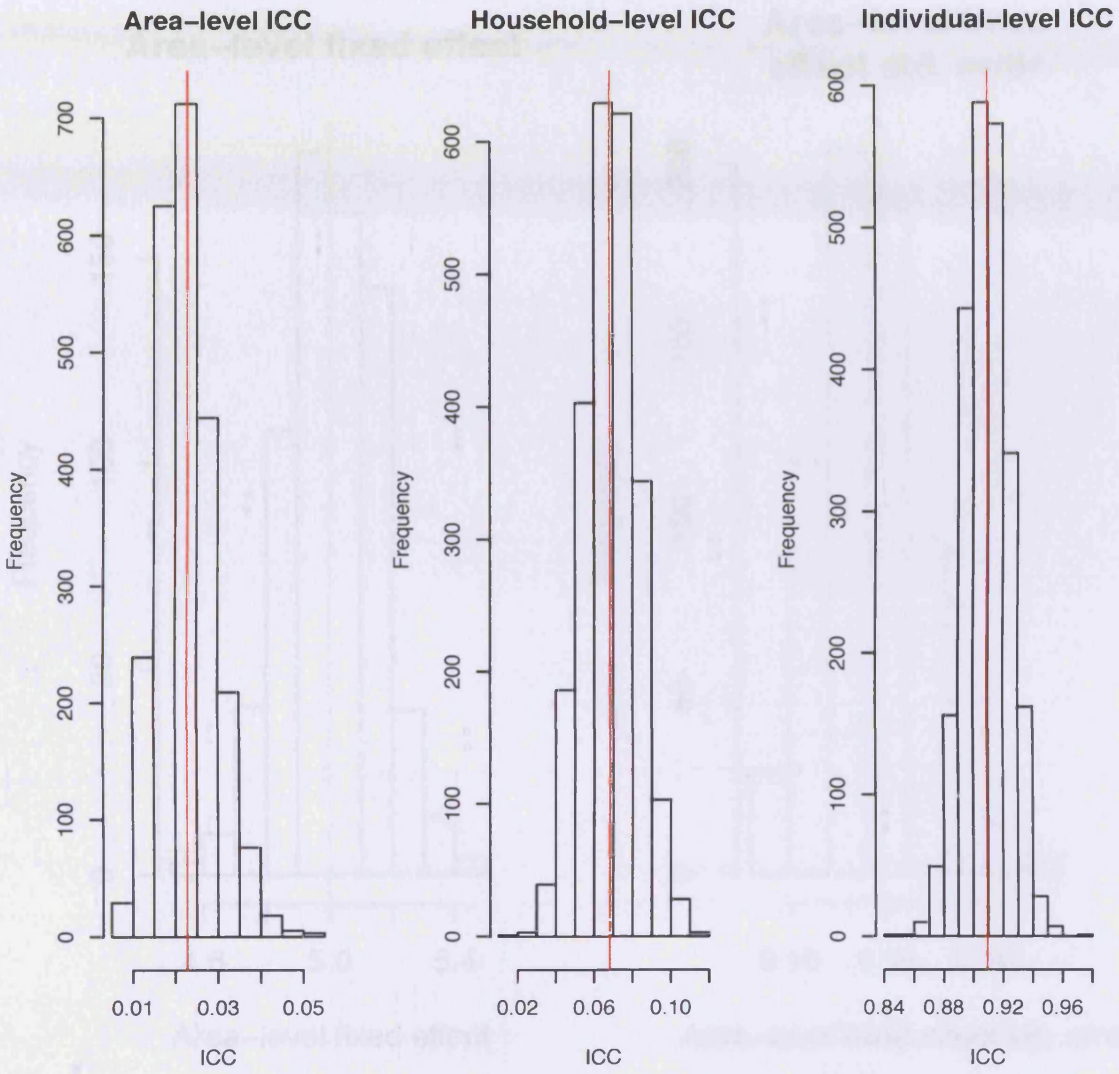


Table 6.40: Summary information for figure 6.36

	Area-level	Household-level	Individual-level
Lower quartile	0.018	0.059	0.899
Upper quartile	0.027	0.077	0.919
Mean	0.02249	0.06849	0.90902
Variance	0.00004	0.00019	0.00022

Figure 6.37: Area-level fixed effect estimation for simulations with an average number of individuals between 1.6 and 1.9, for comparison with literature

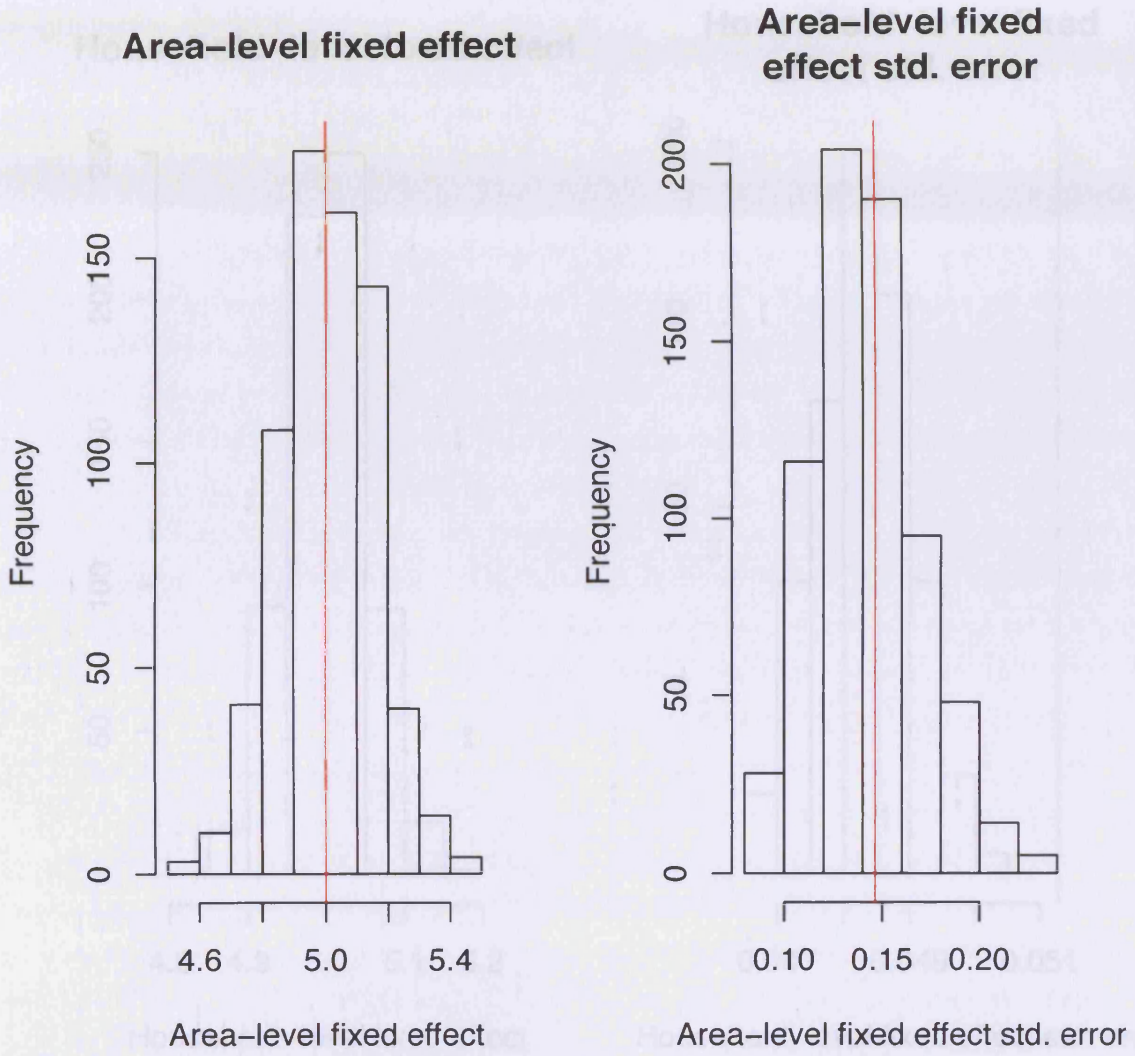


Table 6.41: Summary information for figure 6.37

	Area-level fixed effect	Area-level fixed effect
Lower quartile	4.910	0.125
Upper quartile	5.111	0.159
Mean	5.009	0.143
Variance	0.022	0.001

Figure 6.38: Household-level fixed effect estimation for simulations with an average number of individuals between 1.6 and 1.9, for comparison with literature

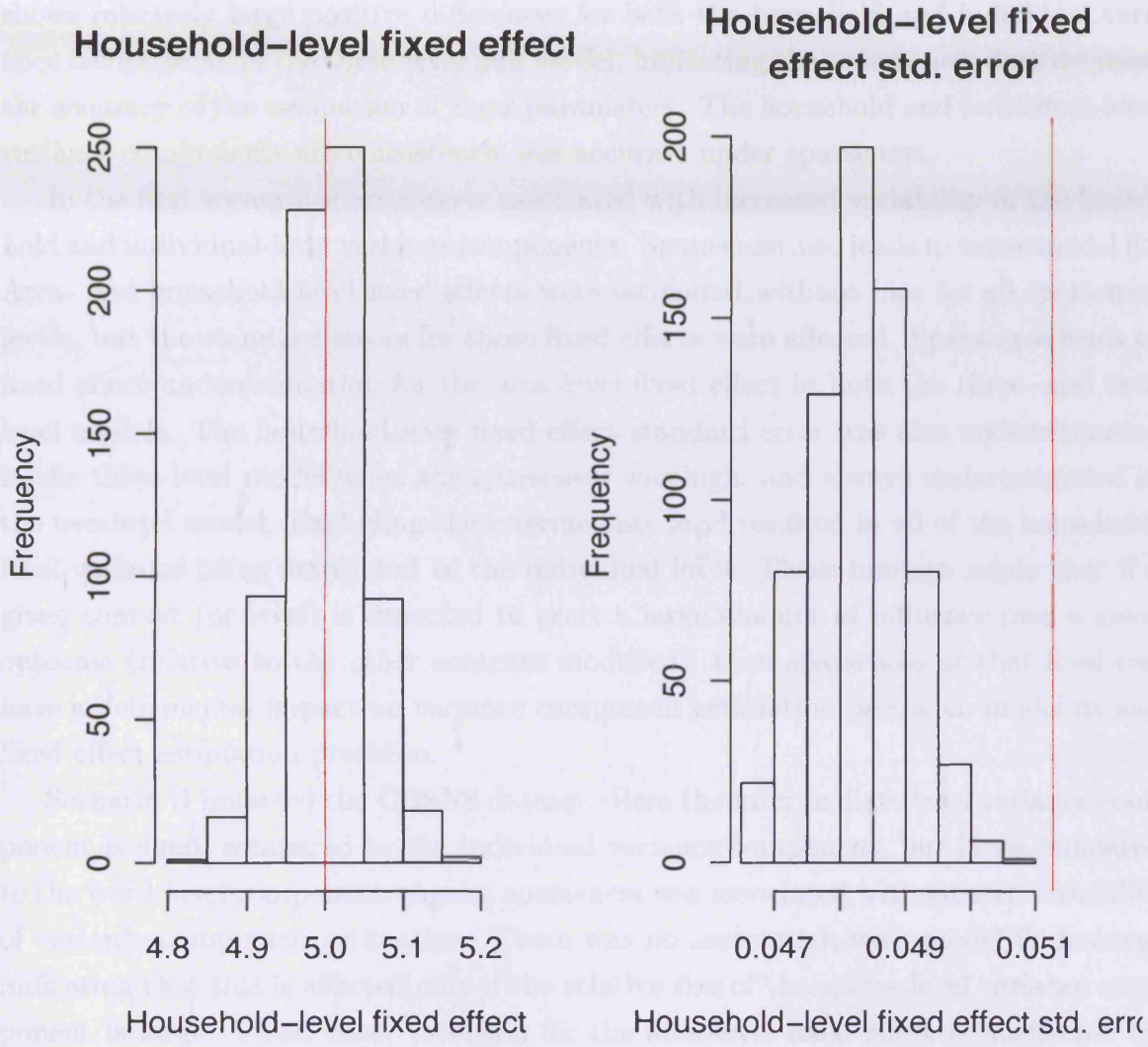


Table 6.42: Summary information for figure 6.38

	Household-level fixed effect	Household-level fixed effect
Lower quartile	4.968	0.048
Upper quartile	5.035	0.049
Mean	5.002	0.048
Variance	0.003	0.000

Since these tables are best examined relative to one another, the difference between the two tables is provided in table 6.45. These differences are calculated by subtracting table 6.43 from table 6.44, so that positive entries indicate that the estimates calculated under sparse conditions are less accurate than their non-sparse equivalents. This table shows relatively large positive differences for both the household and individual variance components in the three-level null model, indicating that sparseness does decrease the accuracy of the estimation of these parameters. The household and individual-level variance components are consistently less accurate under sparseness.

In the first scenario sparseness is associated with increased variability in the household and individual-level variance components. Sparseness also leads to worse model fit. Area- and household-level fixed effects were estimated without bias for all sparseness levels, but the standard errors for those fixed effects were affected. Sparseness leads to fixed effect underestimation for the area-level fixed effect in both the three- and two-level models. The household-level fixed effect standard error was also underestimated in the three-level model when the sparseness was high, and always underestimated in the two-level model. Excluding the intermediate level resulted in all of the household-level variance being attributed to the individual level. These findings imply that if a given context (or level) is expected to exert a large amount of influence over a given outcome (relative to the other contexts modelled), then sparseness at that level can have a detrimental impact on variance component estimation precision, model fit and fixed effect estimation precision.

Scenario B imitated the CHSNS dataset. Here the intermediate level variance component is small compared to the individual variance component, but large compared to the ward-level component. Again, sparseness was associated with greater variability of variance component estimation. There was no association with model fit however indicating that this is affected only if the relative size of the sparse level variance component is large. Fixed effect precision for the area-level fixed effect is unaffected by sparseness, but is perhaps slightly underestimated. Household-level fixed effects however are underestimated in the three-level model when the average per household is less than three, and underestimated for all levels of sparseness in the two-level model. Excluding the sparse level results in the variance attributable to that level being transferred to the individual level. So, even when the sparse level is expected to have a small influence, then including it can lead to inflated individual-level variance components

Scenario C demonstrated that including an uninformative level is not a large concern for variance component estimation if sparseness at that level is not extreme. Since the true ICC is zero the household-level ICC cannot be underestimated. The majority of the household-level ICCs are close to zero, however when the sparseness is extreme the household-level ICC may be seriously overestimated and results in spurious importance being placed on that level. For example, the largest household-level ICC

Table 6.43: Comparison of mean squared error for selected estimates across all four simulation scenarios, when the the sparseness is high (average number of respondents per household is less than or equal to 1.5)

	3-level null model			2-level null model		Fixed effect model	
Scenario	Area V.C.[a]	Household V.C.	Individual V.C.	Area V.C.	Individual V.C.	Coefficient	Std. Err.
A	0.150	0.462	0.400	0.152	9.981	0.144	0.002
B	0.146	0.548	0.599	0.147	1.531	0.146	0.003
C	0.150	0.416	0.498	0.150	0.279	0.134	0.017
D	0.297	1.453	1.669	0.298	1.762	0.204	0.092

[a]V.C. stands for variance component

Table 6.44: Comparison of mean squared error for selected estimates across all four simulation scenarios, when the the sparseness is low (average number of respondents per household is greater than 1.5)

	3-level null model			2-level null model		Fixed effect model	
Scenario	Area V.C.[a]	Household V.C.	Individual V.C.	Area V.C.	Individual V.C.	Coefficient	Std. Err.
A	0.186	0.409	0.153	0.254	9.849	0.158	0
B	0.155	0.177	0.311	0.158	1.513	0.146	0
C	0.144	0.085	0.289	0.144	0.277	0.139	0
D	0.374	0.612	0.984	0.463	1.609	0.234	0

[a]V.C. stands for variance component

Table 6.45: Difference between table 6.43 and table 6.44

	3-level null model			2-level null model		Fixed effect model	
Scenario	Area V.C.[a]	Household V.C.	Individual V.C.	Area V.C.	Individual V.C.	Coefficient	Std. Err.
A	−0.036	0.053	0.247	−0.102	0.132	−0.014	0.002
B	−0.009	0.371	0.288	−0.011	0.018	0.000	0.003
C	0.006	0.331	0.209	0.006	0.002	−0.005	0.017
D	−0.077	0.841	0.685	−0.165	0.153	−0.030	0.092

[a]V.C. stands for variance component

reported for this scenario is as high as 0.15. Caution should be exercised therefore when choosing contexts to model.

Scenario D revisited scenario B, but used a much smaller sample size. The main finding from this simulation is that variance component estimation is much more variable with a smaller sample size. In the situation where the middle level is excluded the variability from that level is distributed between the adjacent levels. Fixed effect precision for this scenario was similar to that observed in scenario B, except the relationships between fixed effect standard errors were more pronounced due to the small sample size. The area-level fixed effect standard errors were largely unaffected by sparseness. The household-level fixed effects however were underestimated when the average per household was less than 3, and always underestimated in the two-level model.

All of this implies that variance component estimation in models where sparseness is an issue is dependent on the relative variance contribution of the sparse level, the degree of sparseness present and the total sample size.

Chapter 7

Derivation of synthetic boundaries and comparison with administrative boundaries

7.1 Introduction

This chapter will introduce, describe and demonstrate the synthetic boundary algorithm, before using the algorithm in a large simulation study to compare the operationalisation of administrative and synthetic boundaries.

When investigating the effects of area-level exposures of any kind on some individual outcome, the natural method of analysis is a hierarchical model. As explained in chapter 5, this is due to the fact that individuals who live in close proximity to one another are likely to be more similar that individuals who live far away from each other, violating the independence assumption implicit in ordinary least squares regression. To employ hierarchical methods, one must decide on the contexts (or levels) to include in the hierarchy. Ideally, this choice will be based upon some underlying theory or hypothesis. However, the impact of the choice of contexts to include as levels in a hierarchical analysis has not been fully addressed. A comprehensive approach would be to include all possible contexts which may impact upon the outcome of interest as levels. Ideally these levels would group people together based on homogeneity of exposure, perhaps using the concept of commonality of living space. A perfect hierarchy would ensure that everyone in a given group would be exposed to exactly the same neighbourhood influences. As well as this the hierarchy would also group people together based upon homogeneity of area-level confounding variables. These confounders could then be satisfactorily controlled for in an analysis. It is difficult to envisage any contiguous boundaries capable of achieving these ideals. The problem is compounded by the fact that "commonality of living space" is ill-defined. Does living space comprise

the building where one lives, or should it include the street outside or even incorporate where one works, shops or goes to relax? Complex, non-nested hierarchies could perhaps capture all of these contexts, however the amount of information required is not available in this study, nor any study to date.

In practice, however, these ideals can never be achieved. As mentioned previously the majority of published studies of mental health and context have used administratively defined areas to act as proxies for neighbourhoods. The reasons to use administrative boundaries are numerous. Firstly, there is often no choice but to use them. This is usually the case if the data are obtained from a governmental body (such as the ONS) that routinely aggregate small area statistics up to larger administrative areas for the purposes of confidentiality. Under such circumstances the researcher is left with little choice but to use administrative boundaries for their analysis. Also, in many situations administrative areas are the easiest boundaries to use. They are well-defined, contiguous areas that provide a partition for any area of interest. Another advantage of administrative boundaries is the fact that much of the information necessary for area-level modelling (e.g. shape files for GIS analyses, census information etc...) are freely available to everyone, facilitating comparisons between studies. Finally, administrative boundaries may be linked to the research question itself, resulting in them being the only correct hierarchy to model. Research questions concerning the effect of regional laws, or the impact of local councillors or elected officials, may find administrative boundaries to be the natural context to model.

However, administrative areas are often not intrinsic to the study question and in such situations they may not be the most apposite hierarchy to employ. Many administrative areas cover regions which contain extremes of affluence and poverty. In Caerphilly, the electoral ward St. James in the south of Caerphilly county borough (see figure 2.3) is one such region. Its boundaries encompass areas of great affluence as well as areas of high deprivation. In this situation an area-level aggregate variable based on this administrative boundary may not convey any meaningful information. More generally, there is evidence to indicate that the most deprived people do not necessarily live in the most deprived areas (Joshi et al., 2001; McLoone, 2001). This poses a problem for area based targeting of resources. Other authors have theorised that administrative boundaries do not operationalise "neighbourhoods" and so are counterproductive to the investigation of neighbourhoods and health (Rice et al., 1998; Macintyre et al., 2002; Diez-Roux, 2003). This chapter will propose creating new areas for hierarchical modelling.

The way to proceed is not straightforward. The theory behind the proposed solution to this problem is that the boundaries employed in a hierarchical analysis should be delineated with the goal of grouping similar people together. Grouping similar people together serves a two-fold purpose. Firstly, it increases the validity of using aggregate

area statistics, since they are more likely to be representative of the area's residents. Secondly, if the residents of an area are homogenous for various socioeconomic and demographic variables (such as social class, household income, council tax bands and educational achievement) it seems more likely that they will be homogenous for other, perhaps unmeasured, variables.

The need for homogenous areas in environment and health studies is also addressed in a paper by Cockings and Martin (Cockings & Martin, 2005), who cite internal homogeneity as a desirable quality for areas to have in studies concerned with hypothesis testing or visualisation of spatial patterns of disease. There have been a number of attempts to create or define "neighbourhood"-sized homogenous areas for similar purposes (Sampson et al., 2002; Propper et al., 2005; Galster, 2001; Assunção et al., 2006; Vickers & Rees, 2007). There are two basic approaches possible. Firstly, large areas can be built up from small areas (or individuals) by combining similar people or areas together. These are usually combined so that they are homogenous, contiguous and possibly of a certain size. The second method takes the opposite approach and is known as "wombling" (Womble, 1951). Wombling involves finding areas that are close together but are substantially different in composition and ensures that they are assigned to different groups. The method presented in this chapter takes the former approach, building large homogenous areas from 2001 census output areas.

So, to summarise, the basic theory of the work is that the best way to automate the production of meaningful area boundaries is to use area homogeneity to delineate areas. If areas are internally homogenous then the ecological fallacy is rendered powerless. Higher level covariates will then provide more meaningful information about the areas to which they correspond. This enables a more objective analysis of the relative importance each context holds in determining the mental health status of an individual to be made.

7.2 Method

Under ideal circumstances individual-level geographical information would be used to create small unit postcode sized areas. As discussed in chapter 5 many hierarchical studies have attributed modest area effect sizes to the fact that the contexts modelled were too large and heterogenous. The geographical position of each individual household would be required for such an undertaking, however, confidentiality issues preclude such information being recorded in the questionnaire survey, and even if it were possible, linking such individual-level information to other datasets may not be possible. Moreover, without complete enumeration of areas (such as in the census) this approach still suffers from the problem that those surveyed may not be representative of those not included in the survey. A different approach is used in this thesis. This

approach involves taking small administrative boundaries (2001 census output areas) and merging these into larger areas, based on the goal of making these large areas as homogenous as possible. These new areas will be referred to as "synthetic areas" henceforth. It should also be noted that the description of the algorithm refers to these areas as OAs, but the algorithm is generalisable and can be applied to any set of areas.

This method relies heavily on being able to compare the residents of one output area with another output area (or set of output areas) and determining how similar they are. The two areas in question may have widely different numbers of residents and the variables to be homogenised may be ratio (e.g. age), ordinal (e.g. social class) or nominal (e.g. gender). Some of the interval variables are highly discretised. For instance, to increase survey response the household income question was phrased as a multiple choice question, with respondents asked to indicate whether their annual household income is less then £5,000, between £5,000 and £11,250 or more than £11,250. This avoided participants having to divulge detailed information on their finances, however it also resulted in a variable with only three possible answers. With such a wide range of variable types it was decided to employ a non-parametric method. The measure chosen was the Mann-Whitney (or Wilcoxon) U statistic. The formulation of this statistic is very intuitive and it provides a dimensionless measure of similarity for two groups based on the ranks of the data. It is also very quick to calculate making it a very useful measure. It is constructed as follows. Consider two areas, A B. Let A and B contain n and m measurements respectively. Compare each measurement from A with each measurement from B. If measurement A_i is less than B_j then let U_{ij} equal 1. If A_i is equal to B_j let U_{ij} equal 0.5 and if A_i is greater than B_j let U_{ij} equal zero. Sum all the U_{ij}s and divide by the maximum possible score, $m \times n$ as in equation 7.1.

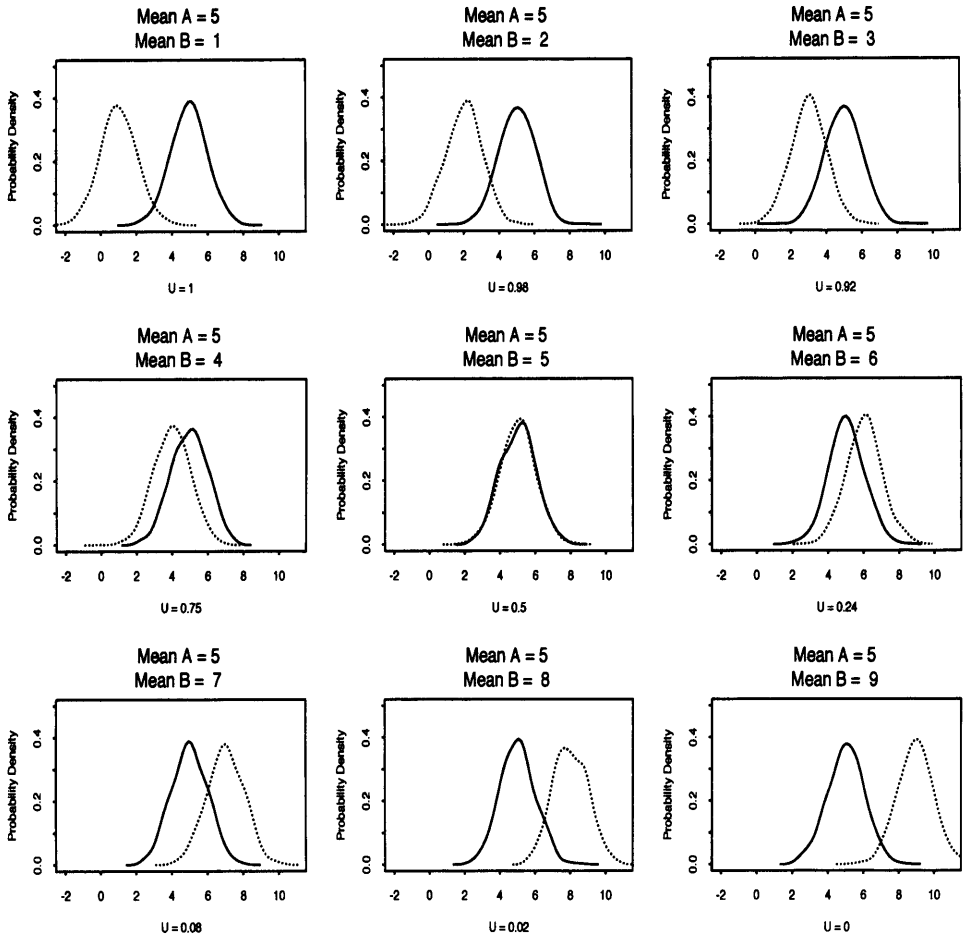
$$
\begin{aligned}
A_i < B_j &\Rightarrow U_{ij} = 1 \\
A_i = B_j &\Rightarrow U_{ij} = \frac{1}{2} \\
A_i > B_j &\Rightarrow U_{ij} = 0
\end{aligned}
$$

$$
U = \frac{\sum_{i=1}^{n} \sum_{j=1}^{m} U_{ij}}{m \times n} \tag{7.1}
$$

This results in a score ranging between 0 and 1. If area A is compared with itself (or any other identical area) the U statistic is 0.5. If every measurement in A is greater than every measurement in B the U statistic will equal 0, while if the reverse is true the U statistic will equal 1. Obviously, for the U statistic to be applicable it is necessary that the measurements be at least ordinal. Dichotomous nominal variables can also be examined, using an arbitrary ordering. This U statistic gives a measure of how close any two areas are in terms of any given variable. Figure 7.1 provides an illustration of

how the U statistic varies as the populations being compared change. Here two popu-

Figure 7.1: Illustration of the Mann-Whitney U statistic



lations are compared using the U statistic. Population A is normally distributed with mean of 5 and standard deviation 1 and is depicted as a solid line, while population B is normally distributed with mean ranging from 1 to 9, as indicated in the title of each individual plot, and a standard deviation of 1. In each case a sample of 1,000 was taken from both populations and compared using the U statistic. The U statistic effect size is recorded underneath each plot. The U statistic lies far away from 0.5 when the two populations are distinct, but is exactly 0.5 when the two populations are identically distributed. Each population is based on 1000 observations. A shift in the mean of the population of one standard deviation (in this case 1) results in a U statistic that lies approximately 0.25 away from 0.5.

The decision to use the Mann-Whitney U statistic was pragmatic, being based on simplicity, applicability and expediency. There are other measures that could have been used however. To the authors knowledge this measure has not been used to derive area boundaries previously.

The variables chosen to homogenise on for this study were educational achieve-

ment, social class, council tax valuation band, tenure, gross household income and employment category. These variables represent both individual and household-level attributes and are all socioeconomic or demographic variables, except for council tax band, which has been shown to be a useful measure of individual socioeconomic status (Beale et al., 2000; Fone et al., 2006b). These variables will be referred to as the "homogenising variables" henceforth. The algorithm requires four different types of information to output a set of synthetic boundaries, and these shall be addressed in turn.

Spatial information

Firstly, the algorithm requires information about the spatial distribution of areas, so that it can ensure that the boundaries produced represent contiguous areas. All that is needed to ensure contiguity is an adjacency matrix detailing which output areas are adjacent. A 2001 census shape file for the UK was obtained from the Office for National Statistics. From this, the boundaries for OAs in Caerphilly county borough were extracted using Arcview. This information was saved in table form, and R (R Development Core Team, 2006) was used to obtain the adjacency matrix.

Compositional information

The algorithm uses information about the composition of small administrative areas (OAs) in order to determine which OAs should belong together in a larger synthetic area. Information regarding the six homogenising variables was used in order to do this.

Initial merging threshold

In order to begin the algorithm needs information on how similar areas must be before they are included in the same synthetic area (or merged). This is called the initial merging threshold and specifies how close to 0.5 the Wilcoxon effect size between two adjacent OAs must be, before they are merged. This is specified at the beginning of the algorithm and then this threshold is relaxed until all of the OAs are assigned.

Initial seed pairs

The final type of information required by the algorithm can be thought of as initial values for the synthetic areas. The algorithm begins with a user-specified number of pairings of OAs. These are called "seed" pairs since the synthetic areas "grow" from them. These can be chosen in any way (provided that the pairings are adjacent).

The seed pairs were chosen so that the largest Wilcoxon effect size (for all six merging thresholds) was below a certain threshold. This was done by constructing a large matrix with 559 columns and 559 rows (559 being the number of output areas) for each of the six homogenising variables. Calling these matrices M1,...,M6, the $M1_{i,j}$ entry, equals the Wilcoxon effect size between the i^{th} OA and the j^{th} OA for the first homogenising variable, if the two OAs are adjacent, and is equal to zero otherwise. Entries in these six matrices were used to determine which pairs of OAs had Wilcoxon effect sizes below a given threshold for the homogenising variables. These pairs of OAs were then used as initial seed pairs. The rationale for using this approach is to ensure that the initial seed pairs represent the most similar pairings possible. Since the primary interest is in comparing the administrative boundaries with the synthetic boundaries, it is desirable that both types of boundaries share equal numbers of highest-level areas. In the 2001 census, there are 110 lower super output areas (as described in chapter 2 and so 130 initial seed pairs were chosen. The algorithm will now be presented.

7.2.1 Algorithm

A description of the algorithm will now be presented. Figure 7.2 illustrates these steps.

1. U statistics are calculated for all pairs of adjacent output areas for all six homogenising variables $(U_{1ij}, U_{2ij}, \ldots, U_{6ij})$. This provides a measure of distance between each pair of adjacent OAs in terms of the homogenising variables. The pairs of OAs resulting in U statistics that all lie within a user-specified distance from 0.5 (i.e. the most similar adjacent OAs) are selected. Each of these pairs represent a "seed" from which the synthetic areas are created. These can be thought of as initial values that need to be input into the algorithm. The user decides how many seeds to input to the algorithm. Since the algorithm has the capacity to remove synthetic areas completely, but not to create them, the resultant number of synthetic areas is likely to be less than the original number of seeds.

2. The second step of the algorithm is the "inclusion" step. Each seed is examined in turn. The order in which the seeds are examined is random, and this means that different runs of the algorithm will produce different sets of boundaries. All OAs adjacent to the seed are identified (using an adjacency matrix). U statistics comparing each of these OAs and the seed (which may comprise many OAs) are calculated for each of the homogenising variables. A merging threshold, denoted by θ, is used to determine whether these U statistics are sufficiently close to 0.5 to warrant merging. If five of the six U statistics lie within the range $[0.5-\theta, 0.5+\theta]$, the adjacent OA is merged with the seed (i.e. the U statistics satisfy the merging

criterion). Merging occurs if at least five of the six U statistics lies within the user-specified radius of 0.5. If more than one of the six U statistics lies outside the merging threshold then the OA in question remains unassigned to the seed pair. This OA is thus eligible to be "recruited" by any other seed to which it is adjacent. OAs that belong to a seed are not eligible to be recruited by a different seed. The algorithm continues in this fashion until it has examined each unrecruited OA that was adjacent to the initial seed.

3. The third step is the exclusion step. This step is a further precaution to ensure the internal homogeneity of the new areas as they grow larger and larger. At this step the "fledgling" synthetic areas are examined in turn. Each constituent OA in a synthetic area is compared with the entire synthetic area less itself. For example, if at this stage a synthetic area consisted of ten OAs, then each of these OAs would be compared to the synthetic area created by the other nine OAs. Again, all six U statistics are computed, and if more than one of them lies outside the user-specified range the OA is removed from the synthetic area. This OA is now eligible to be recruited by any other synthetic area (or indeed the same synthetic area, should its composition change sufficiently in subsequent iterations to re-allow its inclusion). This step is in place to allow for the fact that the criterion inclusion into a synthetic area is not transitive, i.e. using the approximate equivalence sign ($\approx$) to denote U-statistics which are sufficiently similar to allow merging, if we have three areas A, B and C, then $A \approx B$ and $B \approx C$ does not imply that $A \approx C$.

4. Next the algorithm makes a check on the contiguity of the synthetic areas. The exclusion step of the algorithm permits the creation of "islands" consisting of single OAs belonging to a synthetic area, but not lying adjacent to any of its constituent OAs. If the exclusion of an OA at the third step isolates another OA from the synthetic area, this second OA should also be removed, preserving contiguity. This "island removing" step is also invoked if a synthetic area consists of only one OA. This removes the possibility of single OA synthetic areas being produced by the algorithm.

5. At this point the algorithm switches back to step two and examines all of the OAs that are now adjacent to each synthetic area. Steps two and three are cycled between until equilibrium is attained. It is difficult to determine conclusively when equilibrium has been achieved, hence an equivalent proxy is monitored. This proxy is the number of OAs left unassigned after the i^{th} U statistic comparison (N_i). Equilibrium is deemed to have been achieved if either the variability of the last twenty N_is is zero, or if the last 200 comparisons merely cycle between

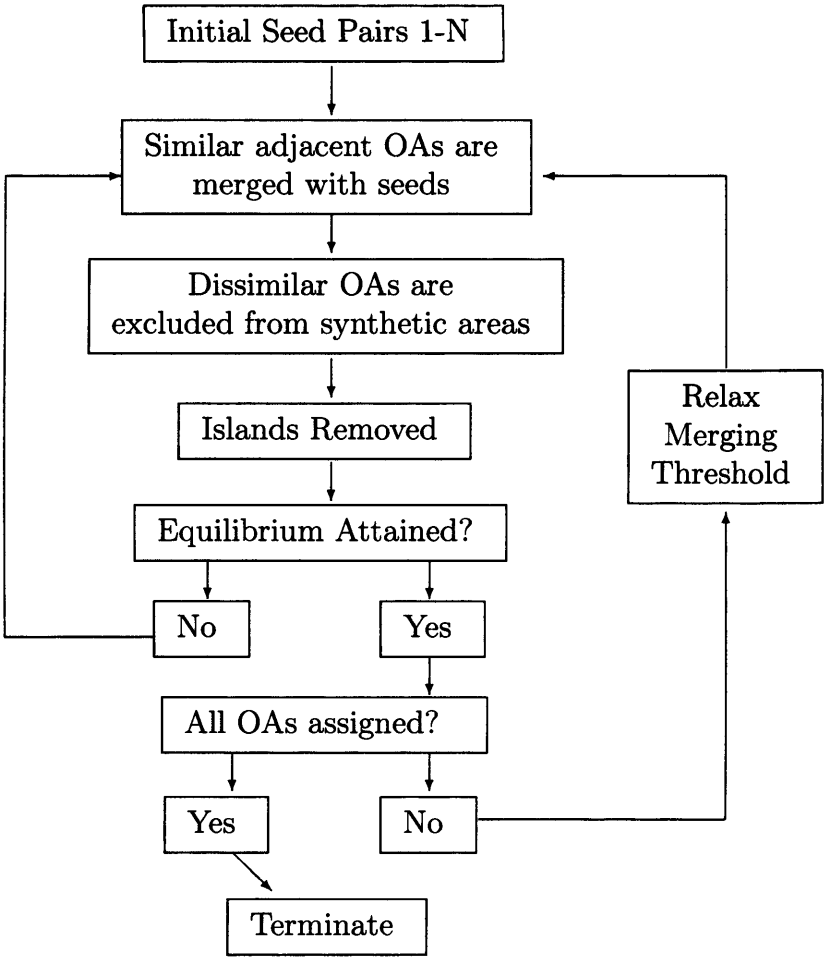
10 or less distinct values of N_i. The first criterion is satisfied if and only if $N_{i-19} = N_{i-18} = \ldots = N_i$, indicating that the last 20 cycles of steps two and three have failed to yield any mergings. The second criterion is satisfied if the number of OAs unassigned after the last 200 cycles of step one and two have belonged to an infinite loop of length 10 or less. If the merging criterion is strict, equilibrium may be obtained before the majority of OAs are recruited. If the merging criterion is lenient the algorithm may run until all of the OAs are assigned to a synthetic area. Typically the merging criterion is set to be strict at the start. It cannot be set too tight however, since an extremely tight criterion may be too strict for even the initial seed pairs to satisfy. In this scenario, the inclusion step may result in no OAs being recruited and the exclusion step may remove all of the seed pairs. With no seed pairs left, the algorithm cannot assign any OAs to any synthetic area and it will terminate with all the OAs unassigned. On the other hand, if the merging criterion is set to be too lenient, then any OA adjacent to a seed pair will satisfy the inclusion criterion. In the extreme case where the merging criterion is set to the maximum (0.5), so that all U statistics result in merging, the only criterion necessary for merging is adjacency. As a consequence the synthetic areas produced will not be based on internal homogeneity but rather spatial proximity. There is a tradeoff to be made then between selecting a merging criterion strict enough to result in meaningful synthetic areas, while being permissive enough to allow the synthetic areas to grow.

6. Once equilibrium has been attained a check is made on N_i to determine how many OAs are left unassigned. If there are any OAs unassigned (i.e. if $N_i > 0$), then the merging criterion is relaxed by a user-specified increment and the algorithm recommences with step two. This process continues until all of the OAs are recruited.

7.3 Result of the synthetic boundary algorithm

This section illustrates the type of output the algorithm can produce. Figure 7.3 shows the progression of the algorithm as it attempts to partition Caerphilly county borough into contiguous and internally homogenous areas. To illustrate that the algorithm works with any areas, it will be demonstrated using 1991 census EDs. These are also larger than OAs and so are better suited for visual inspection. Here enumeration districts are used as the building block instead of OAs. For the purposes of illustration, the number of initial seeds was four. These are shown by the coloured EDs in the first map of Caerphilly, with all other EDs remaining white. As the algorithm progresses these seed pairs grow, until all of the EDs in the borough are assigned to a synthetic

Figure 7.2: Flowchart of the synthetic boundaries algorithm



area. As can be seen in the final map of Caerphilly, the synthetic areas produced are contiguous. It is also obvious that the synthetic area sizes vary considerably, with the synthetic area number 1 taking up most of the borough, and the synthetic area number 4 taking up a very small number of EDs. In order to demonstrate how the algorithm works in practice a larger number of initial seed pairs can be chosen. Figure 7.4 compares 1991 electoral wards with a set of synthetic boundaries produced from 50 initial seed pairs. There are 36 1991 census electoral wards, and 36 synthetic areas produced. This set of synthetic boundaries was produced essentially by trial and error. Fifty seed pairs were input to the algorithm and many boundaries produced. One of the boundaries that comprised 36 areas was selected. This is one way the algorithm could be used to create a single set of boundaries. There is no guarantee, however, that any set of boundaries produced by the algorithm is in any sense optimal. Instead, many sets of boundaries can be created and compared, and the "best" (whatever that may mean in a given setting) chosen.

7.4 Critique

The major drawback of this method is a consequence of the data itself. Since individual geographical positioning is not available the synthetic areas are created using output areas as building blocks. Output areas contain 300 individuals on average. As such, they may be large enough to be internally heterogenous. This hampers the homogenising process and provides a limiting factor for how homogenous the synthetic areas can be. From a homogenisation point of view, a smaller unit such as unit postcode would be preferable. As described earlier however, this approach would have its own difficulties. One drawback to using unit postcodes is that information at such a small level is difficult to obtain due to confidentiality issues. Moreover, if it is obtainable, such data are often randomly perturbed to prevent the identification of individuals. Conversely, information at OA-level is more freely available and is not usually subjected to such data protection techniques.

Secondly, the algorithm takes some time to run. For each set of synthetic boundaries created, many thousands of comparisons are made between OAs and synthetic areas. As mentioned earlier it takes approximately eight minutes to create a single set of synthetic boundaries. Since each set of boundaries is a product of the merging threshold as well as the initial seeds, many different sets of boundaries need to be created in order to assess the influence of each of these criteria. The computing time required for such an exercise is considerable. Once again the Condor parallel computing system (Litzkow et al., 1988) was used in order to facilitate the creation of many thousands of boundaries. Fitting models including these new boundaries was also time-consuming, and this was also done using Condor.

Figure 7.3: Illustration of the algorithm in progress

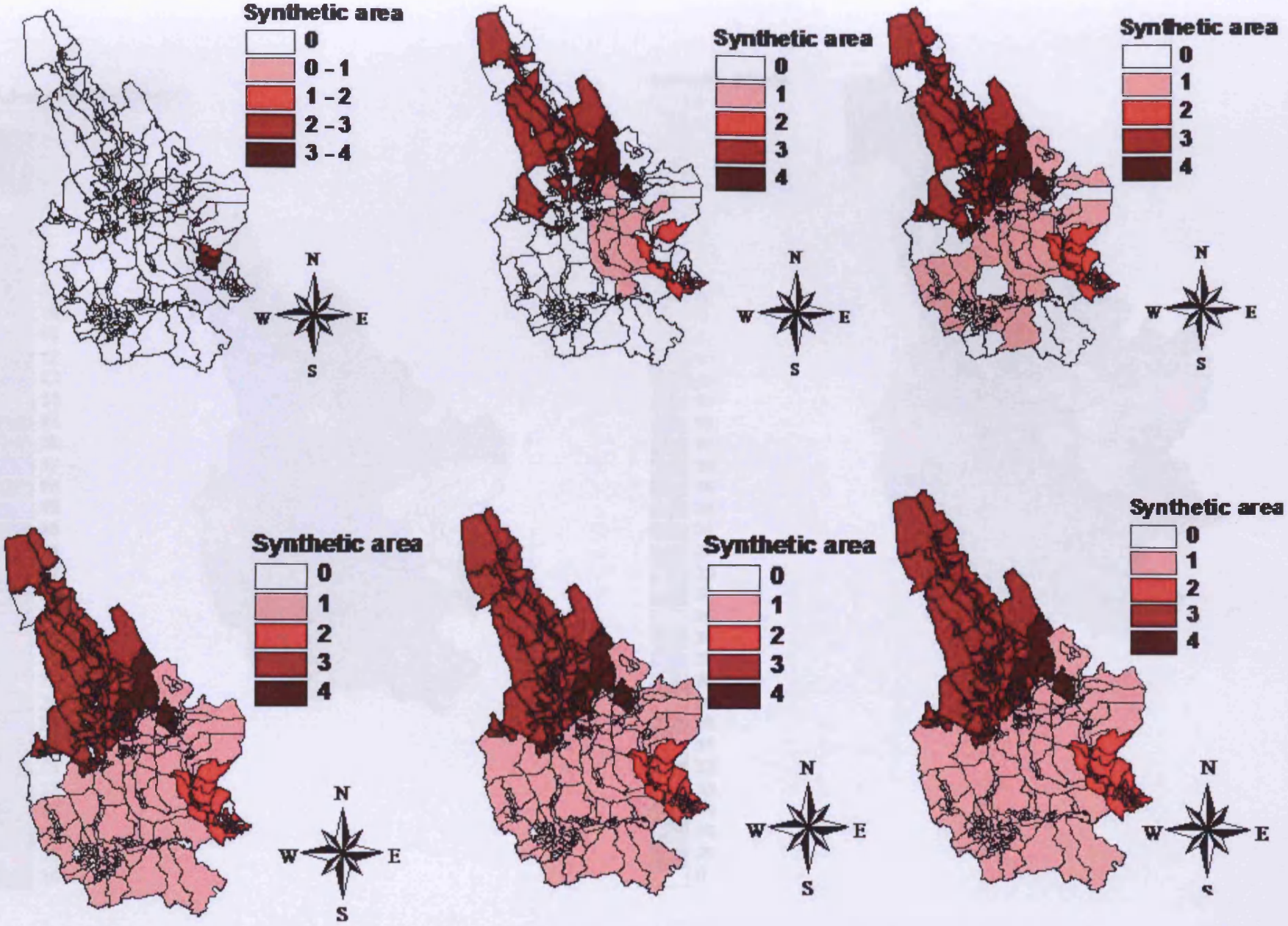
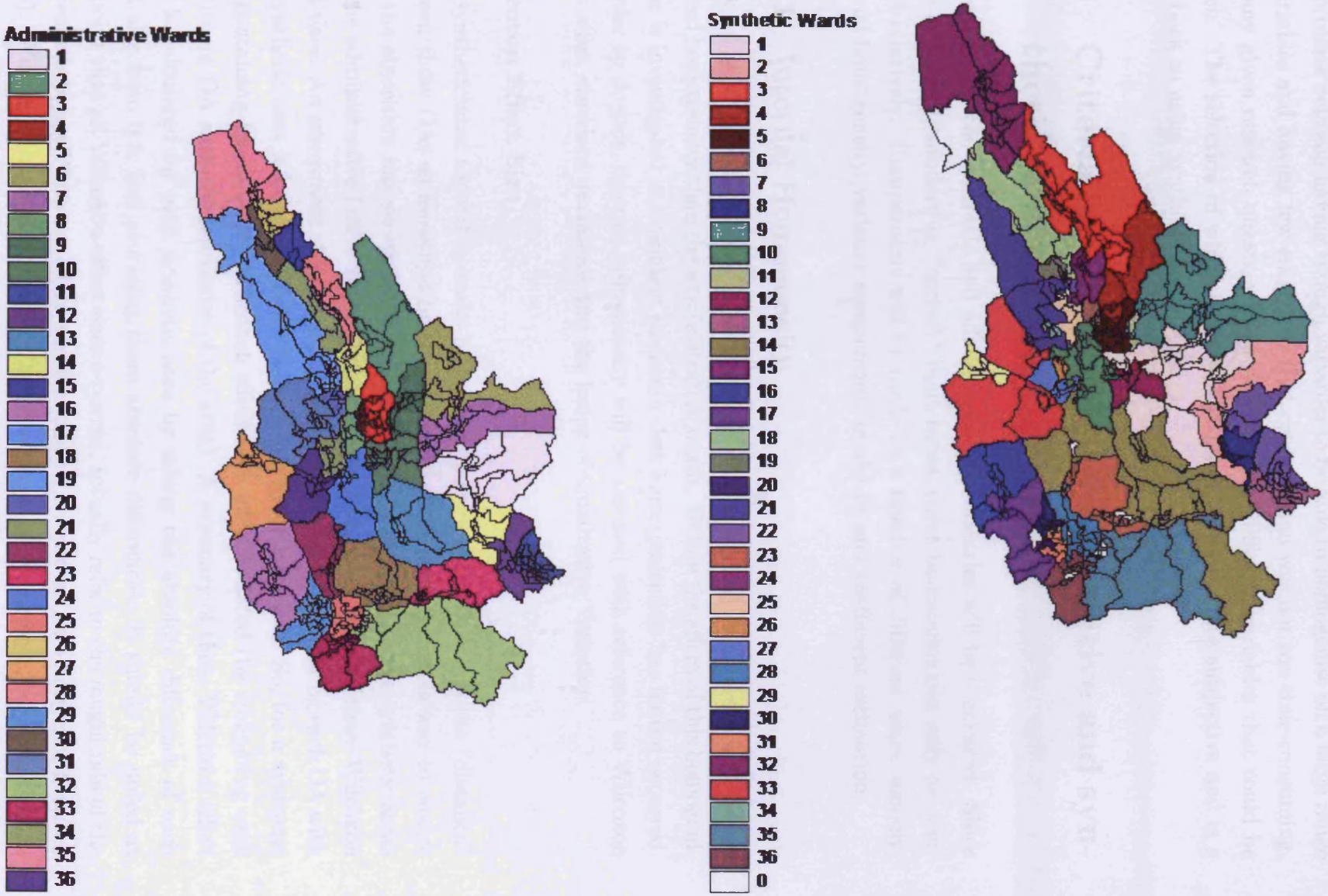


Figure 7.4: Comparison of administrative wards and a set of synthetic boundaries



Thirdly, how to choose the homogenising variables is not straightforward. Here variables known to be related with mental health were chosen. Six were chosen as a compromise between having enough variables to be able to homogenise on a large range of variables and having few enough so that computation was not too time-consuming. For any given research question there will be many different variables that could be chosen. The selection of which of these will be included is quite subjective and is a drawback to using synthetic boundaries.

7.5 Criteria for comparing administrative and synthetic boundaries

In this section the synthetic and administrative boundaries will be compared. Since there is no gold standard or "correct" boundaries, these boundaries can only be compared relatively. Comparisons will be made in a number of different ways, namely: internal homogeneity, variance components, model fit and coefficient estimation.

7.5.1 Internal Homogeneity

The goal of the synthetic area algorithm is to produce LSOA sized areas with better internal homogeneity than the administrative wards. Before the effect of this homogenisation is investigated it is prudent to ensure that homogenisation has indeed occurred. In order to do this, internal homogeneity will be assessed with reference to Wilcoxon effect sizes, standard deviations and the Index of Qualitative Variation.

Wilcoxon Effect Size

The synthetic area algorithm works by merging output areas based on the "distance" between these OAs as measured by the Wilcoxon effect size. A simple way to ensure that the algorithm has succeeded in this endeavour is to compare the synthetic areas and the administrative Lower Super Output Areas (LSOAs) in terms of these Wilcoxon effect sizes. An assessment of internal homogeneity is made by comparing each OA with the synthetic area it belongs to with respect to a given covariate. So, for a synthetic area containing ten OAs, ten Wilcoxon effect sizes are calculated (by comparing each constituent OA with the remainder of the area). A summary of these Wilcoxon effect sizes is calculated for each synthetic area by taking the absolute difference of each effect size from 0.5, and averaging these absolute differences. It should be noted at this point that all Wilcoxon effect sizes reported, actually refer to the magnitude of the difference between Wilcoxon effect and 0.5 (since 0.5 represents no difference between areas). Wilcoxon thresholds therefore, refer to range of Wilcoxon effect sizes that lie

within a certain range of values either side of 0.5

Index of Qualitative Variation

However, it is not enough to demonstrate that the resultant synthetic boundaries are more homogenous with respect to Wilcoxon effect sizes. After all, Wilcoxon effect size is the measure used to determine whether two adjacent areas should be assigned to the same synthetic area or not. It is important therefore to utilise some other measure of homogeneity to compare the synthetic and administrative boundaries to demonstrate that the algorithm has had a homogenising influence that is not limited to the Wilcoxon effect size. A measure called the Index of Qualitative Variation (IQV) (Lieberson, 1969) was used for this purpose. As the name suggests the IQV provides a measure of the variability of qualitative or categorical variables. The IQV is shown in equation 7.2, where K is the number of categories in the variable, and P_i is the proportion of the dataset belonging to category i.

$$\text{IQV} = 1 - \sum_{i=1}^{K} P_i^2 \tag{7.2}$$

The summation term gives the probability that a randomly chosen pair of observations belong to the same category. Subtracting from one gives the probability of a non-matching pair. The probability that a randomly chosen pair of observations will differ on any characteristic is highly dependent on the number of categories that characteristic represents. If comparisons were to be made between different characteristics it would be important to correct for this dependency, however in this situation the synthetic and administrative areas will be compared on the same characteristics and so no correction is necessary. In fact, it is distinctly undesirable to correct for the number of categories in a characteristic since this would mean that the index no longer represents the probability of a non-matching pair. Since the IQV is a probability it ranges between 0 and 1, with 0 indicating no variation (i.e. every individual belongs to the same category) and 1 indicating maximum variation (i.e. no two individuals belong to the same category).

Equation 7.2 defines the IQV for a single variable with K categories. It can be extended to assess the homogeneity of a sample based on two or more variables according to equation 7.3, denoted by GIQV (generalised index of qualitative variation). Here N is the number of possible combinations (i.e. for j variables with $n_1, n_2, ..., n_j$ categories, $N = \prod_{i=1}^{j} n_i$), C_i refers to the proportion of the sample belonging to combination i, C_{ij} denotes $C_i \times C_j$ and W_{ij} refers to the proportion of shared characteristics between combinations i and j. The GIQV can be interpreted as the proportion of characteristics two randomly selected individuals can be expected to differ upon. The GIQV

is a useful method to assess the homogeneity of the synthetic areas, for three reasons: 1. it is not the method by which the synthetic areas are created (avoiding a circular argument), 2. it is designed for use with categorical variables and 3. it provides a measure of the homogeneity of all six merging merging criteria simultaneously. Since all comparisons using this measure are made on multiple variables the GIQV will be referred to simply as the IQV.

$$\text{GIQV} = 1 - \sum_{i=1}^{N} C_i^2 + 2 \sum_{i=1}^{N} \sum_{j>i}^{N} C_{ij} W_{ij} \tag{7.3}$$

The first summation calculates the probability that two randomly chosen individuals match on all of the characteristics in question. The second summation allows for the fact that individuals who share some characteristics, but not all, also contribute to homogeneity. The probability of this partial matching is weighted by the proportion of shared characteristics.

7.5.2 Variance Components

As introduced in chapter 5, the relative importance of each level is calculated by expressing the variance component at that level as a fraction of the sum of all of the variance components. This fraction is called the Intra Class Correlation (ICC) coefficient. The ICC can be used to assess the effectiveness of the homogenisation, since if homogenisation does indeed produce areas that are more meaningful, or at least more amenable to hierarchical modelling, then this should be reflected in the variance components. Essentially, the higher level variance components (in this scenario OAs and synthetic areas) should be larger for the synthetic hierarchy than for the equivalent administrative hierarchy in the null model. This is because the synthetic hierarchy should convey more information than the administrative hierarchy when no other variables are included. This in turn implies that the ICC coefficients for the synthetic boundaries will be larger than for the administrative boundaries. This is a direct consequence of homogenisation and implies that knowing an individual's position in the synthetic hierarchy conveys more information than knowing their position in the administrative hierarchy. Variance components for four different models were fitted. Firstly, a three-level null model was fitted, as described in equation 7.4, with individuals nested within OAs, nested within synthetic areas. This is called Model 1.

$$\text{Mental Health}_{ijk} = \beta_0 + \nu_k + \mu_j + e_{ijk} \tag{7.4}$$

The next model, Model 2, uses the same hierarchical structure, but included individual-level socioeconomic variables (age (modelled as a cubic), gender, social class, employment status, gross income, tenure, and council tax band) and an area-level explanatory

variable. This was the percentage of people claiming disability benefits at area level. Model 2 is given in equation 7.5. Here, the βs either reference scalar quantities (for the continuous variables like age and percent disability) or vectors (for the categorical variables like gender, social class, employment status, gross income, tenure and council tax band). Once individual-level variables are included, the ICC attributable to the synthetic area-level should be reduced. This reduction should be greater than the equivalent reduction using administrative boundaries, if the synthetic boundaries are more homogenous.

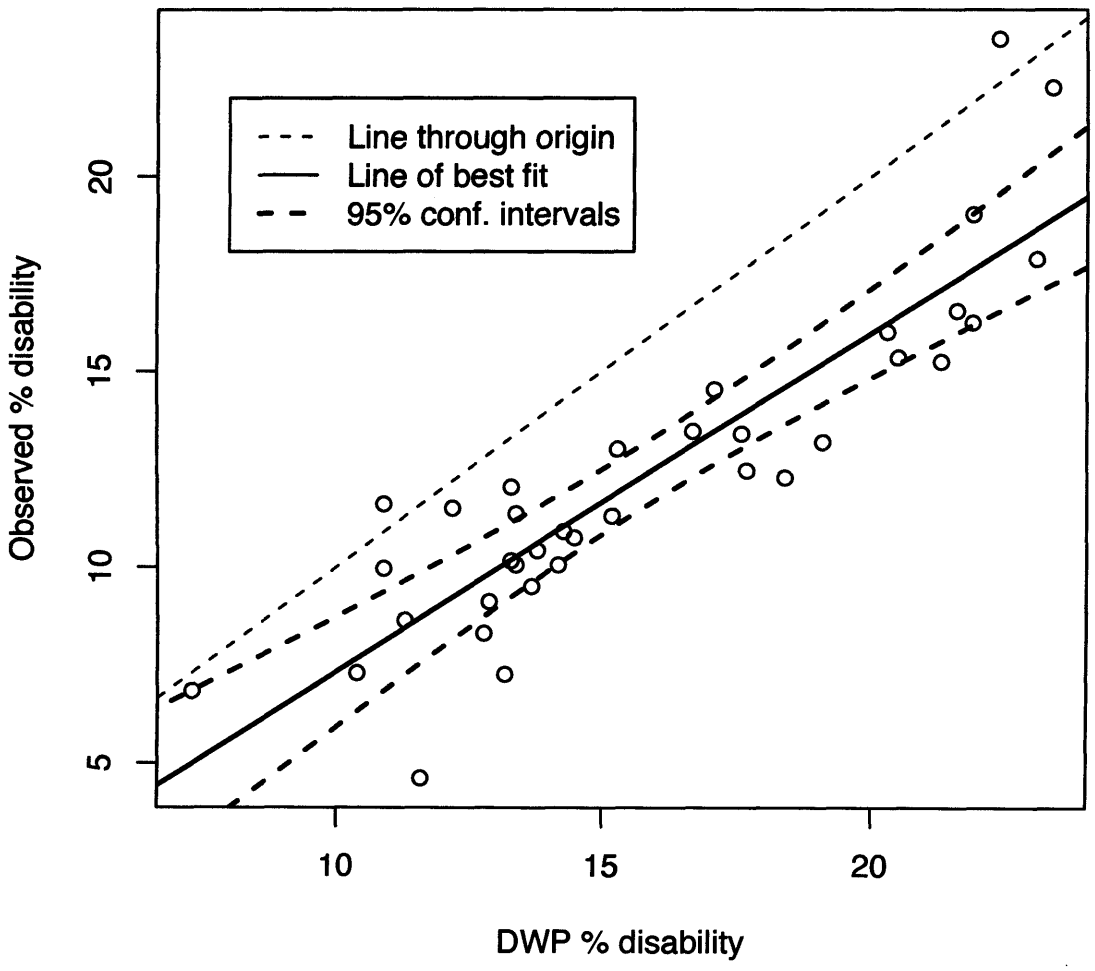
$$\begin{aligned} y_{ijk} = \beta_0 + \beta_1 \text{Age}_i + \beta_2 \text{Age}_i^2 + \beta_3 \text{Age}_i^3 + \beta_4 \text{Gender}_i + \beta_5 \text{Social Class}_i \\ + \beta_6 \text{Employment Status}_i + \beta_7 \text{Gross Income}_i + \beta_8 \text{Tenure}_i \\ + \beta_9 \text{Council Tax Band}_i + \beta_{10} \text{Percent Disability}_k + \nu_k + \mu_j + e_{ijk} \end{aligned} \tag{7.5}$$

In order to investigate the impact of including area-level fixed effects with different synthetic areas, an area-level variable that can be calculated for any area needs to be chosen. It was decided to choose the variable with the largest effect size from the CHSNS (Fone, 2005), namely being permanently unable to work due to disability. This can be aggregated for any area to give a percent disability in the area, representing the proportion of the sample from a given area that are claiming incapacity benefits. Since it is an area-level variable it has the subscript k. This model uses the synthetic hierarchy specified to calculate the percentage of individuals residing in each synthetic area who are disabled. This model can be used to investigate the impact of changing the hierarchy on a higher level area variable.

Since this percent disability variable is an aggregation of individual-level responses it is important to ascertain how accurately it represents an area. Using information from the Department of Work and Pensions (DWP) on the percentage of disability claimants in each ward, we can assess how good a proxy for area-level disability, our aggregated disability variable is. Figure 7.5 compares the two sources of information. The line through the origin indicates where the data should lie if the two variables were identical. Instead it seems that the observed percent disability consistently underestimates the percentage of disability claimants in wards. The relationship is approximately linear however ($R^2 = 0.79$) and the correlation between the two variables is high at 0.89, which provide some justification for the use of an aggregated disability variable as a proxy for area-level disability.

The next two models replicate models 1 and 2, but instead of modelling just one hierarchy, these models include both the administrative hierarchy and the synthetic hierarchy in a cross-classified approach. This allows a further comparison to be made between the hierarchies. With both sets of hierarchies being fitted simultaneously, it allows the relative importance of each to be measured and compared. To avoid

Figure 7.5: Relationship between the DWP percent disability figures and the observed percent disability figures



unnecessary repetition, only variance components will be investigated for these two models.

7.5.3 Model Fit

Model fit will be assessed using the Akaike Information Criterion (AIC) as introduced in section 5.19. Lower AIC values indicate better model fit.

7.5.4 Coefficient Estimation

The effect of changing the hierarchy on the coefficient estimates, as well as their associated standard errors, is unclear. In this section coefficient estimation is only examined for a synthetic area-level covariate and so the percent disability variable from Model 2 is used. As described earlier, this model controls for the same basic socioeconomic variables (age (modelled as a cubic), gender, social class, employment status, tenure and council tax band), but also includes the proportion of people unemployed due to disability. Coefficients for this variable along with its associated standard error are calculated for the administrative LSOAs and for the synthetic areas. The initial merging thresholds investigated range between 0.060 and 0.500 in steps of 0.005. Initial thresholds higher than 0.300 can be thought of as producing randomly delineated synthetic areas, since most pairwise comparisons of OAs produce Wilcoxon effect sizes smaller than this. Hence, the main condition necessary for two OAs to be merged when the initial merging threshold is higher than 0.3, is that they be adjacent.

7.6 Technical details of the comparison process

As described earlier, the algorithm requires four sets of information in order to produce a set of synthetic boundaries: adjacency information for the areas, composition information for the areas (in this case the six homogenising variables), an initial merging threshold indicating how similar areas need to be for them to be merged and a set of seed pairs representing initial values for the algorithm. The adjacency information is a constant for the CHSNS dataset, as is the composition information (given the six homogenising variables). The initial merging threshold was varied between 0.06 and 0.5. A number of different choices of seed pairs were investigated. For this thesis however, results are presented for when 130 initial seed pairs are used.

In order to compare the results produced by the synthetic area algorithm with administrative boundaries, it was necessary to create a large number of synthetic boundaries. As described earlier, the initial merging threshold is varied between 0.060 and 0.500 in steps of 0.005. At each merging threshold approximately 300 synthetic hierarchies were created, resulting in a total of over 25,000 simulated hierarchies. To these

hierarchies four multilevel models were fitted and parameters of interest extracted. To create a single set of synthetic boundaries and fit the necessary models to it, took approximately eight minutes. The total CPU time required for this work was approximately 20 weeks. Again, the Condor parallel computing facility (Litzkow et al., 1988) was utilised in order to expedite this process.

7.7 Results using 130 initial seed pairs

Firstly the initial seed pairs which act as initial values for the algorithm are mapped in figure 7.6.

7.7.1 Internal Homogeneity

Wilcoxon Effect Size

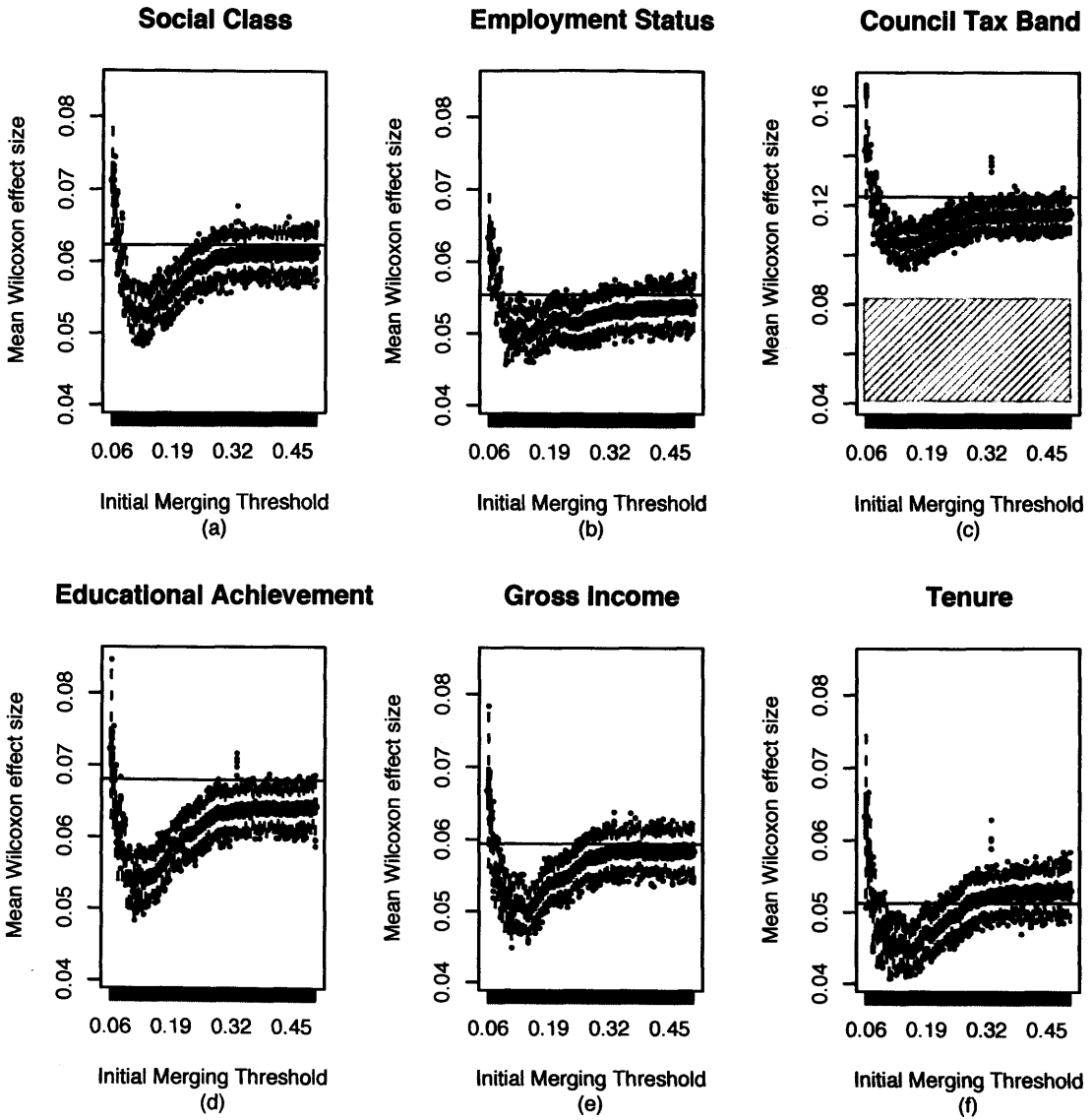
Average Wilcoxon effect sizes for each of the six homogenising variables (for each OA compared with the synthetic area to which it is assigned) are plotted against the initial merging threshold in figure 7.7.

The y-axes are the same for five of the six homogenising variables. The council tax band y-axis however, is plotted on a different scale, since the smallest average Wilcoxon effect size reported for the council tax band variable is nearly twice the size of the largest average effect sizes reported for the other variables. The grey shaded area on the council tax band plot indicates the y-axis range of the other five homogenising variables. Such large mean Wilcoxon effect sizes indicate that the distribution of council tax bands is not particularly homogenous. As a result, even when the merging threshold is at its strictest there is still considerable variability in this variable.

All six variables have a similar shaped relationship with the initial merging threshold. Since lower average Wilcoxon effect sizes imply greater internal homogeneity, it is clear that the optimum merging threshold is close to 0.11 (the optimum is attained at an initial merging threshold of 0.11 for four of the six variables). Decreasing the initial merging threshold below this value results in a sharp increase in average Wilcoxon effect size. Above this value there is a less steep increase in average Wilcoxon effect sizes, which levels off as the initial merging threshold reaches 0.5 (the maximum possible). This V-shaped relationship is a consequence of smaller thresholds resulting in fewer synthetic areas being produced. A low merging threshold means that adjacent areas must be quite similar before they will be merged. If the merging threshold is too low however, few adjacent areas will be merged and in extreme cases entire seeds may be removed by the algorithm, reducing the number of synthetic areas produced. Fewer synthetic areas mean larger synthetic areas, which in turn mean more heterogenous synthetic areas. There is a tradeoff to be made therefore between a merging threshold

Figure 7.6: 130 initial seed pairs

Figure 7.7: Relationship between the initial merging threshold and the resulting mean Wilcoxon effect size for each of the six homogenising variables when there are 130 seed pairs



strict enough to ensure good internal homogeneity, but relaxed enough to ensure that sufficient numbers of synthetic areas are output.

The horizontal lines on each plot indicate the average Wilcoxon effect size for the 2001 LSOA census boundaries. For most of the variables, the synthetic areas produce lower average Wilcoxon effect sizes than for the administrative boundaries. This indicates that practically all of the synthetic boundaries produce greater homogeneity for most of the variables than for the administrative boundaries. The tenure variable is the only exception, with initial merging thresholds greater than 0.275 producing more heterogenous areas than for the administrative boundaries. This in turn implies that the 2001 administrative boundaries do not group similar people together any better than randomly chosen boundaries.

In order to properly interpret the v-shaped relationship between the initial merging threshold and the mean Wilcoxon effect size, the relationship between the initial merging threshold and the resulting number of synthetic areas needs to be considered. Figure 7.8 displays this relationship and indeed shows that smaller thresholds consistently produce smaller numbers of synthetic areas. As the threshold is increased so does the number of synthetic areas produced.

To provide a frame of reference for the initial merging criterion, all Wilcoxon effect sizes resulting from all pairwise comparisons of adjacent OAs in the borough for all six homogenising variables are plotted in figure 7.9. The vertical line indicates 0.3 (this will be relevant for interpreting the graphs in section 7.7.2).

Index of Qualitative Variation

The IQV is calculated for each synthetic area for a given set of synthetic boundaries. These IQVs are then averaged and plotted against both the initial merging threshold and the number of synthetic areas, as shown in figures 7.10 and 7.11. IQVs are interpreted as the proportion of characteristics two randomly selected individuals from the same area will differ on. We can see that the strictest merging threshold gives higher proportions (indicating higher heterogeneity) than the more relaxed merging thresholds. This seems counterintuitive until we recognise that strict merging thresholds are associated with fewer synthetic areas. The more synthetic areas there are the more chance there is of those areas being small enough to be relatively homogenous. It is important to note however that the administrative area IQV, indicated by the horizontal line, is 0.980, lower than most of the IQVs produced by the synthetic area algorithm. The range of IQVs produced is from 0.979 to 0.995.

Figure 7.8: Relationship between the number of synthetic areas and the initial merging threshold

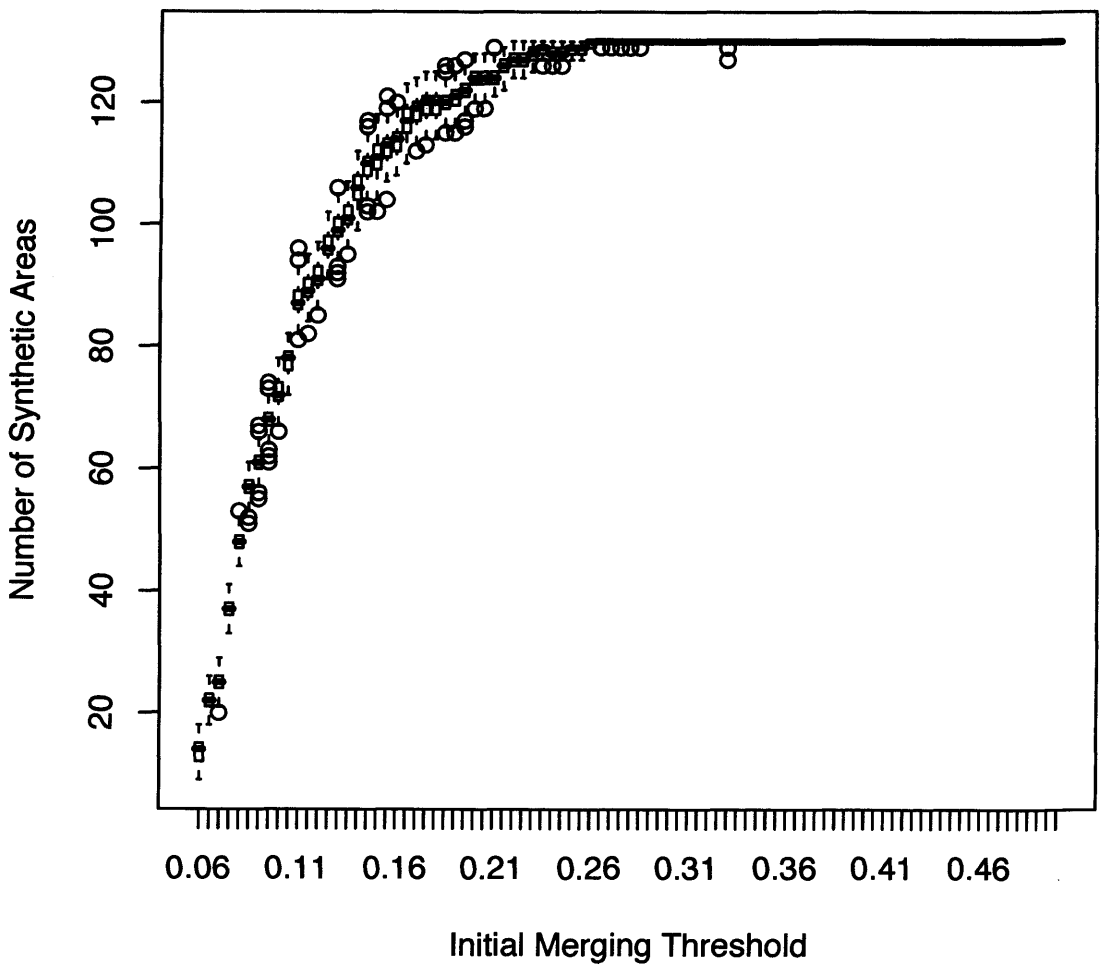
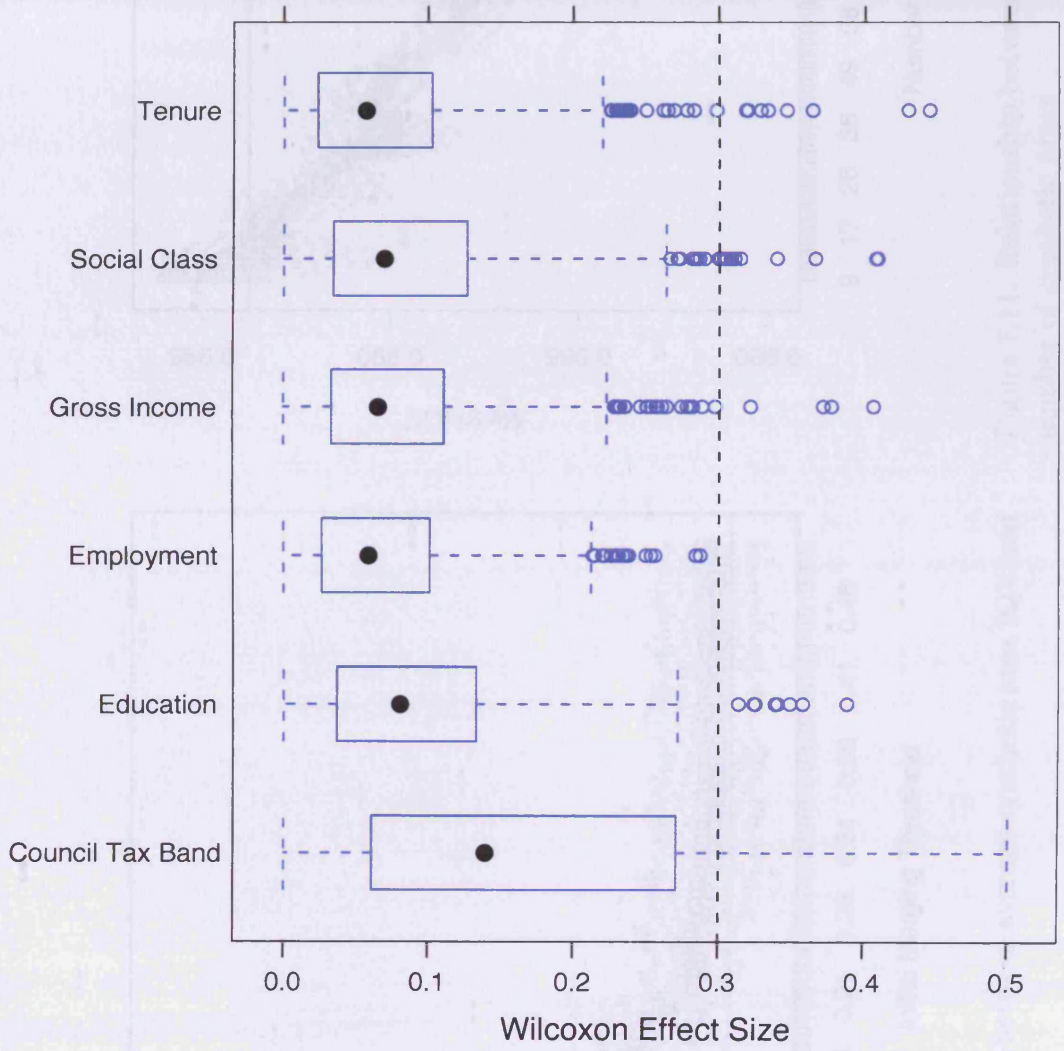


Figure 7.9: Wilcoxon effect sizes for all pairwise comparisons of adjacent OAs, for all six homogenising variables



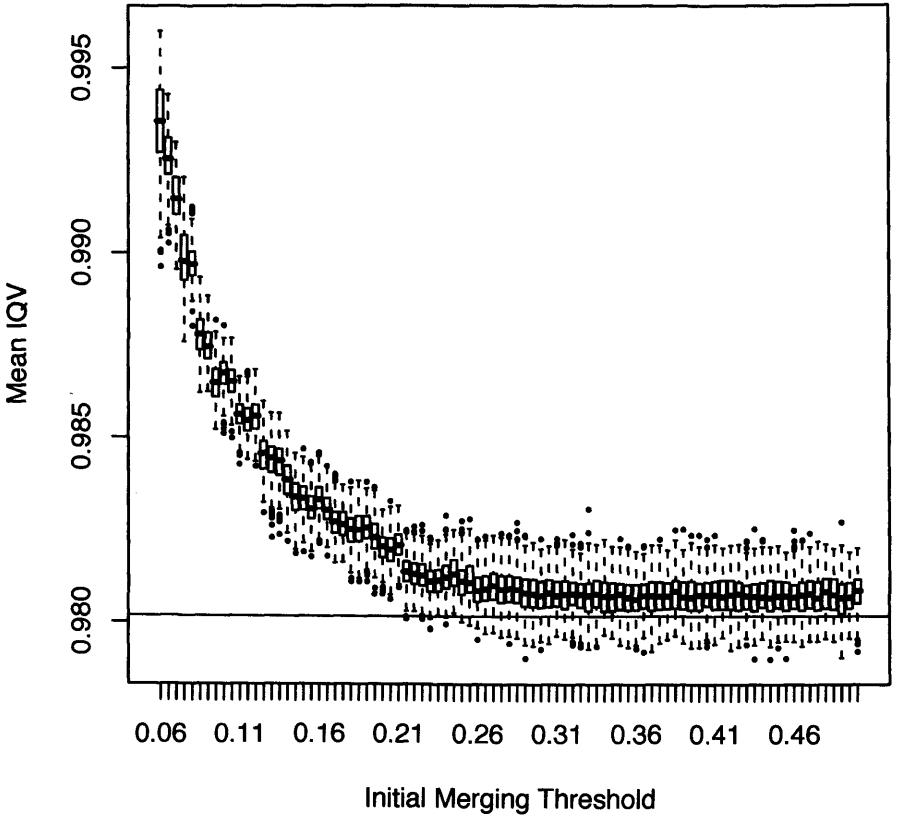


Figure 7.10: Relationship between average synthetic area IQV and initial merging threshold

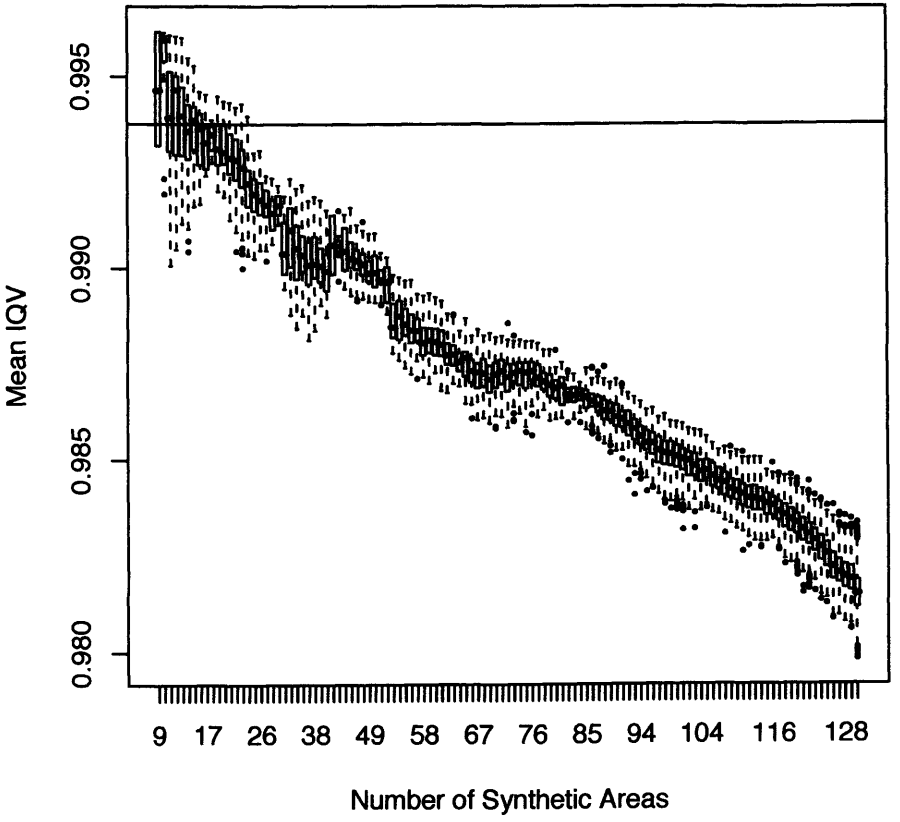


Figure 7.11: Relationship between average synthetic area IQV and number of synthetic areas

7.7.2 Variance components

This section will summarise the four models outlined in section 7.5.2, in terms of their variance components. Variance components will be illustrated via ICC coefficients. As mentioned previously, models 3-4 (the cross-classified models) will only appear in this section.

Variance Components

The relationship between the synthetic area-level ICC coefficient for the null model (as given in equation 7.4) and both the initial merging threshold and the resulting number of synthetic areas is displayed in figure 7.12. The horizontal line indicates the 2001 LSOA census boundaries null model ICC coefficient (0.026). This can be interpreted as being the proportion of the variability attributable to the LSOA of residence of an individual when only area of residence and household and individual identifiers are available. The synthetic area algorithm produces larger ICCs than the administrative boundaries for only one initial merging threshold; 0.11. Even then, the increase in ICC is not large. This indicates that the 2001 LSOAs produce ICC coefficients as large as any of the boundaries produced by the synthetic area algorithm. It is interesting to note however that figure 7.12.(b) shows that the synthetic area ICCs larger than 0.026 were produced by hierarchies containing between 83 and 92 synthetic areas. This is fewer areas than the administrative boundaries which have 110 LSOAs. This indicates that the same amount of highest-level variation present in the administrative areas was captured in fewer areas using the synthetic area algorithm.

The relationship between the synthetic area-level ICC coefficient for the percent disability model (equation 7.5) and both the initial merging threshold and the resulting number of synthetic areas is displayed in figure 7.13. The synthetic level ICC coefficient does not appear to be strongly related to either the merging threshold or the number of synthetic areas produced. The administrative boundaries produce an ICC coefficient of 0.0022. Almost 90% of the synthetic boundaries produce lower ICCs than this. Note the difference in y-axis scales between figures 7.12 and 7.13, indicating that the individual-level information explains more of the synthetic area-level variation than the administrative area variation.

Next, cross-classified models are examined. Again, two models are fitted: a null model, and a model with individual-level covariates and an area-level covariate. These models include both the synthetic and administrative hierarchies in the same model and so can be used to compare how important each context is. Figure 7.14 illustrates the variance contribution from each hierarchy as the initial merging threshold is varied in the null model. The synthetic area-level ICC is close to zero when the merging threshold exceeds 0.3 (on average 0.003). Synthetic levels produced at such high merging

Figure 7.12: Relationship between the synthetic area-level ICC coefficient (for the null model) and both the initial merging threshold and the resulting number of synthetic areas

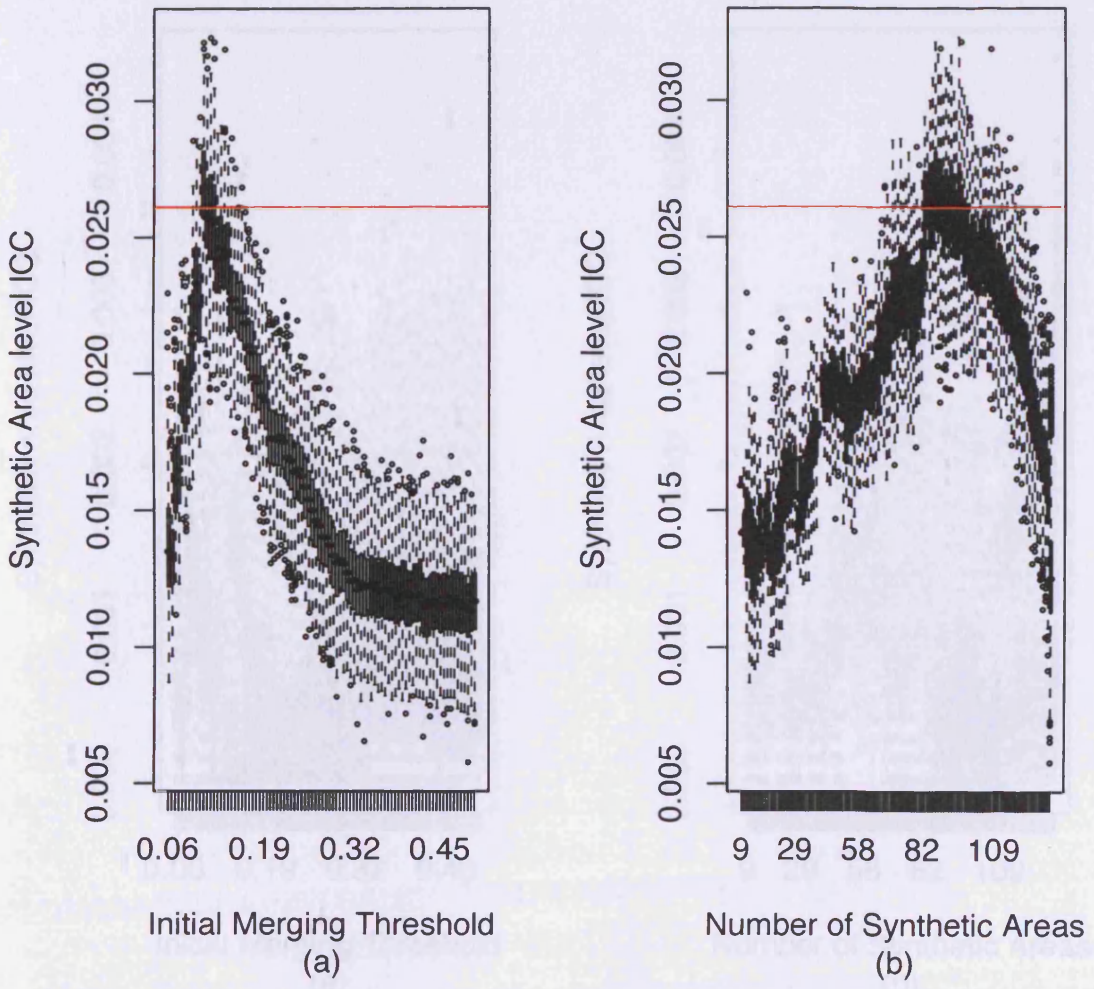


Figure 7.13: Relationship between the synthetic area-level ICC coefficient (for the percent disability model) and both the initial merging threshold and the resulting number of synthetic areas

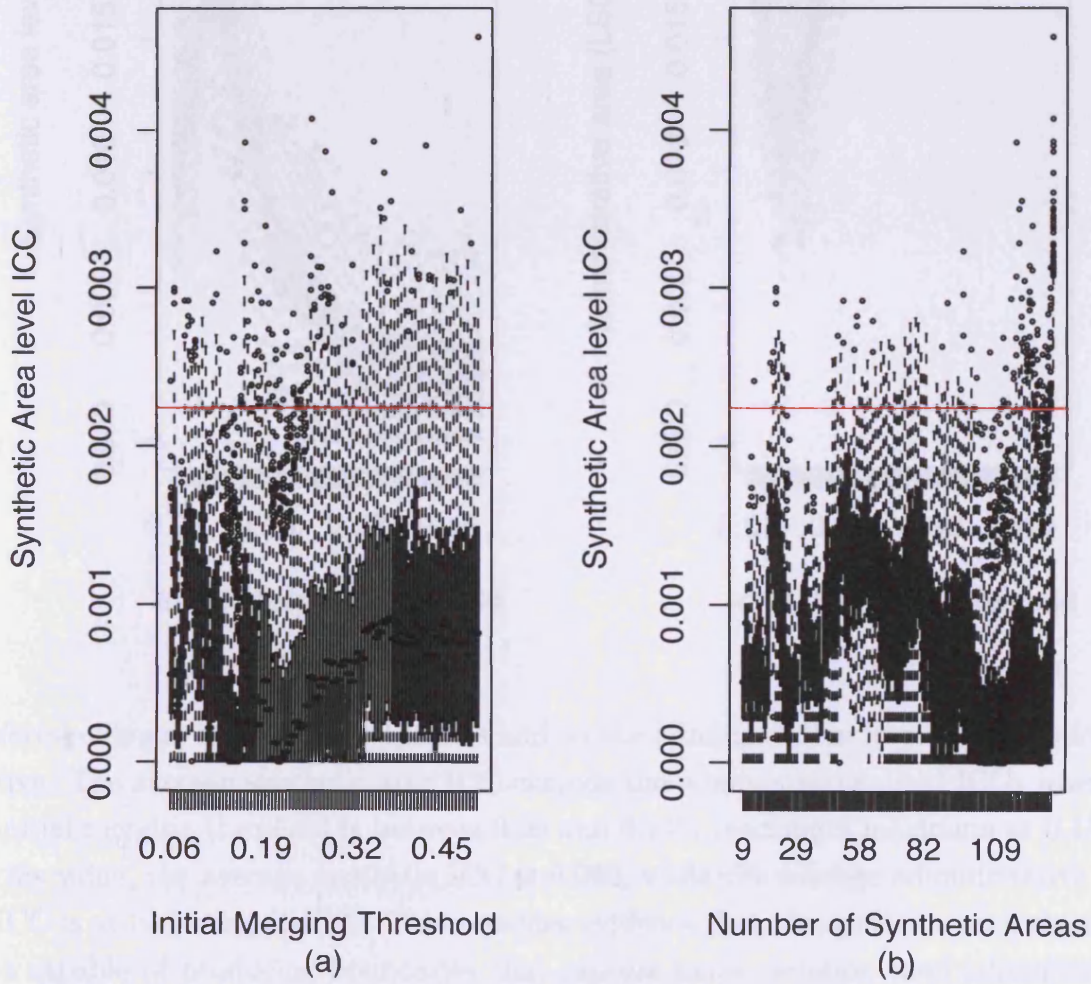
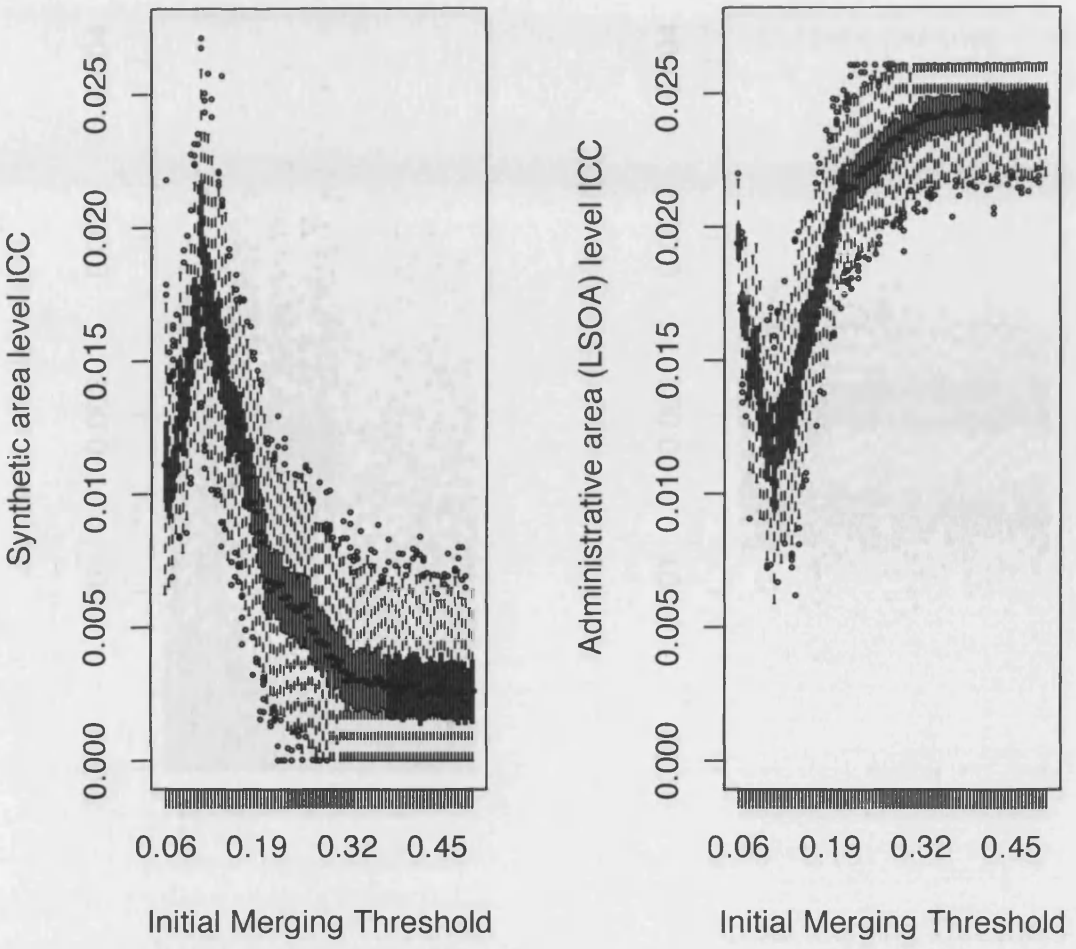


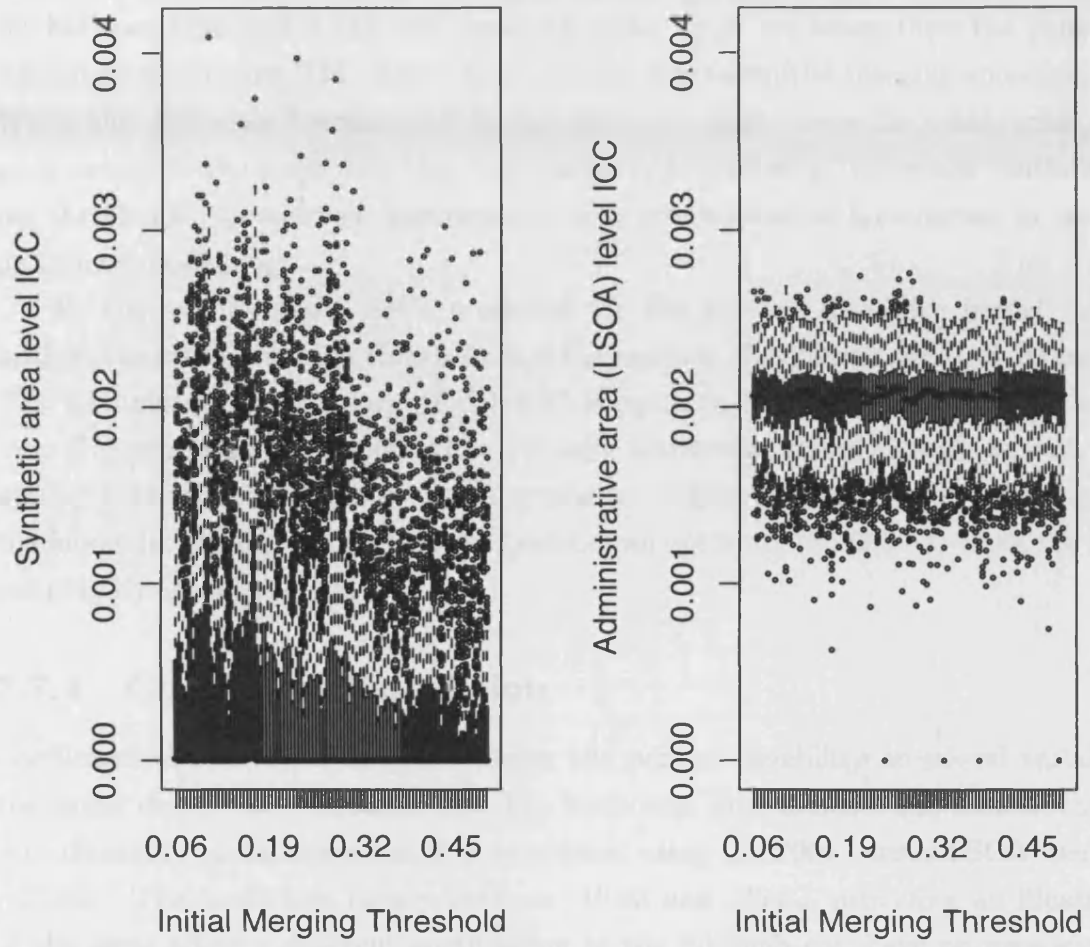
Figure 7.14: Comparison of synthetic and administrative hierarchies ICCs in a cross-classified null model



thresholds are not very homogenous and so the administrative level is more informative. The average synthetic area ICC exceeds the administrative level ICCs when the initial merging threshold is between 0.09 and 0.145, reaching a maximum at 0.11. At this value, the average synthetic ICC is 0.020, while the average administrative level ICC is at its lowest (0.010). This provides evidence that the synthetic area algorithm is capable of producing boundaries that capture more variation than administrative boundaries.

Figure 7.15 shows the variance components for the cross classified percent disability model. Once individual-level covariates and the percent disability covariate are included both the synthetic level ICC and the administrative level ICC are on average reduced (when compared with the null models). It is clear that the variance contribution of the synthetic area is much more variable than for the administrative areas, since

Figure 7.15: Comparison of synthetic and administrative hierarchies fixed effect estimation in a cross-classified percent disability model



the synthetic areas change while the administrative areas do not. The average administrative area ICC is 0.02, which is greater than the vast majority of synthetic area ICCs (a little over 1.5% of the synthetic boundaries produce ICCs higher than 0.02). There is indication that stricter merging thresholds produce more variable synthetic ICCs. It is important to note however, that in the previous graph (figure 7.14) average synthetic level ICCs were greater than the administrative level ICCs when the initial merging threshold lay between 0.09-0.145. Here however, the average administrative ICC is greater than the average synthetic ICC for all merging thresholds. This implies a greater amount of the variation attributed to the synthetic level ICC is explained by the addition of compositional explanatory covariates, than for the administrative level ICC.

7.7.3 Model Fit

The relationship between the AICs for the null model (equation 7.4) and both the initial merging threshold and the number of synthetic areas is shown in figure 7.16. The horizontal line denotes the AIC for the model with administrative boundaries (95637). Lower AIC values indicate better model fit. When the initial merging threshold is set between 0.08 and 0.145, the resulting mean AICs are lower than the equivalent administrative model AIC. The lowest of these was for initial merging threshold 0.11. While the difference between the model fit is not large (note the y-axis scale), this adds weight to the possibility that boundaries produced with "optimum" initial merging thresholds represent an improvement over administrative boundaries, in terms of multilevel modelling.

We also compare the AICs produced for the percent disability model, plotted against the initial merging threshold and the number of synthetic areas, in figure 7.17. The administrative boundary model AIC is equal to 85,096. Few of the AICs (just over 1%) produced by models with synthetic hierarchies reach as high as that. This implies that the vast majority of the synthetic boundaries have consistently improved the model fit. Again, the average AIC reaches an optimum (85,063.53) when the initial merging threshold is equal to 0.12.

7.7.4 Coefficient Estimation

Coefficient estimation is examined using the percent disability area-level variable in the model described in equation 7.5. The horizontal lines indicate the value of the percent disability coefficient when it is calculated using the 2001 census LSOA hierarchy (-23.94). The coefficient ranges between -10.93 and -35.22, providing an illustration of the large effect a different partitioning of the borough can have on area-level covariates. In this instance, all of these coefficients are large enough to attain statistical significance (the largest standard error overall is 2.41), however it is possible that in another situation this range of values could attain statistical significance at one extreme and not at the other. There also appears to be an upward trend for both 7.18.a and 7.18.b. Since synthetic boundaries created with high initial merging thresholds can be thought of as randomly created contiguous areas, this implies that the coefficient of the percent disability variable is attenuated when the boundaries do not group similar people together. The average value of the percent disability coefficient when the merging threshold is 0.11 is -25.66.

Figure 7.16: Relationship between AIC for the null model and both initial merging threshold and number of synthetic areas

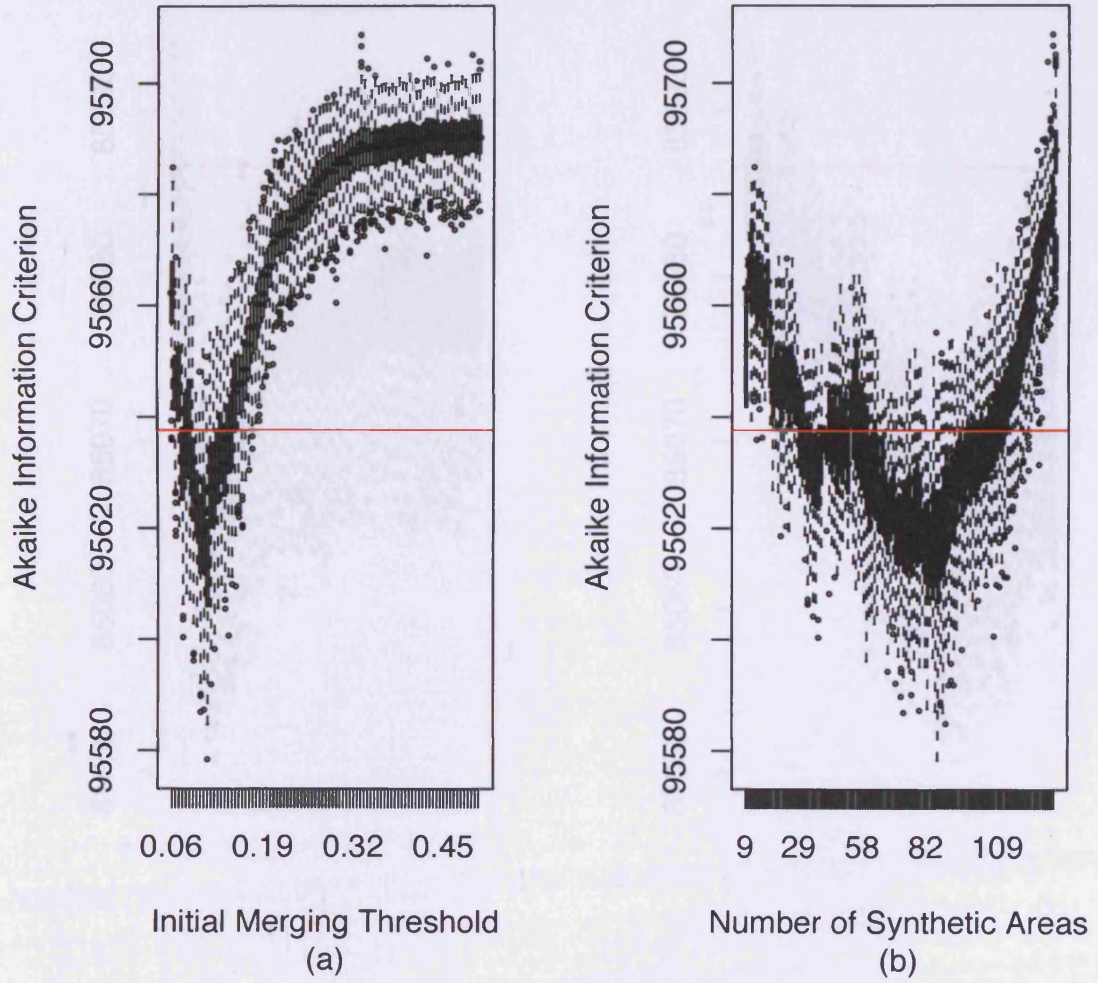


Figure 7.17: Relationship between AIC for the percent disability model and both initial merging threshold and number of synthetic areas

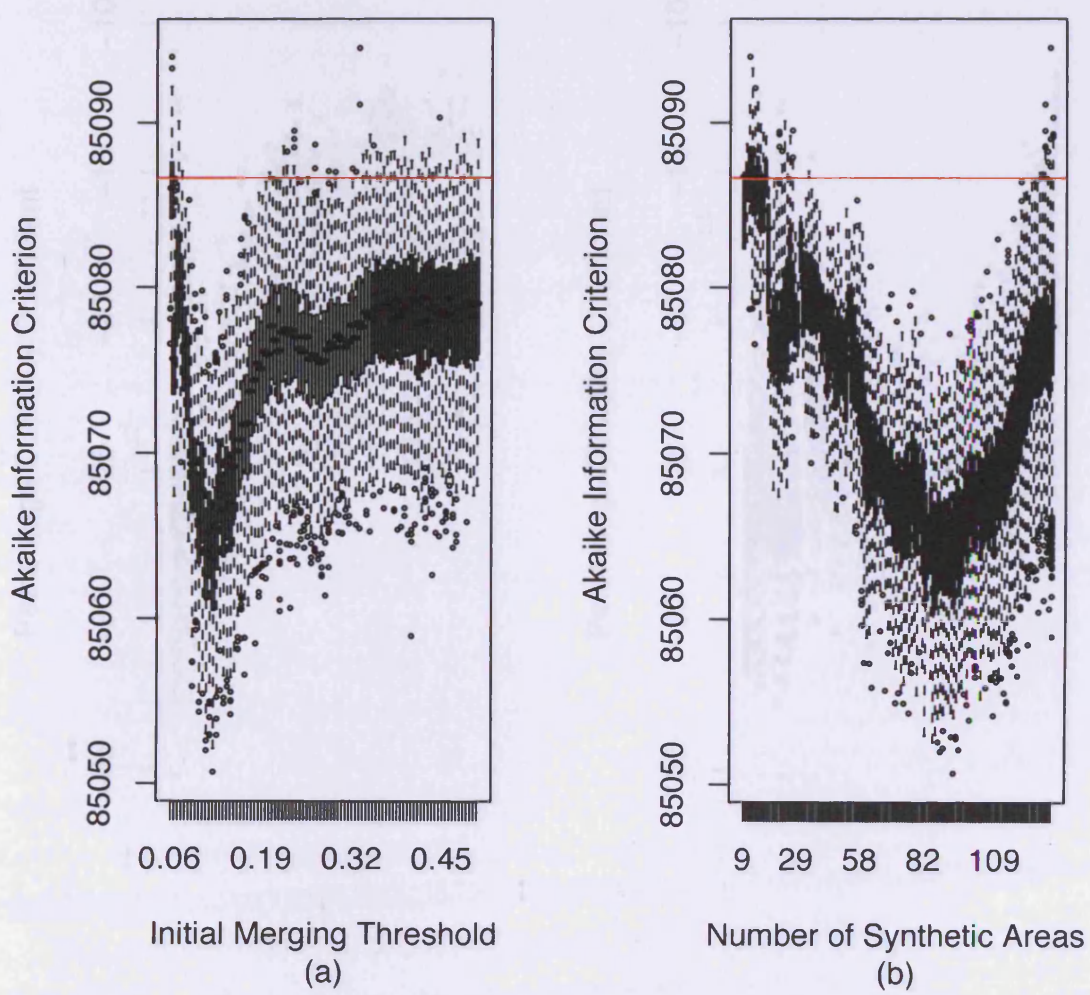
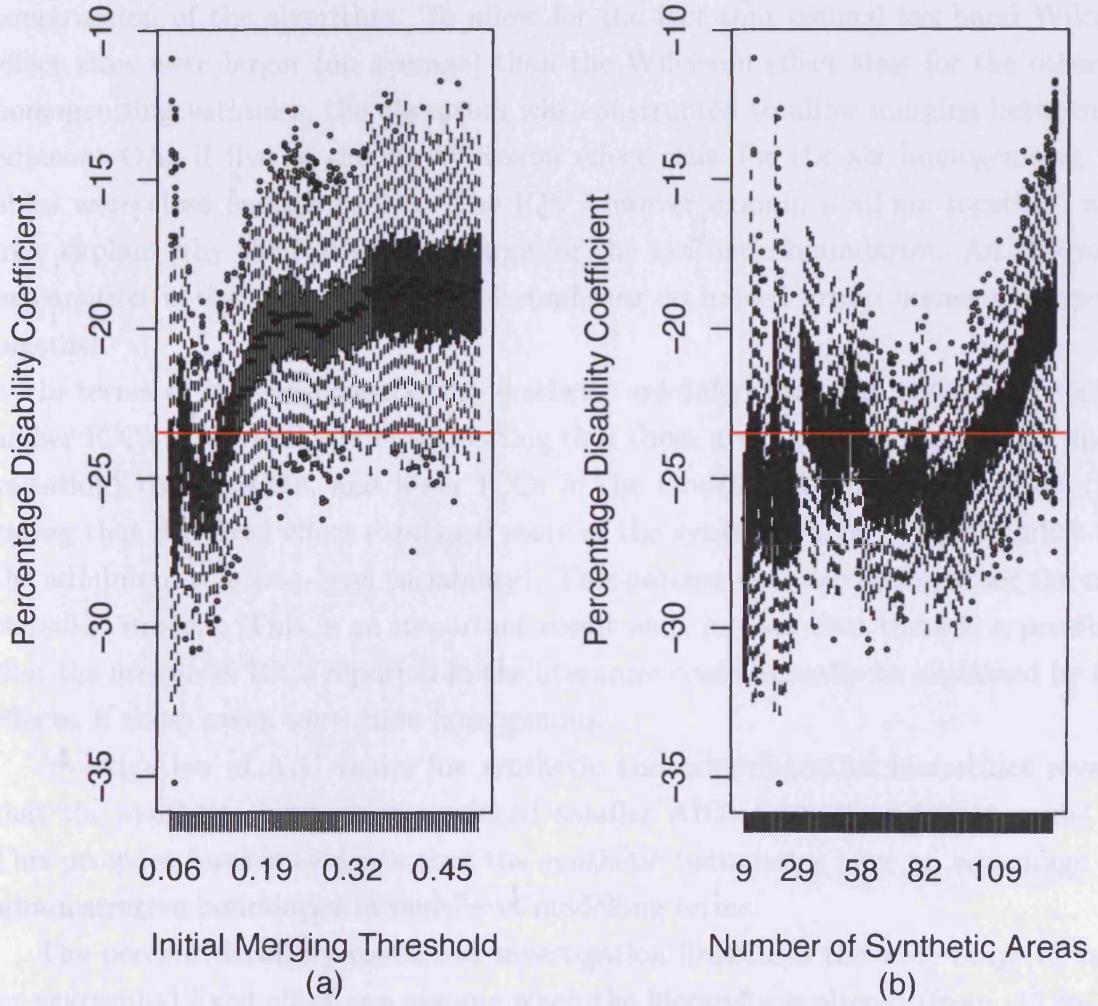


Figure 7.18: Relationship between percent disability coefficient and both initial merging threshold and number of synthetic areas



7.8 Discussion

The previous results demonstrate a number of points. Firstly, the synthetic area algorithm has been demonstrated to be a robust process capable of producing meaningful areas from smaller areas. The synthetic area algorithm was applied 25,000 times and produced a partition of Caerphilly county borough each time.

The algorithm was also shown to produce areas with lower internal average Wilcoxon effect sizes for all of the homogenising variables. This is a good indication that the synthetic areas produced are more homogenous than administrative boundaries. In terms of the index of qualitative variation however, the administrative areas were shown to be more homogenous. This is a surprising result, but is possibly explained by the construction of the algorithm. To allow for the fact that council tax band Wilcoxon effect sizes were larger (on average) than the Wilcoxon effect sizes for the other five homogenising variables, the algorithm was constructed to allow merging between two adjacent OAs if five of the six Wilcoxon effect sizes for the six homogenising variables were close enough to 0.5. The IQV however examines all six together, which may explain why the IQVs are so large for the synthetic boundaries. An alternative explanation is that the 2001 census boundaries do indeed group homogenous people together.

In terms of ICC coefficients, the synthetic area algorithm can produce areas with higher ICCs in the null model (indicating that those areas capture more of the spatial variation) than LSOAs, and lower ICCs in the model including a fixed effect (indicating that the fixed effect explained more of the synthetic area-level variability than the administrative area-level variability). This pattern was also observed for the cross-classified models. This is an important result as it implies that there is a possibility that the area-level ICCs reported in the literature could actually be explained by fixed effects, if those areas were more homogenous.

Investigation of AIC values for synthetic and administrative hierarchies revealed that the synthetic hierarchies produced smaller AICs (indicating better model fit). This provides further evidence that the synthetic boundaries have an advantage over administrative boundaries in multilevel modelling terms.

The percent disability coefficient investigation illustrates the wide range of values an aggregated fixed effect can assume when the hierarchy is altered (from -35 to -11). This provides a stark reminder to researchers about the consequences and impact of choosing a given hierarchy for a multilevel analysis.

Finally, it should be emphasised that the algorithm is a general one and can be altered for different situations. Different building blocks can be used to create synthetic areas (so long as adjacency information is available for those areas). Different homogenising variables can be included depending on the focus of the analysis. It

would even be possible to include different (or multiple) measures for comparing areas. This allows an assessment to be made of how suitable administrative boundaries are for a given analysis.

Chapter 8

Synthesis of thesis findings applied to the Caerphilly Health and Social Needs Study

8.1 Introduction

The previous chapters introduced the literature researching area effects on mental health, described the CHSNS dataset and investigated various methodological issues surrounding its analysis. These comprised the following sections.

- Firstly, the measure of mental health status, the MHI-5, was presented and critiqued. Various methods of analysis were investigated and cutpoints defining a case of common mental disorder were produced.
- Secondly, the issue of whether to include a sparse household level in multilevel analyses of this type was examined using a simulation study.
- Thirdly, the modifiable areal unit problem Openshaw (1984) was explored using an algorithm that partitions an area into smaller contiguous areas in chapter 7. This algorithm will be used to produce a partition of Caerphilly county borough that will be used as an alternative to administrative boundaries in this chapter.

This chapter will draw on the results of these sections of the thesis and combine them to investigate the research question: does where you live affect your mental health, as formulated in the CHSNS (Fone, 2005). Chapter 6 found that including a sparse level of context results in less bias of fixed effect standard errors than excluding a sparse level of context. The household level in the CHSNS is just such a level. It was excluded from the original analyses (Fone, 2005; Fone et al., 2007c), but in this chapter the effect of including it will be examined. Similarly, the CHSNS used administrative boundaries to delineate area of residence. Chapter 7 found that synthetic boundaries can

provide more internal homogeneity than administrative boundaries as well as showing the extent of the impact changing the hierarchy can have on the results of a multilevel analysis. This chapter, guided by the findings of this thesis, will provide a more evidence-based approach to analysing the CHSNS dataset than has previously been attempted in order to better elucidate the link between area of residence and mental health.

8.2 Objective

This chapter applies the findings from the previous chapters regarding MHI-5 cutpoints, the household level, and synthetic boundaries to investigate the contextual determinants of mental health and compares the results with the original CHSNS model (Fone, 2005; Fone et al., 2007c) (called the CHSNS model here). The models chosen for comparison purposes (called models 1 to 4) are now described and justified.

CHSNS model

Previous work based on the CHSNS (Fone, 2005; Fone et al., 2007c) modelled the mental health score as a binomial response using a cutpoint. This was done to avoid the problem of the skewness of the mental health score (discussed in chapter 3). This previous work by Fone et al (referred to all the CHSNS model) will now be summarised. The cutpoint used was 60, chosen because it produces a case prevalence closest to that reported by Caerphilly county borough in the Health in Wales 1996 survey (Kingdon et al., 1998) (32.0%).

The cutpoints for the MHI-5 found in chapter 3 were 76 (from the Youden Index and (0,1) methods), 60 (from the misclassification rate method) and 68 (from the minimax criterion and prevalence matching methods). As discussed in section 3.4.3, since the data is from a relatively small area and comparisons are not being made across areas with widely different case prevalences, minimising the misclassification rate is the most sensible criterion to employ. By chance, this provides the same cutpoint as the CHSNS of 60. This coincidence is not particularly remarkable since the MHI-5 has only a small number of possible values (20). All the models that are compared in this chapter use the same cutpoint.

The response variable is binomial indicating whether a given individual's MHI-5 score was less than or equal to 60, or not. The explanatory variables will now be described. Age was modelled as a cubic polynomial. The following variables were included as individual-level indicators: gender, social class (I or II, IIINM, IIIM, IV and V, Other, Missing), employment status (employed, seeking work, home or carer, student/training, permanently sick/disabled, retired, missing), gross household income

(low, medium, high), tenure (owner-occupier, non-owner occupier, missing), council tax band (A-B, C-H, missing). This approach could be criticised since the last three of those variables are actually measured at the household level, yet the household level is excluded from the analysis. Interactions between age and each of the following were investigated: female gender, being permanently sick/disabled and low income. An interaction between age squared and low income was also investigated. Ward-level variables were also incorporated, comprising the incapacity claimant ratio and an interaction between the incapacity claimant ratio and individual disability. The latter is a cross-level interaction, as introduced in section 5.2.

Each of the next four models changes this first model in a different way based on the findings of the thesis. These changes are cumulative, so that the final model incorporates all of the thesis findings.

Model 1

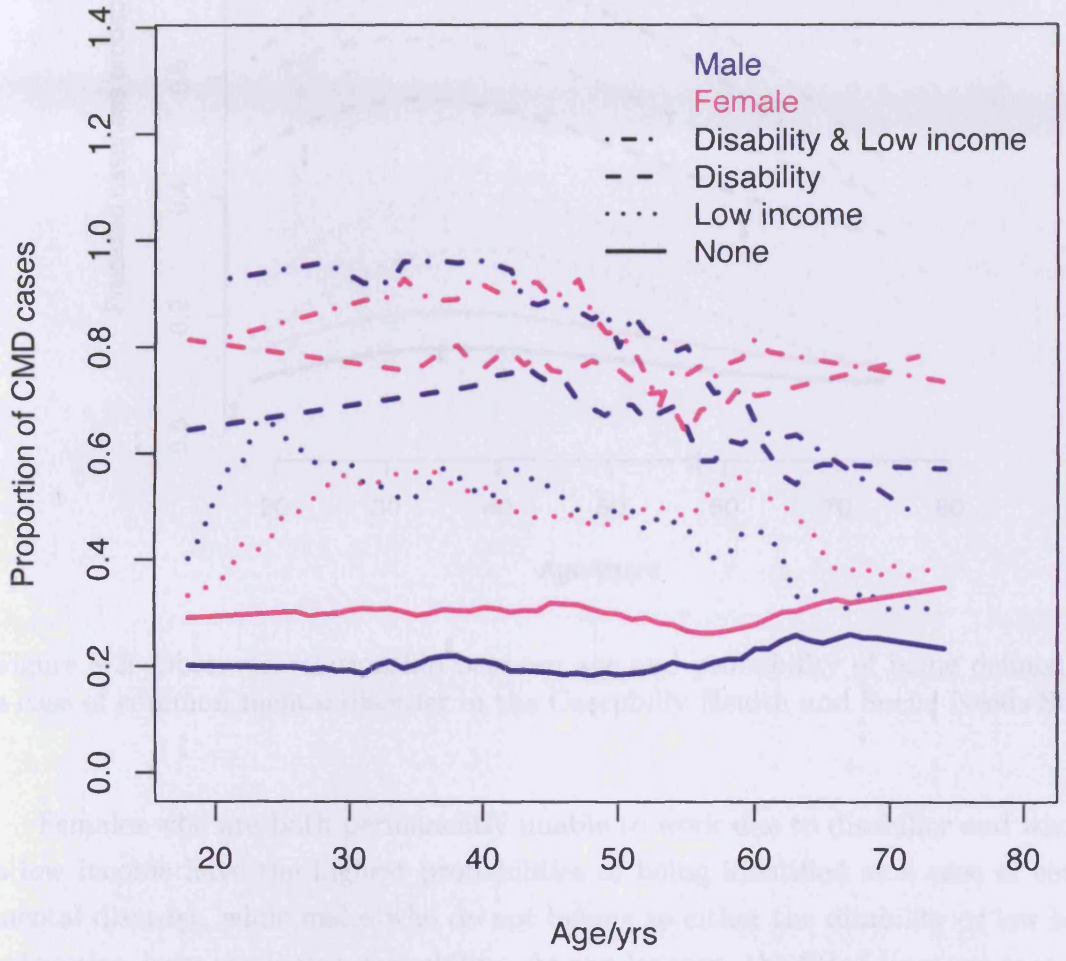
Model 1 modifies the CHSNS model by replacing the Incapacity Claimant Ratio with the reported percent disability coefficient introduced in section 7.5.2. The reason for this is that the ICR was calculated using ward-level information supplied by the DWP making it unsuitable for inclusion in a model which employs synthetic (and not administrative ward) boundaries. The percent disability variable on the other hand can be calculated for any area based on postcodes. This is because it is based on aggregated individual-level data which is geo-coded to postcodes. This makes it a more suitable variable to include when comparing models with different hierarchies. Since the ICR was calculated using a z-transformation (subtracting the mean and dividing by the standard deviation), the percent disability coefficient is also z-transformed. The original percent disability variable ranged between 1% and 28%, while under the z-transformation it ranges between -2.085 and 3.026. The percent disability variable was compared with the ICR in section 7.5.2, and showed that the two variables were highly related.

Model 2

Model 2 modifies model 1 by changing the way age is modelled. It is perhaps instructive to illustrate the raw data here, in order to inform the choice of how best to model this relationship. Figure 8.1 shows the relationship between age in years and the proportion of each subset of the data that is a case. The pink lines (indicating females) tend to be higher than their blue counterparts (indicating males). There is no strong relationship between the two variables for any of the subsets. There is some indication that the probability of being a case increases between 40 and 60; however there are a number of inexplicable peaks and troughs throughout the age range. This plot shows that

whatever relationship exists between these variables, it is not a strong one, and is quite complicated.

Figure 8.1: Proportion of caseness for each age, with overlayed smoothed line



Lines are smoothed lines based on the raw data

The relationship between the response variable and age in the CHSNS model is questionable, since polynomials are notoriously unreliable in the tails. Moreover, the age variable features in no less than four interaction terms (the age variable is interacted with female gender, being on permanently sick/disabled and having a low income, and there is a fourth interaction between the square of age and having a low income) making its effect difficult to interpret. The relationship between age and mental health found in the CHSNS model is plotted in figure 8.2, and shows the effect of each of the interaction terms on the probability of being defined as a case of common mental disorder.

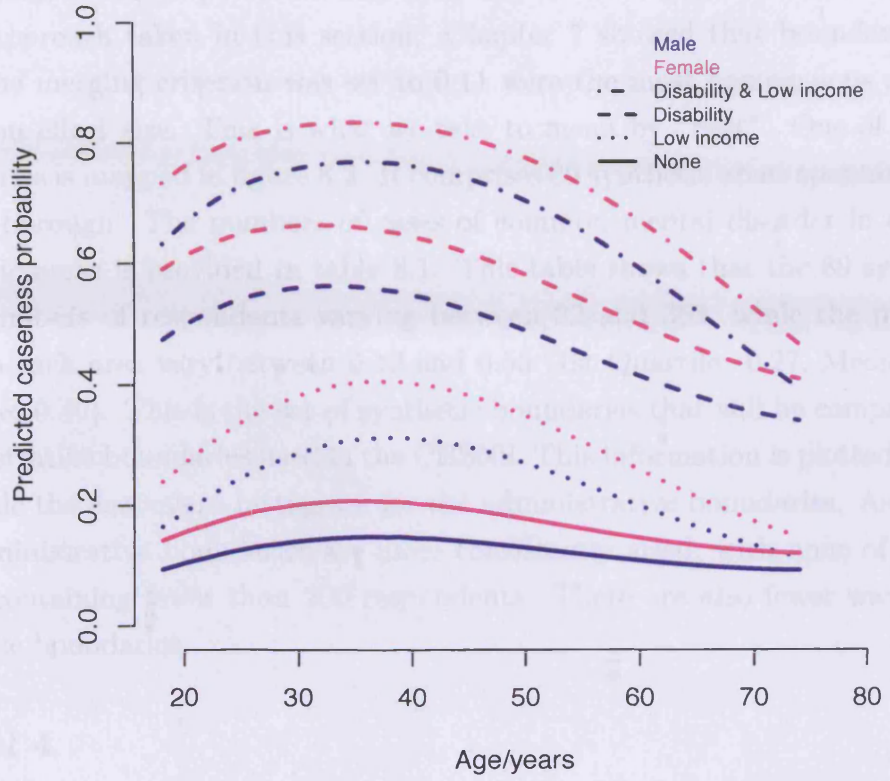


Figure 8.2: Observed relationship between age and probability of being defined to be a case of common mental disorder in the Caerphilly Health and Social Needs Study

Females who are both permanently unable to work due to disability and who have a low income have the highest probabilities of being identified as a case of common mental disorder, while males who do not belong to either the disability or low income categories, have the lowest probability. As can be seen, the fitted lines are reasonably close to what is observed in the raw data plotted in figure 8.1, however there is still room for improvement. The approach taken in Model 2 is to split the age variable into 5-year age groups and model it as a categorical variable. This approach makes no assumptions about the nature of the relationship between age and mental health.

Model 3

Model 3 is the same as model 2 except that the top level in the hierarchy is changed from electoral ward to synthetic boundary. Chapter 7 produced many sets of boundaries in order to assess how robust the results of multilevel models are when the top-level is partitioned in different ways. Here these boundaries are used in a different way. As

discussed in section 7.3, the boundaries produced by the synthetic ward algorithm are not optimal solutions. If a single set of boundaries is required by a user, a large number of boundaries can be produced and then the "best" set of boundaries chosen, which is the approach taken in this section. Chapter 7 showed that boundaries produced when the merging criterion was set to 0.11 were the most homogenous as defined by Wilcoxon effect size. This is what we take to mean by "best". One of these sets of boundaries is mapped in figure 8.3. It comprises 89 synthetic areas spanning Caerphilly county borough. The numbers of cases of common mental disorder in each of these synthetic areas is provided in table 8.1. This table shows that the 89 synthetic areas have numbers of respondents varying between 22 and 363, while the proportions of cases in each area vary between 0.13 and 0.53 (1st Quartile: 0.27, Median: 0.32, 3rd Quartile: 0.40). This is the set of synthetic boundaries that will be compared with the administrative boundaries used in the CHSNS. This information is plotted in figure 8.4, alongside the equivalent histogram for the administrative boundaries. As can be seen, the administrative boundaries are more consistently sized, with none of the electoral wards containing fewer than 200 respondents. There are also fewer wards (36) than synthetic boundaries.

Model 4

This is the final integrated model and includes household as a level. Household is included since chapter 6 found that even when the sparseness is as extreme as in the CHSNS dataset (where there is an average of 1.08 responses per household) including the sparse level results in less biased fixed effect standard errors.

Summary

Table 8.2 summarises the various models that will be compared in this chapter.

8.3 Comparison of results

Instead of attempting to compare every term in each model, just a few crucial parameters are compared: the AIC (assess model fit) , the variance components (to compare the variance components attributable to the synthetic boundaries with the administrative ones, as well as the variance component associated with the household level), the percent disability coefficient (the area-level covariate), the interaction between percent disability and individual disability (a cross-level interaction) and belonging to council tax bands A or B (in order to assess the effect of including the household level on a household-level variable). These parts of the model are most likely to be sensitive to

Figure 8.3: Set of synthetic boundaries produced with the initial merging criterion set to 0.11

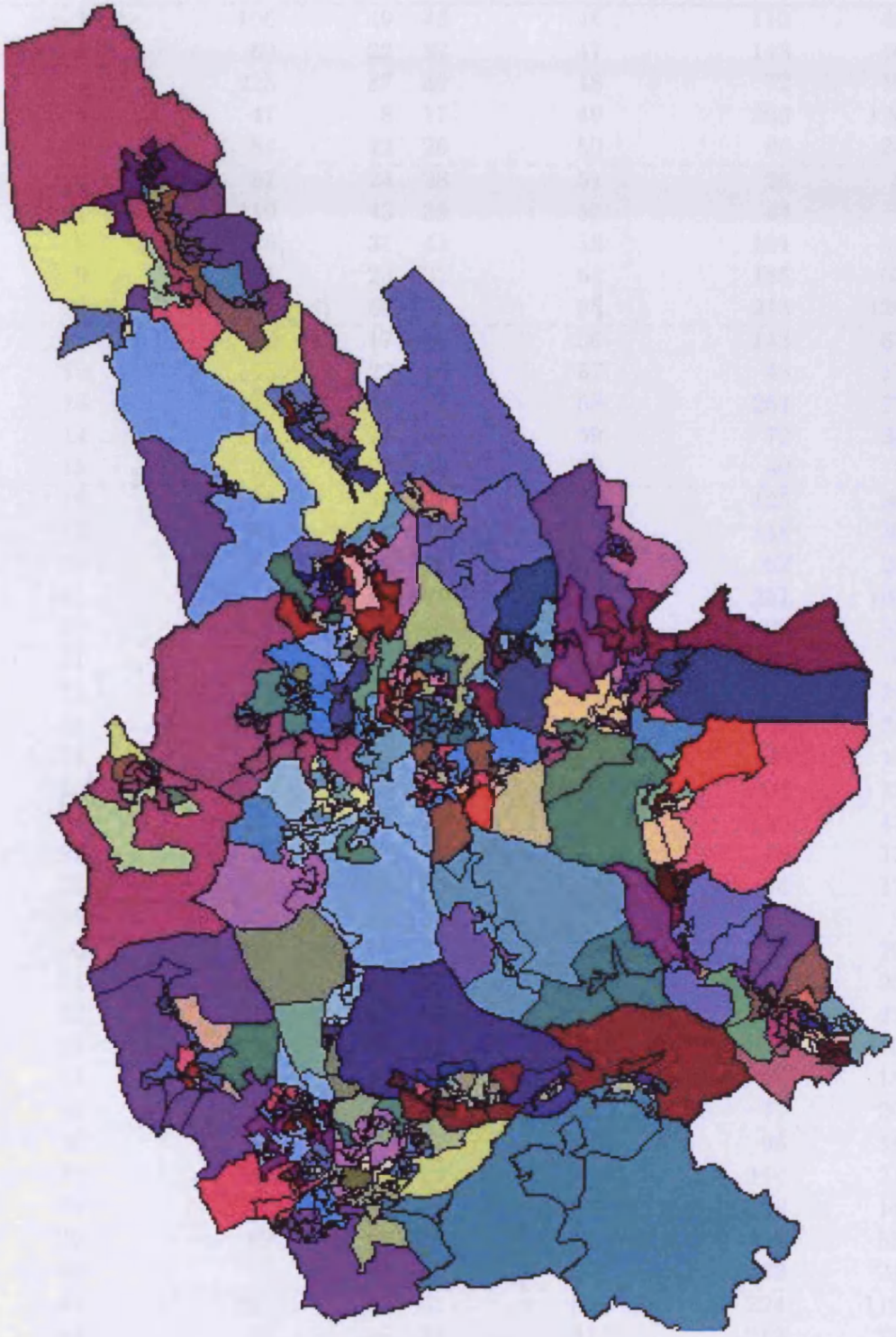


Table 8.1: Number of cases in each synthetic area

Synthetic area number	No. of respondents	No. of cases	%	Synthetic area number	No. of respondents	No. of cases	%
1	106	49	46	46	110	40	36
2	60	22	37	47	143	46	32
3	225	87	39	48	72	19	26
4	47	8	17	49	363	126	35
5	84	22	26	50	86	24	28
6	87	24	28	51	26	9	35
7	110	43	39	52	44	7	16
8	75	31	41	53	101	32	32
9	71	23	32	54	185	68	37
10	167	65	39	55	313	126	40
11	43	17	40	56	143	65	45
12	55	22	40	57	43	17	40
13	173	54	31	58	254	72	28
14	155	74	48	59	72	30	42
15	62	20	32	60	40	8	20
16	136	35	26	61	125	36	29
17	153	34	22	62	146	30	21
18	41	6	15	63	62	26	42
19	53	22	42	64	257	104	40
20	115	34	30	65	96	51	53
21	65	30	46	66	40	8	20
22	25	4	16	67	100	29	29
23	239	51	21	68	47	21	45
24	78	29	37	69	48	15	31
25	120	46	38	70	94	27	29
26	65	25	38	71	125	42	34
27	297	77	26	72	49	13	27
28	212	61	29	73	66	17	26
29	211	74	35	74	26	9	35
30	48	11	23	75	78	29	37
31	283	80	28	76	90	25	28
32	208	80	38	77	101	42	42
33	161	64	40	78	22	10	45
34	97	42	43	79	64	18	28
35	56	17	30	80	77	20	26
36	62	23	37	81	96	38	40
37	348	101	29	82	110	21	19
38	271	59	22	83	110	14	13
39	69	18	26	84	199	51	26
40	75	24	32	85	122	24	20
41	261	80	31	86	224	112	50
42	42	16	38	87	212	75	35
43	112	46	41	88	230	63	27
44	97	36	37	89	49	16	33
45	73	34	47				

Figure 8.4: Ranked number of respondents in each administrative and synthetic area, with case proportions and associated confidence intervals in red

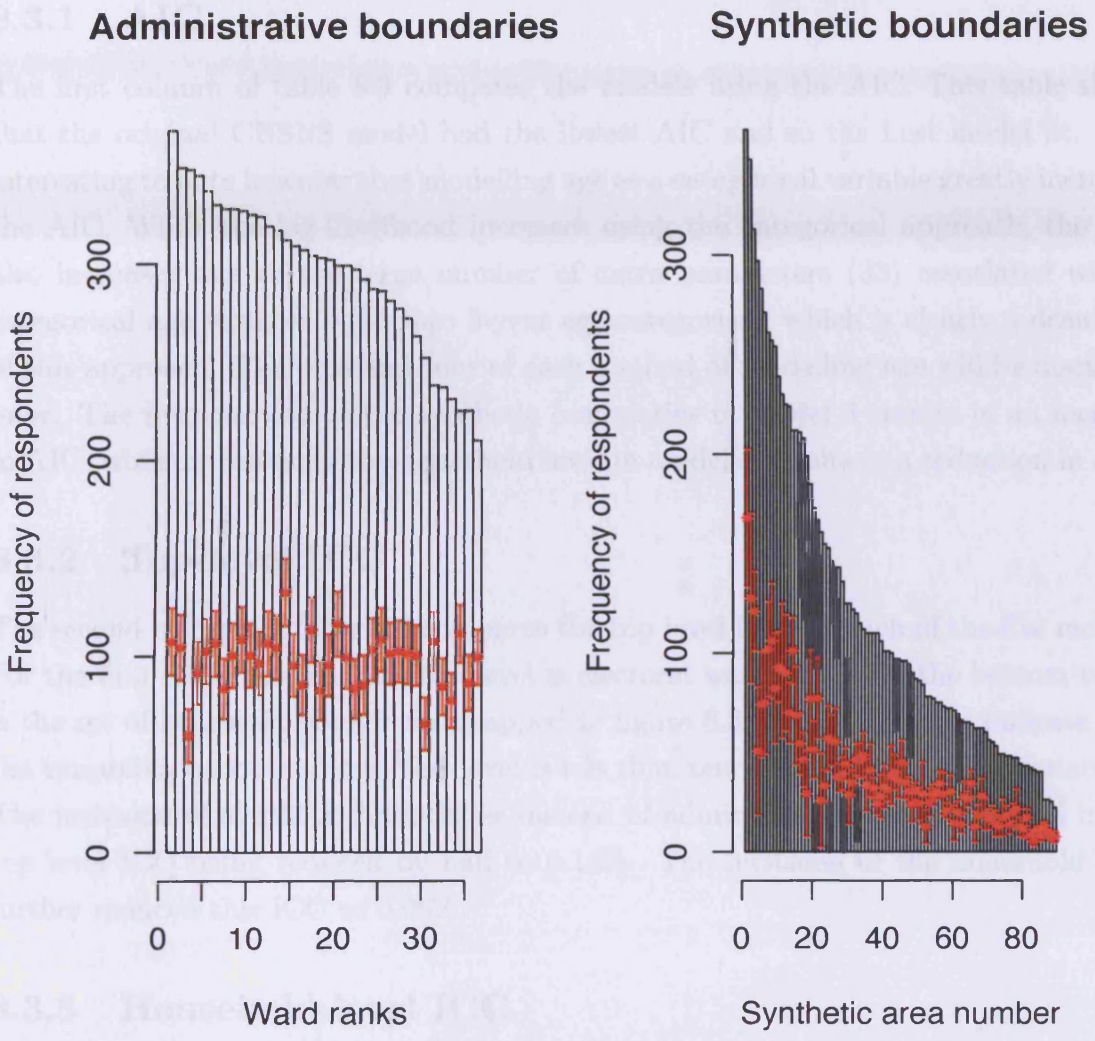


Table 8.2: Model summaries

Model name	Model description
CHSNS	Original CHSNS study model
Model 1	Replaces ICR with Reported Percent Disability
Model 2	Replaces cubic age with categorical age
Model 3	Replaces ward level with synthetic area level
Model 4	Includes household level

the changes made. The results for these variables are summarised in table 8.3. The relationship between age and mental health between model 4 and the CHSNS model will also be examined.

8.3.1 AIC

The first column of table 8.3 compares the models using the AIC. This table shows that the original CHSNS model had the lowest AIC and so the best model fit. It is interesting to note however that modelling age as a categorical variable greatly increases the AIC. While the log-likelihood increases using the categorical approach, the AIC also increases due to the large number of extra parameters (33) associated with a categorical age variable (split into 5-year age categories), which is clearly a drawback of this approach. The pros and cons of each method of modelling age will be discussed later. The introduction of the synthetic boundaries in model 3 results in an increase in AIC, while introducing the household level in model 4 results in a reduction in AIC.

8.3.2 Top-level ICC

The second column of table 8.3 compares the top level ICC for each of the five models. For the first three models, this top level is electoral ward, while for the bottom two it is the set of synthetic boundaries mapped in figure 8.3. All five models indicate that the variability attributable to this level is less than one percent of the total variability. The inclusion of synthetic boundaries instead of administrative wards resulted in the top level ICC being reduced by half to 0.14%. The inclusion of the household level further reduced this ICC to 0.06%.

8.3.3 Household-level ICC

In model 4 the household level ICC is 6.78%. This is a much larger ICC than for the top level, and indicates that households are a more important context to model than areas. Studies which exclude household from the hierarchy will produced overestimated area level ICCs, since some of the variation that should be attributed to households will be erroneously attributed to areas. This finding should be interpreted with caution however, since figure 6.13 showed that when the sparseness is as high as it is in the CHSNS (1.08 respondents per household on average), the household level variance component is estimated with large variability.

The household-level variance component found in model 4 is also in line with the literature recommending that household is an important context to include in multilevel modelling (Chandola et al., 2003, 2005; Propper et al., 2005; Weich et al., 2003b, 2006, 2005). The aforementioned literature however indicates that the size of the household-

Table 8.3: Comparison of CHSNS study with the new analysis

	AIC	Top level ICC (%)	Household level ICC (%)	% disability OR			% disability*Indiv. disability OR			Council Tax Band A & B		
				Est.	L[a]	U[b]	Est.	L	U	Est.	L	U
CHSNS	11,952	0.27		1.06	1.00	1.12	1.18	1.04	1.35	1.29	1.15	1.44
Model 1	11,955	0.27		1.04	0.99	1.11	1.17	1.02	1.34	1.30	1.16	1.46
Model 2	12,000	0.31		1.05	0.99	1.11	1.17	1.02	1.34	1.29	1.16	1.45
Model 3	12,006	0.14		1.05	1.00	1.11	1.15	1.00	1.32	1.28	1.15	1.44
Model 4	12,002	0.06	6.78	1.05	0.99	1.11	1.16	1.01	1.33	1.30	1.16	1.47

Estimates in bounded boxes indicate statistical significance at the 5% level

[a]L indicates the lower 95% confidence limit

[b]U indicates the upper 95% confidence limit

level variance component is between9% and 29%, whereas this work estimates it to be about 7%. This is not a large discrepancy however, especially considering the caveat, mentioned above, that this level's variance component is estimated imprecisely due to the sparseness.

8.3.4 Percent disability coefficient

The estimate of the percent disability coefficient varies little across all five models, with its odds ratio being consistently estimated to be about 1.05, and never attains significance at the 5% level. The confidence intervals around this odds ratio vary only at the second decimal place. This is not to say however that changing the hierarchy will never affect fixed effect estimation, since figure 7.18 illustrated the large range of coefficients that can be produced by changing the hierarchy. The fact that this coefficient changes only slightly in Model 3, compared with the CHSNS model, despite the fact that the ICR is calculated for 36 electoral wards in the latter, while the percentage disability variable is calculated for 89 synthetic areas in the former provides further evidence that these two variables are measuring the same thing .

8.3.5 Cross-level interaction between percent disability and individual disability

The interaction between percent disability and individual disability is similarly unaffected by the changing models, and consistently attains significance at the 5% level. This interaction was only of borderline significance in the CHSNS model and remains so as the model is adjusted according to the thesis findings.

8.3.6 Council Tax Band

This household-level variable is also consistently estimated across the models. This is perhaps surprising, since including the household as a level (model 4) should increase the standard errors on all the parameters. This is not observed, with the confidence intervals around this parameter being no wider in Model 4 than in any of the other models. This is most likely due to fact that there are very few multiple response households in the dataset (792 out of 9,827), and so the difference between treating this variable as an individual- or household-level variable is not substantial.

8.3.7 Age

The relationship between age and mental health was modelled differently between models 1 and 2. The difference in predicted values between the original CHSNS model

and model 4 is illustrated in figures 8.5 (this is the same as figure 8.2, but reproduced here for comparison purposes) and 8.6. The original CHSNS model fitted a cubic polynomial to the age variable, as well as fitting four interaction terms involving age. Model 2 on the other hand split the age variable into 5-year age groups and treated it as a categorical variable. Interactions between this categorical age variable and female gender, individual disability and low household income were retained. Both approaches have their strengths and weaknesses. The cubic polynomial approach can be criticised for being too restrictive, in that it forces a certain form on the relationship. Polynomials are also unreliable in the tails. The categorical approach however involves the loss of age information. Moreover, as can be seen from figure 8.6, this approach treats each age category completely independently of the age categories adjacent to it. This can lead to jumps between age categories which may be due to small sample sizes. The polynomial approach avoids this by fitting a prediction line across the entire age range simultaneously. This may have more biological plausibility than the categorical approach. For instance the low income subgroups prediction lines for model 4 have three local maxima. There is no obvious explanation for why mental health caseness probabilities might dip around the ages of 25, 35 and 50. It would appear however that broadly speaking the polynomial approach is a reasonable approximation for what is observed in the data.

8.4 Strengths and Limitations

The strengths and limitations of each model compared in this chapter will be discussed in this section.

8.4.1 CHSNS model

The original CHSNS dataset was generally well-suited to the analysis of the contextual determinants of mental health. It is based on a large sample size of 10,653, set in a well-defined and contiguous study area, has a good response rate of 62.7%, collected a wide range of demographic information on the respondents and strengthened this information by linking with other data sources such as the DWP. The model itself acknowledges the presence of clustering the data by fitting a two-level hierarchy to the data.

Having said that, the model still suffers from a number of methodological limitations which motivated this thesis. These limitations included a skewed response variable which was dichotomised using a cutpoint based on prevalence matching, a sparse contextual level that was excluded from the analysis and a reliance on administrative areas to serve as proxies for area of residence. Another criticism that could be

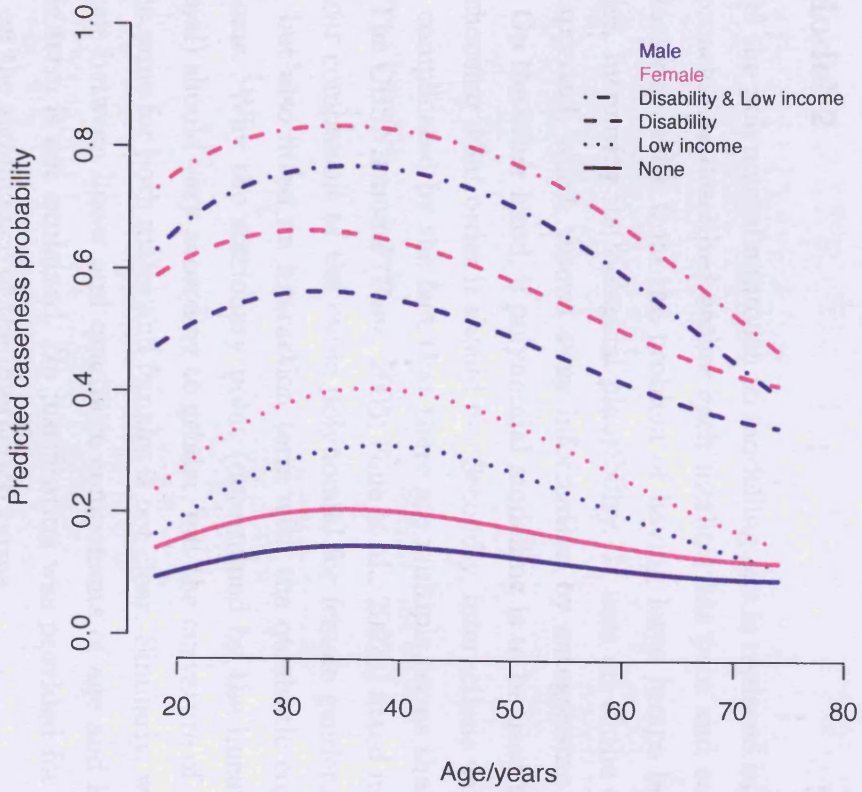


Figure 8.5: Predicted relationship between age and probability of being defined to be a case of common mental disorder in the Caerphilly Health and Social Needs Study

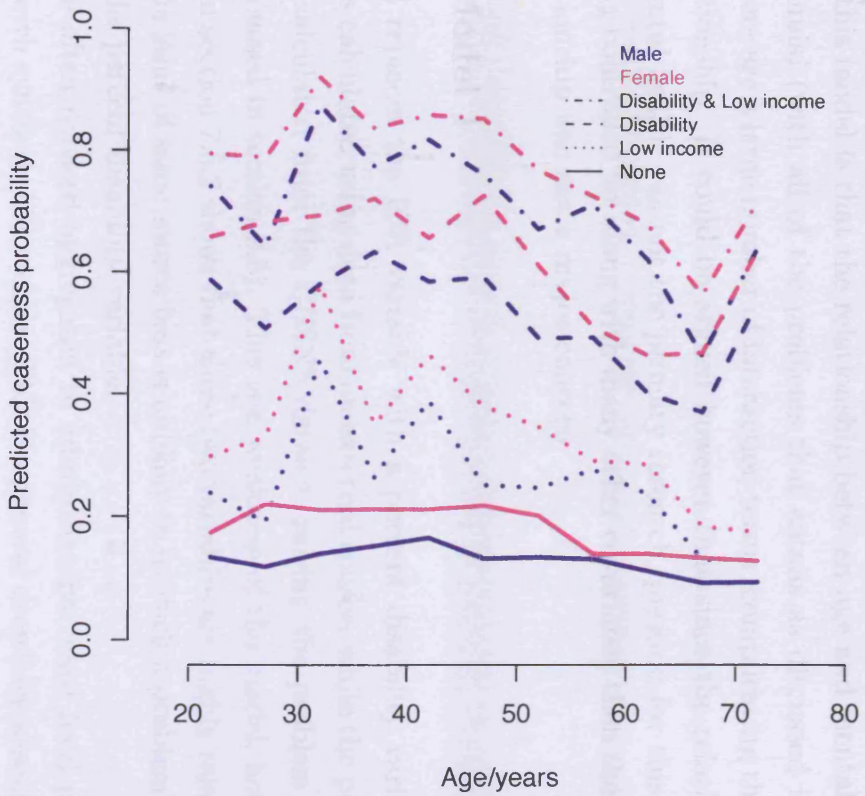


Figure 8.6: Predicted relationship between age and probability of being defined to be a case of common mental disorder in model 4

levelled at this model is that the relationship between age and mental health is fitted as a polynomial (with all of the problems that entails as discussed in section 8.3.7) and that there are a large number of interaction terms, complicating the interpretation of this relationship. It could be argued however that since the relationship between age and mental health was not the primary research question for this study, and was merely being controlled for along with many other covariates, then the interpretability of this relationship was not a major concern.

8.4.2 Model 1

This model replaces the ICR variable with a percent disability variable. The ICR variable was calculated using data from an external source, while the percent disability variable is calculated from the CHSNS dataset, raising the problem of same source bias (as discussed in section 2.5). This is a weakness of the model, however, the work illustrated in section 7.5.2 shows that these two variables are highly related (correlation of 0.89). The issue of same source bias is unlikely to be such a problem as to invalidate the use of the percent disability variable.

Data are often released aggregated to administrative ward level precluding them being used with synthetic hierarchies. This is a general drawback associated with using bespoke boundaries rather than administrative ones.

8.4.3 Model 2

In this model the polynomial approach to modelling age is replaced in favour of a categorical approach. As described earlier each method has pros and cons. Polynomial modelling does not suffer from the problem of having large jumps between adjacent age categories, increasing its biological plausibility. It uses all of the data, unlike the categorical approach which throws away information by categorising age into 5-year age groups. On the other hand, if polynomial modelling is to be used then there is the problem of choosing what order it should be. Secondly, interactions with a polynomial variable are complicated by the fact that there are multiple terms that need to be investigated. The CHSNS model (Fone, 2005; Fone et al., 2007c) fitted interaction terms with the linear component of the cubic polynomial for female gender, incapacity and low income, but also fitted an interaction term with the quadratic component of age and low income. Why the stationary point (determined by the linear component of the polynomial) should vary according to gender, but the curvature of the relationship should be the same for both males and females is not clear. Similarly, why there should be interactions between linear and quadratic components of age and low income, but not the cubic term is not explained. No justification was provided for this other than being based on the significance of the interaction terms.

This model avoids these problems by treating the age variable as categorical. Of course, the drawback to this approach is the large number of extra parameters that need to be estimated, especially when interaction terms are included. With a large sample size however this is less of a problem, as there is enough information there to properly estimate these extra parameters. The two approaches produce reasonably similar results, however the categorical approach is more easily defended than a polynomial approach, since it does not assume anything about the nature of the relationship between age and mental health.

8.4.4 Model 3

This model replaces the administrative ward level, with the synthetic ward level. There are a number of weaknesses associated with using synthetic wards. As discussed earlier, using bespoke boundaries often necessitates the use of other bespoke measures to describe them. Moreover if a researcher decides to create synthetic boundaries for use in a specific analysis, then other researchers may have difficulty repeating their analysis, due to not having access to the synthetic boundaries. This makes it harder to compare between studies. Researchers using administrative boundaries have no such problems, due to the fact that synthetic boundaries are freely available.

Another problem with the synthetic boundary approach is that it is not obvious how to choose which set to use. The approach used here is that a large number of boundaries were created and the "best" chosen. A criterion to measure how good a set of boundaries must then be selected, which is a subjective process. In this chapter it was decided that boundaries created when the merging criterion was set to 0.11 produced the most homogenous boundaries. This in turn was based on another subjective process; choice of homogenising variables. As discussed in section 7.4 the choice of which variables to use to create the synthetic boundaries is crucial, but there is little to guide researchers about how this should be done. In this study homogenising variables were chosen based on whether they were known to be useful predictors of mental health. The number of homogenising variables (6) was chosen to provide a compromise between having a comprehensive set of mental health predictors and having an algorithm that ran quickly.

Furthermore, the computation time for creating a large number of boundaries from which the best is chosen is not trivial. The CPU time taken to create the 25,000 sets of boundaries used in chapter 7 was 20 weeks as described in section 7.6. This approach may be ruled out in many situation therefore due to time constraints.

The final weakness of the synthetic boundary approach used in this model is that the number of top-level areas are not the same across the models. There are 36 administrative wards in Caerphilly, but there are 89 synthetic areas used in this model.

The synthetic boundaries are therefore smaller and so more likely to be homogenous. The synthetic boundaries are not constrained by size either, meaning that some contain large numbers of respondents and some small numbers. This is in contrast with the administrative boundaries which are much more consistent in size (see 8.4). What effect this might have on the precision of the results of the models is uncertain. If these synthetic boundaries meaningful contextual levels that affect mental health, then the differential sample sizes are not a problem as they reflect reality. If they do not represent meaningful groupings however, the unbalanced sample sizes are likely to reduce precision.

There are strengths associated with Model 3 also. Firstly, there have been calls from the literature for hierarchies based on homogeneity instead of hierarchies based on administrative boundaries (Rice et al., 1998; Macintyre et al., 2002; Diez-Roux, 2003). The synthetic boundary approach is more theoretically based than using administrative boundaries which are unrelated to the question of interest.

More generally, Model 3 has answered this call from the literature, created synthetic boundaries and applied them to a real-life dataset in order to assess how large an impact the use of bespoke boundaries has. Models like this can be used to provide sensitivity analyses when a parameter estimate is based on an arbitrary hierarchy, and researchers want to be sure that the results they are uncovering are not specific to the hierarchy they are using, but instead say something more general about area effects on health.

8.4.5 Model 4

Model four includes the sparse household level in the multilevel hierarchy. The strengths of this model are that it uses more of the information collected by the CHSNS, acknowledges a level of context that is present in the data so that standard errors are not underestimated at that level, and is furthermore acknowledging a call from the literature (Chandola et al., 2003, 2005; Propper et al., 2005; Weich et al., 2003b, 2006, 2005) that the household level be included in studies of area effects on mental health.

A practical drawback to the modelling of the sparse household level is that the time taken for the models to run is greatly increased. The computing time in R was tripled compared with the two level hierarchies, and in MLwiN (Rasbash et al., 2003), the models fail to converge.

Another weakness of this model is that while chapter 6 showed that including a sparse level resulted in better estimates of fixed effect standard errors than excluding a sparse level, it also demonstrated that when the sparseness of the included level is very high (as it is here) that the precision of the variance component for that level is very low. Hence the ICC quoted in table 8.3 must be treated with caution.

Nonetheless, Model 4 is the culmination of the work of the thesis and provides the most methodologically sound and comprehensive analysis of the CHSNS data that is available.

8.5 Conclusion

This chapter incorporated the various findings of this thesis in a single analysis of the CHSNS and essentially found that they did not make a large difference to elucidating the association between area of residence and mental health. The recommendation with the largest impact was including the household level, which reduced the top-level ICC, and resulted in 7% of the variation in the response being attributed to that level. In order to properly interpret this finding it is important to view it in conjunction with the findings of the individual chapters. While this particular analysis found that the effect of including the sparse household level was not observed to affect the household-level covariate standard error (expressed here through a confidence interval around the odds ratio) chapter 6 found that it can do.

This chapter also adds to the literature regarding the suitability of administrative areas for use in the hierarchy of analyses investigating associations between mental health and area of residence (Rice et al., 1998; Macintyre et al., 2002; Diez-Roux, 2003). This chapter investigated the use of synthetic boundaries for the CHSNS data and found that using them produced very similar results to those produced when administrative hierarchies were used, indicating that administrative areas may perform better than has previously been assumed. However the simulation in chapter 7 found that it can potentially produce large differences. The message remains that studies which employ "convenience" hierarchies (i.e. hierarchies unrelated to the research question) would do well to consider performing sensitivity analyses with alternate hierarchies to assess the robustness of their results.

Chapter 9

Conclusion

This chapter will summarise the results of the thesis, with specific reference to the stated objectives from chapter 1, presented below:

1. To investigate the spatial variation of common mental disorders in Caerphilly county borough in a Bayesian framework.

2. To investigate properties surrounding the distribution of the mental health score used in the study, as well as evaluating a cutpoint to identify cases of common mental disorder.

3. To investigate the robustness of multilevel modelling techniques to sparse levels of data.

4. To develop an algorithm that can partition an area into internally homogenous areas, using data from the Caerphilly Health and Social Needs Study as an example.

5. To compare (quantitatively) the operationalisations of area for both administrative and synthetic boundaries

Furthermore, the work will be placed in context for researchers who are interested in the practical implications of the work. Finally, areas requiring further work will be highlighted and suggestions for how those further investigations might proceed will be provided.

9.1 Summary of results

9.1.1 Investigating the spatial variation of common mental disorders in Caerphilly county borough

The CHSNS dataset provided excellent motivation for this thesis. With a large sample size, and a large amount of information, it was generally well-suited to the analysis of the contextual determinants of mental health. The CHSNS analysis (Fone, 2005) was repeated using the R programming language and the results were found to agree closely with the results produced by MLwiN.

Plotting the raw data showed that there was spatial variation in mental health in Caerphilly county borough. Bayesian methods were introduced and used to smooth this raw data, to provide a more reliable picture of the spatial variation in mental health. The smoothed information was mapped and provided evidence that the spatial variation in mental health was not an artefact of small sample sizes. The existence of spatial variation in Caerphilly county borough provided crucial motivation for the rest of the thesis.

9.1.2 Modelling mental health

The stated objective for this section was to investigate the distributional properties of the MHI-5 mental health scale as well as evaluating a cutpoint to identify a case of common mental disorder. Chapter 3 began with a description of the main mental health measure used in the study, the MHI-5. The next section explored the validity and reliability of the MHI-5, firstly introducing some important concepts and methods, and then summarising some of the validation literature. The MHI-5 was shown to be a valid and reliable scale, and a useful measure of mental health.

Having introduced the mental health scale, different methods of modelling it were investigated. A Box-Cox transformation was performed indicating that a square transform is the best way to normalise the response. The reduction in skewness however was not deemed sufficient to justify the increased complication of parameter interpretation and so this approach was not used. Ordinal modelling was introduced as an alternative. Various ordinal modelling approaches were described and the cumulative logit method was illustrated with reference to the CHSNS dataset. Ordinal modelling was not used because of the strict assumptions and difficulty in interpreting the results. Finally, different methods for dichotomising the MHI-5 scale as well as the SF-36 Mental Component Summary score were compared and contrasted. The dichotomisation was performed using ROC curve analysis. This procedure was introduced and explained along with various criteria for determining a cutpoint. These comprised the most common methods used in ROC curve analysis, namely, the Youden Index, the shortest

distance to the upper left corner, the misclassification rate, the minimax criterion, and prevalence matching. This indicated that cutpoints of less than or equal to 76 and less than or equal to 51.7 for the MHI-5 and MCS are the most generalisable cutpoints. For UK populations , cutpoints of less than or equal to 60 and less than or equal to 44.8 for the MHI-5 and MCS minimise the missclassification rate. Furthermore, the Youden Index and the point closest to the upper left corner were identified as the methods least dependent on population prevalence. These were recommended as the most suitable methods to provide cutpoints on scales intended for use across large geographical areas where the prevalence of cases (whatever a case may mean in the given context) is likely to vary widely.

This chapter investigated alternative approaches to modelling the scores from the MHI-5 measure as a Normal variable. Each had advantages and disadvantages. It was decided that for the investigations into multilevel modelling in the following chapters it was most appropriate to model the variable as a Normal one.

9.1.3 Impact of sparse levels on hierarchical modelling

An investigation of sparse levels in multilevel analysis was presented in chapter 6. The objective was to investigate the robustness of multilevel modelling to sparse levels of data. Essentially, this comprised both questions regarding the effect on the results of a multilevel analysis of including a sparse level, and excluding a sparse level. This was investigated using four simulation studies: where both household and individual contributed equal variance components (scenario A), where the household variance component was small (scenario B), where the household variance component was zero (scenario C), and where the household variance was small and the total sample size was small (scenario D). The results of each of these scenarios are summarised in detail at the ends of sections 6.4.1, 6.4.2, 6.4.3, and 6.4.4. Here an overview of the results of all of these scenarios is provided. The overall findings of this simulation study are now summarised briefly under the following headings: effect of relative size of the sparse level's variance component, effect of total sample size, effect of including or excluding a sparse level on fixed effect estimation, effect of including an uninformative sparse level.

Effect of relative size of the sparse level's variance component

Scenarios A and B modelled similar situations, but the important difference between them was the relative contribution of the sparse level (large in A, and small in B). With respect to ICC estimation, the smaller ICC in scenario B was estimated less precisely under sparseness than the larger ICC in scenario A. When the sparse level was excluded the area-level was overestimated in scenario A, but not in scenario B. Household-level fixed effects were more seriously underestimated when the contribution

from the household-level was large.

Effect of total sample size

Scenarios B and D utilised the same variance structure (with the area, household and individual variance components being 0.5, 1.5 and 20 respectively), except scenario B had a sample size of 10,000, while scenario D had a sample size of 1,000. As expected, the smaller sample size accentuated the problems observed when the sample size was larger. Household-level ICCs were more variable at sparse levels. The underestimation of the household-level standard errors was also of a larger magnitude.

Effect of including or excluding a sparse level on fixed effect estimation

In general, the simulation study suggests that it is better to include sparse levels (even with small variance contributions and high sparseness) into a multilevel analysis, since the underestimation of household-level fixed effect standard errors is larger in the models which exclude a sparse contextual level. Area-level fixed effect standard errors were much less affected by sparse household levels.

Effect of including an uninformative sparse level

ICC coefficients and model fit were investigated for scenario C, where the household level was uninformative. While most of the simulations attributed only a tiny proportion of the variability to the household-level, some of the ICC coefficients for an extremely sparse yet uninformative level were almost as high as 0.15. The results of this scenario investigation indicate that including an uninformative level does not lead to hugely overestimated ICCs at that level (they must always be somewhat overestimated since the true value is zero), except for when the average number of respondents per household is less than 1.5.

9.1.4 Synthetic area algorithm

An algorithm for partitioning areas into contiguous sub-areas was developed and introduced in chapter 7. The algorithm uses information on the spatial geography and composition of areas to create homogenous areas. It is a generalisable algorithm and can be applied to any geographical area with adjacency information. An illustration of how the algorithm operates was provided in figure 7.3. A set of boundaries was produced for comparison with the 1991 census electoral wards and plotted in figure 7.4, providing an example of how the algorithm could be used to produce a single set of boundaries. This algorithm was then used in a simulation study to compare the

synthetic boundaries with the administrative boundaries, both in terms of homogeneity and their implementations in hierarchical models. The results of this chapter have already been summarised in section 7.8. Here a brief summary will be provided under the following headings: internal homogenity, variance components, model fit and coefficient estimation.

Internal homogenity

The synthetic boundaries demonstrated better internal homogeneity than the administrative ones as measured by Wilcoxon effect sizes of the homogenising variables. In terms of IQVs however, the administrative boundaries performed remarkably well, having a smaller LSOA IQV (calculated based on the six homogenising variables) than the majority of synthetic boundaries. There are two possible explanations for this, as discussed in section 7.8: firstly, the administrative areas could indeed be more homogenous in terms of IQVs than the synthetic boundaries produced. Alternatively, the inclusion of council tax band as a homogenising variable may have restricted the homogeneity of the synthetic areas due to the fact that the average Wilcoxon effect sizes for the council tax band were larger than for the other five homogenising variables (as shown in figure 7.9).

Variance components

The variance components are expressed through ICC coefficients in this section. Four sets of variance components, resulting from the two models described in equations 7.4 and 7.5, as well as the two cross-classified models (which are the same as the aforementioned models except they contain both synthetic and administrative hierarchies) were investigated. The single hierarchy models illustrated that the synthetic boundaries created larger area-level ICC coefficients in the null model than the administrative boundaries, when the initial merging threshold was set below 0.3. This indicates that the synthetic boundaries are indeed capturing more area homogeneity than the administrative areas. When individual- and area-level covariates are included the synthetic ICC was reduced more than the administrative ICC. The fact that in the null model the synthetic ICC was larger and in a fuller model it was smaller, indicates that it is possible that the large size of the administrative ICC is a result of heterogeneity at that level meaning that the area-level explanatory variable is less meaningful than it could be, and so explains rather less variation than it should.

Cross-classified models display a similar pattern and again imply that the large administrative ICC may be a product of the area-level covariate being less meaningful for the administrative areas than for the synthetic areas.

Model fit

Model fit, as assessed by the AIC was examined for the models described in equations 7.4 and 7.5. The AICs for the null model with synthetic hierarchies were lower (indicating better model fit) when the initial merging threshold for the algorithm was set close to 0.11. These low AICs were achieved even with fewer synthetic areas than the 110 LSOAs against which they are compared. For the model with individual- and area-level covariates the AICs from the models using synthetic hierarchies were almost always smaller than the administrative boundaries. This represents evidence that the synthetic areas improve the operationalisation of area of residence in multilevel modelling analyses.

Coefficient estimation

The percent disability coefficient was investigated for the model in equation 7.5. The relationship between the sparseness and the magnitude of this fixed effect was striking in figure 7.18. Lower merging thresholds tended to produce lower coefficients. There is no true coefficient to be compared against here, but there is still an important message contained in this figure. By essentially shuffling the middle-level units into different higher-level units, the percent disability coefficient ranged between -35 and -11. This is a huge range and provides graphic evidence that the choice of hierarchy is crucial in multilevel modelling.

9.1.5 Area effects on health

Chapter 8 addressed the fundamental research question that motivated the work of the thesis; what are the contextual determinants of mental health.

9.2 Implications for researchers

There are implications for researchers from the three methodological objectives of the thesis. These implications are now described under the following headings: modelling mental health using the MHI-5, recommendations for sparse levels of context in multilevel analyses, and homogenous synthetic boundaries.

Modelling mental health using the MHI-5

Chapter 3 investigated a number of different ways to deal with a skewed response variable. None of these methods were a perfect solution and all approaches complicated the interpretation of parameters, however researchers should investigate these approaches when dealing with a skewed response. Binomial modelling was investigated with the

goal of deriving cutpoints to define a case of common mental disorder on the MHI-5 and MCS. Moreover, the methods of deriving cutpoints on ROC curves were compared and the Youden Index and (0,1) methods were recommended as methods that were less prevalence dependent than the other methods examined. If researchers are attempting to identify a cutpoint on a scale for use in areas where the case prevalence may vary widely (international studies for instance), then the aforementioned methods are most suitable. The work of chapter 8 showed that the evidence for contextual determinants of mental health above the level of household is not strong.

Recommendations for sparse levels of context in multilevel analyses

The overall message from the household simulation chapter is that in most cases, even if the variance contribution of the sparse level is small and the degree of sparseness high, it is better to include that level in a multilevel analysis. This results in less underestimation of household-level fixed effect standard errors. This comes at the expense of greater variance component estimation variability.

Homogenous synthetic boundaries

The synthetic boundaries derived for this thesis demonstrate that it is possible to derive synthetic areas with the goal of creating homogenous areas for multilevel research. Moreover, it has been demonstrated that these synthetic boundaries represent an improvement over administrative boundaries in many ways. The work of chapter 7 also serves as a reminder to researchers of the large impact the choice of hierarchy can have on the results of a multilevel modelling analysis. In particular, variance components and fixed effect parameters were shown to vary substantially between models with different hierarchies. This should encourage researchers to spend more time deciding on the hierarchy they will employ.

The overall theme of the results of this thesis are that the potential impact of the choice of hierarchy on the results of a multilevel modelling analysis is large. Including or excluding contextual levels in an analysis can have important implications depending on the characteristics of the level itself. On a practical note, the simulation in chapter 6 indicated that sparse levels with more than 1.5 sub-units per unit on average should be included in analyses.

9.3 Further research

There are three main themes to the study that need investigating further. Firstly, there is the work on modelling the MHI-5. Secondly, there is the work on sparse levels in multilevel analysis. Finally, there is the synthetic boundaries investigation.

9.3.1 Mental health

Various approaches to dealing with the problem of a skewed response variable were taken in chapter 3. None of these were approaches were used in the simulation studies, with Normal modelling being used for ease of interpretation. It would be useful to conduct a simulation study investigating the effect of modelling a skewed response as a normal variable on the results of a multilevel analysis.

This chapter also made use of existing information from the British Household Panel Survey to compare the MHI-5 with the GHQ-12. In order to identify a definitive cutpoint to identify CMD for the MHI-5, it would need to be compared with a diagnostic/clinical interview schedule. This could provide a clinically valid cutpoint for the MHI-5. A difficulty in doing this would be achieving a large enough sample size with a resource-intensive interview schedule.

9.3.2 Sparse levels

The investigation of sparseness in chapter 6 could be extended in many ways. One of the more important considerations would be to model the household responses using different distributions, to see how the results change. Currently, they are modelled using a Poisson distribution since that fits well with the CHSNS dataset. It would not fit as well to the household responses in the BHPS.

The situation modelled in scenario C, where an uninformative level was included in a multilevel analysis could be further investigated. Assessing the magnitude of the variance component likely to be attributed to an uninformative level could be used to augment the current methods of assessing "significance" of variance components. This could be approached using a simulation study, but may be better investigated analytically.

9.3.3 Synthetic boundaries

Again, there are many possible suggestions for extending the work of the synthetic boundaries algorithm. One of the more useful features that should be incorporated into the algorithm is to include a component which restricts the growth of synthetic

areas in terms of population size, once they attain a certain size. This could be used to produce areas with similar populations. The exact method of restricting the growth, and the size at which this restriction should occur are both difficult questions.

Secondly, the effect of the choice of homogenising variables could be further investigated. Ideally, a different outcome would be investigated, for which the homogenising variables would be different to those investigated here. Thirdly, the structure of the algorithm could be altered so that a different measure for determining merging is used.

The algorithm uses seed pairs as initial values to begin the analysis. In the analyses presented in section 7.2 these seed pairs were chosen to be those OAs which were most similar. Another approach would be to generate random seed pairs for each iteration. Alternatively, seed pairs could be chosen in a stratified way so that different sections of the area to be partitioned are equally represented. Yet another way would be to chose seed pairs so that different types of area based on some criterion are equally represented. The effect of such changes on the results of the work presented here would be of interest.

Another area for further research involves the Wilcoxon effect sizes for the six homogenising variables. As was shown in section 7.7.2 the effect sizes for the council tax band were larger than for the other variables. To counteract this, only 5 of the 6 homogenising variables were required to be below the merging threshold, in order for two areas to be merged. An alternative approach would be to standardise these Wilcoxon effect sizes so that each contributed equally. This may lead to even more homogenous boundaries.

9.4 Final summary

The objectives stated in section 1.4 were achieved in this thesis. The overall theme of the results of this thesis are that the impact of the choice of hierarchy on the results of a multilevel modelling analysis is large. Researchers should spend time considering the different levels of context that should be included in a model. These levels should be determined based on theory and not purely convenience.

Bibliography

Andresen, E., Bowley, N., Rothenber, B., et al. (1996) Test-retest performance of a mailed version of the Medical Outcomes Study 36-item Short-Form Health Survey among older adults. *Medical Care*, **34**, 1165–70.

Assunção, R., Neves, M., Câmara, G., et al. (2006) Efficient regionalization techniques for socio-economic geographical units using minimum spanning trees. *International Journal of Geographical Information Science*, **20**, 797–811.

Avery, A., Betts, D., Whittington, A., et al. (1998) The mental and physical health of miners following the 1992 national pit closure programme: a cross sectional survey using General Health Questionnaire GHQ-12 and Short Form SF-36. *Public Health*, **112**, 169–173.

Beale, N., Baker, N., Straker-Cook, D. (2000) Council tax valuation band as marker of deprivation and of general practice workload. *Public Health*, **114**, 260–264.

Bebbington, P., Dunn, G., Jenkins, R. (1998) The influence of age and sex on the prevalence of depressive conditions: report from the National Survey of Psychiatric Morbidity. *Psychological Medicine*, **28**, 9–19.

Berwick, D., Murphy, J., Goldman, P., et al. (1991) Performance of a five-item mental health screening test. *Medical Care*, **29**, 169–176.

Besag, J., York, J., Mollie, A. (1991) Bayesian image-restoration, with 2 applications in spatial statistics. *Annals of the Institute of Statistical Mathematics*, **43**, 1–20.

Birkes, D. (1998) Likelihood ratio tests. In *Encyclopaedia of Biostatistics*, volume 3, pp. 2245–2248, Wiley.

Blakely, T., Woodward, A. (2000) Ecological effects in multi-level studies. *Journal of Epidemiology and Community Health*, **54**, 367–374.

Bland, J., Altman, D. (1986) Statistical methods for assessing agreement between two methods of clinical measurement. *Lancet*, **30**, i:307–310.

Bland, J., Altman, D. (**1997**) Cronbach's Alpha. *British Medical Journal*, **314**, 572.

Bowling, A., Bond, M., Jenkinson, C., et al. (**1999**) Short Form 36 (SF-36) health survey questionnaire: which normative data should be used? Comparisons between the norms provided by the Omnibus Survey in Britain, the Health Survey for England and the Oxford Healthy Life Survey. *Journal of Public Health Medicine*, **21**, 255–70.

Box, G., Cox, D. (**1964**) An analysis of transformations. *Journal of the Royal Statistical Society Series B*, **26**, 211–246.

Brayne, C., Gill, C., Paykel, E., et al. (**1995**) Cognitive decline in an elderly population- a two wave study of change. *Psychological Medicine*, **25**, 673–683.

Brazier, J. E., Harper, R., Jones, N., et al. (**1992**) Validating the SF-36 health survey questionnaire: new outcome measure for primary care. *British Medical Journal*, **305**, 160–165.

Brooks, S., Gelman, A. (**1998**) Alternative methods for monitoring convergence of iterative simulations. *Journal of Computational and Graphical Statistics*, **7**, 434–455.

CACI (**1999**) Paycheck: Targeting by income. **http://www.caci.co.uk/pdfs/caci_paycheck.pdf**.

Chandola, T., Bartley, M., Wiggins, R., et al. (**2003**) Social inequalities in health by individual and household measures of social position in a cohort of healthy people. *Journal of Epidemiology and Community Health*, **57**, 56–62.

Chandola, T., Clarke, P., Wiggins, R., et al. (**2005**) Who you live with and where you live: setting the context for health using multiple membership multilevel models. *Journal of Epidemiology and Community Health*, **59**, 170–175.

Cockings, S., Martin, D. (**2005**) Zone design for environment and health studies using pre-aggregated data. *Social Science and Medicine*, **60**, 2729–2742.

Congdon, P. (**2001**) *Bayesian Statistical Modelling*. Probability and Statistics, Wiley.

Cortina, J. (**1993**) What is coefficient alpha? An examination of theory and applications. *Journal of Applied Psychology*, **78**, 98–104.

Croft-Jefferys, C., Wilkinson, G. (**1989**) Estimated costs of neurotic disorder in uk general practice. *Psychological Medicine*, **19**, 549–58.

Cummins, S., Macintyre, S., Davidson, S., et al. (**2005**) Measuring neighbourhood social and material context: generation and interpretation of ecological data from routine and non-routine sources. *Health and Place*, **11**, 249–260.

De Silva, M., McKenzie, K., Harpham, T., et al. (2005) Social capital and mental illness: a systematic review. *Journal of Epidemiology and Community Health*, **59**, 619–627.

De Vellis, R. (1991) *Scale Development: Theory and Application.* Sage.

Department of Public Health Medicine (1993) *Annual Report on Health in Herefordshire.* Department of Public Health Medicine.

Diez Roux, A. (2001) Investigating neighbourhood and area effects on health. *American Journal of Public Health*, **91**, 1783–1789.

Diez-Roux, A. (2003) The examination of neighborhood effects on health: conceptual and methodological issues related to the presence of multiple levels of organization. In **Kawachi, I., Berkman, L.**, eds., *Neighborhoods and Health*, pp. 56–57, Oxford University Press.

Drukker, M., van Os, J. (2003) Mediators of neighbourhood socioeconomic deprivation and quality of life. *Social Psychiatry and Psychiatric Epidemiology*, **38**, 698–706.

DSM IV (2000) *Diagnostic and Statistical Manual of Mental Disorders: Fourth Edition.* American Psychiatric Association.

Dukes, K. (1999) Cronbach's Alpha. In **Armitage, P., Colton, T.**, eds., *Encyclopedia of Biostatistics*, volume 2, Chichester: Wiley.

Duncan, G., Raudenbush, S. (1999) Assessing the effects of context in studies of child and youth development. *Educational Psychologist*, **34**, 29–41.

Efron, B. (1986) Why isn't everyone a Bayesian? *The American Statistician*, **40**, 1–5.

Ellen, I., Mijanovich, T., Dillman, K. (2001) Neighborhood effects on health: Exploring the links and assessing the evidence. *Journal of Urban Affairs*, **23**, 391–408.

Emslie, C., Fuhrer, R., Hunt, K., et al. (2002) Gender differences in mental health: evidence from three organisations. *Social Science and Medicine*, **54**, 621–624.

Etter, J., Perneger, T. (1997) Analysis of non-response bias in a mailed health survey. *Journal of Clinical Epidemiology*, **50**, 1123–1128.

Fears, T., Benichou, J., Gail, M. (1996) A reminder of the fallibility of the wald statistic. *The American Statistician*, **50**, 226–227.

Fone, D. (2005) *People, places and mental health in Caerphilly County Borough: a multilevel modelling analysis.* Ph.D. thesis, University of Wales College of Medicine.

Fone, D., Christie, S., Lester, N. (2006a) Comparison of perceived and modelled geographical access to accident and emergency departments: a cross-sectional analysis from the Caerphilly Health and Social Needs Study. *International Journal of Health Geographics*, **5**, 1–16.

Fone, D., Dunstan, F. (2006) Mental health, place and people: A multilevel analysis of economic inactivity and social deprivation. *Health and Place*, **12**, 332–344.

Fone, D., Dunstan, F., Christie, S., et al. (2006b) Council tax valuation bands, socio-economic status and health outcome: A cross-sectional analysis from the Caerphilly Health and Social Needs Study. *BMC Public Health*, **6**.

Fone, D., Dunstan, F., John, A., et al. (2007a) Associations between common mental disorders and the Mental Illness Needs Index in community settings. *British Journal of Psychiatry*, **191**, 158–163.

Fone, D., Dunstan, F., Lloyd, K., et al. (2007b) Does social cohesion modify the association between area income deprivation and mental health? A multilevel analysis. *International Journal of Epidemiology*, **36**, 338–345.

Fone, D., Dunstan, F., Williams, G., et al. (2007c) Places, people and mental health: A multilevel analysis of economic inactivity. *Social Science and Medicine*, **64**, 633–645.

Fone, D., Farewell, D., Dunstan, F. (2006c) An ecometric analysis of neighbourhood cohesion. *Population Health Metrics*, **4**, 1–17.

Fone, D., Jones, A., Watkins, J., et al. (2002) Using local authority data for action on health inequalities: the Caerphilly Health and Social Needs Study. *British Journal of General Practice*, **52**, 799–804.

Friedman, B., Heisel, M., Delavan, R. (2005) Validity of the SF-36 five-item mental health index for major depression in functionally impaired, community-dwelling elderly patients. *Journal of the American Geriatrics Society*, **53**, 1979–1985.

Galster, G. (2001) On the nature of neighbourhood. *Urban Studies*, **38**, 2111–2124.

Gamerman, D. (1997) *Markov Chain Monte Carlo: stochastic simulation for Bayesian inference.* Texts in statistical science, Chapman and Hall.

Garratt, A. M., Ruta, D. A., Abdalla, M. I., et al. (1993) The SF-36 health survey questionnaire - an outcome measure suitable for routine use within the NHS. *British Medical Journal*, **306**, 1440–1444.

Gilbody, S., House, A., Sheldon, T. (2001) Routinely administered questionnaires for depression and anxiety: systematic review. *British Medical Journal*, **322**, 406–409.

Gilbody, S., House, A., Sheldon, T. (2006) Screening and case finding instruments for depression. *The Cochrane Database of Systematic Reviews.*

Glennerster, H., Lupton, R., Noden, P., et al. (1999) Poverty, Social Exclusion and Neighbourhood: Studying the area bases of social exclusion. CASE paper 22.

Glover, G., Robin, E., Emami, J., et al. (1998) A needs index for mental health care. *Social Psychiatry and Psychiatric Epidemiology*, **33**, 89–96.

Goldberg, D., Gater, G., Sartorius, N., et al. (1997) The validity of two versions of the GHQ in the WHO study of mental illness in general health care. *Psychological Medicine*, **27**, 191–197.

Goldberg, D., Huxley, P. (1992) *Common Mental Disorders: A Biosocial Model.* London: Routledge.

Goldberg, D., Williams, P. (1988) *A User's Guide to the General Health Questionnaire.* Windsor: NEFR-Nelson.

Goldstein, H. (2003) *Multilevel Statistical Models.* Hodder Arnold, ISBN 0340806559.

Goldstein, H., Browne, W., Rasbash, J. (2002) Partitioning variation in multilevel models. *Understanding Statistics*, **1**, 223–231.

Hand, D. (1987) Screening vs Prevalence Estimation. *Applied Statistics*, **36**, 1–7.

Hayes, V., Morris, J., Wolfe, C., et al. (1995) The SF-36 health survey questionnaire - Is it suitable for use with older adults? *Age and Ageing*, **24**, 120–125.

Hill, S., Harries, U., Popay, J. (1996) Is the short form 36 (SF-36) suitable for routine health outcomes assessment in health care for older people? evidence from preliminary work in community based health services in England. *Journal of Epidemiology and Community Health*, **50**, 94–98.

Hoeymans, N., Garssen, A., Westert, G., et al. (2004) Measuring mental health of the Dutch population: a comparison of the GHQ-12 and the MHI-5. *Health and Quality of Life Outcomes*, **2**, 23–29.

Holmes, W. (**1998**) A short, psychiatric, case-finding measure for HIV seropositive outpatients. *Medical Care*, **36**, 237–243.

Jenkins, R. (**1985**) Minor psychiatric disorder in employed young men and women and its contribution to sickness absence. *British Journal of Psychiatric Medicine*, **42**, 147–54.

Jenkinson, C., Coulter, A., Wright, L. (**1993**) Short form 36 (SF36) health survey questionnaire:normative data for adults of working age. *British Medical Journal*, **306**, 1437–40.

Jenkinson, C., Layte, R., Lawrence, K. (**1997**) Development and testing of the medical outcomes study 36-item short from health survey summary scale scores in the United Kingdom: Results from a large-scale survey and clinical trial. *Medical Care*, **35**, 410–416.

Jenkinson, C., Stewart-Brown, S., Petersen, S., et al. (**1999**) Assessment of the SF-36 version 2 in the United Kingdom. *Journal of Epidemiology and Community Health*, **53**, 46–50.

Joshi, H., Wiggins, R., Bartley, B., et al. (**2001**) Putting health inequalities on the map: does where you live matter, and why? In **Graham, H.**, ed., *Understanding health inequalities*, Buckingham: Open University Press.

Kawachi, I., Berkman, L. (**2003**) *Neighborhoods and Health.* New York: Oxford University Press.

Keller, S. D., Ware, J. E., Bentler, P. M., et al. (**1998**) Use of structural equation modeling to test the construct validity of the SF-36 Health Survey in ten countries: Results from the IQOLA Project. *Journal of Clinical Epidemiology*, **51**, 1179–1188.

Kingdon, A., Roberts, C., Tudor-Smith, C. (**1998**) Lifestyle changes in Wales: results from the Health in Wales Survey 1985-1996. Technical Report 27, Health Promotion Wales.

Lawson, A., Biggeri, A., Böhning, D., et al., eds. (**1999**) *Disease mapping and risk assessment for public health.* John Wiley and Sons.

Li, L., Wang, H. M., Shen, Y. (**2003**) Chinese SF-36 Health Survey: translation, cultural adaptation, validation, and normalisation. *Journal of Epidemiology and Community Health*, **57**, 259–263.

Lieberson, S. (**1969**) Measuring population diversity. *American Sociological Review*, **34**, 850–862.

Litzkow, M., Livny, M., Mutka, M. (**1988**) Condor - A Hunter of Idle Workstations. In *Proceedings of 8th IEEE International Conference on Distributed Computing Systems 1988 (ICDCS8)*, pp. 104–111, San Jose, California, USA: IEEE.

Lumley, T., Diehr, P., Emerson, S., et al. (**2002**) The importance of the normality assumption in large public health datasets. *Annual Review of Public Health*, **23**, 151–169.

Lunn, D., Thomas, A., Best, N., et al. (**2000**) WinBUGS – a Bayesian modelling framework: concepts, structure, and extensibility. *Statistics and Computing*, **10**, 325–337.

Macintyre, S., Ellaway, A., Cummins, S. (**2002**) Place effects on health: how can we conceptualise, operationalise and measure them? *Social Science and Medicine*, **55**, 125–139.

Macintyre, S., Maciver, S., Sooman, A. (**1993**) Area, class and health: Should we be focusing on places or people? *Journal of Social Policy*, **22**, 213–234.

Marmot, M. (**2004**) *Status Syndrome: How your social standing affects your health and life expectancy.* London: Bloomsbury Publications.

McCabe, C. J., Thomas, K. J., Brazier, J. E., et al. (**1996**) Measuring the mental health status of a population: A comparison of the GHQ-12 and the SF-36 (MHI-5). *British Journal of Psychiatry*, **169**, 517–521.

McCallum, J. (**1995**) The SF-36 in an Australian sample - Validating a new, generic health-status measure. *Australian Journal of Public Health*, **19**, 160–166.

McCulloch, A. (**2001**) Ward-level deprivation and individual social and economic outcomes in the British Household Panel Study. *Environment and Planning A*, **33**, 667–684.

McHorney, C. A., Ware, J. E., Lu, J. F. R., et al. (**1994**) The MOS 36-item short-form health survey (SF-36) .3. Tests of data quality, scaling assumptions, and reliability across diverse patient groups. *Medical Care*, **32**, 40–66.

McHorney, C. A., Ware, J. E., Raczek, A. E. (**1993**) The MOS 36-item short-form health survey (SF-36) .2. Psychometric and clinical-tests of validity in measuring physical and mental-health constructs. *Medical Care*, **31**, 247–263.

McLoone, P. (**2001**) Targeting deprived areas within small areas in Scotland: population study. *British Medical Journal*, **323**, 374–5.

Merlo, J. (**2003**) Multilevel analytical approaches in social epidemiology: measures of health variation compared with traditional measures of association. *Journal of Epidemiology and Community Health*, **57**, 550–552.

Moerbeek, M. (**2004**) The consequence of ignoring a level of nesting in multilevel anaysis. *Multivariate Behavioural Research*, **39**, 129–149.

Morris, R., Carstairs, V. (**1991**) Which deprivation? A comparison of selected deprivation indices. *Journal of Public Health and Medicine*, **13**, 318–26.

Murray, C., Lopez, A. (**1996**) Global and regional descriptive epidemiology of disability: Incidence, prevalence, health expectancies and years lived with disability. In **Murray, C., Lopez, A.**, eds., *The Global Burden of Disease*, Cambridge, Mass: World Health Organisation.

National Assembly for Wales (**2000**) *Welsh Index of Multiple Deprivation.* Cardiff: Government Statistical Service.

National Assembly for Wales (**2001**) *Communities First Guidance.* Cardiff: National Assembly for Wales.

Nerenz, D., Repasky, D., Whitehouse, F., et al. (**1992**) Ongoing assessment of health status in patients with diabetes mellitus. *Medical Care*, **30**, MS112–MS123.

O'Campo, P. (**2003**) Advancing theory and methods for multilevel models of residential neighbourhoods and health. *American Journal of Epidemiology*, **157**, 9–13.

Office for National Statistics (**1999**) Great britain '91 census geography. http://www.census.ac.uk/cdu/Datasets/1991_Census_datasets/Area_Stats/General_Topics/Geography/GB_91_Census_geography/.

Office for National Statistics (**2003**) Census 2001- Health, disability and provision of care. http://www.statistics.gov.uk/census2001/profiles/commentaries/health.asp.

Office for National Statistics (**2004**) Percentage of people whose income is below various fractions of median income: Social trends 34. http://www.statistics.gov.uk/statbase/ssdataset.asp?vlnk=7446.

Office for National Statistics (**2006a**) Confidentiality protection in the census and household surveys. http://www.statistics.gov.uk/about/data/guides/LabourMarket/methods/Confidentiality/\\census_household_surveys.asp.

Office for National Statistics (**2006**b) Super Output Areas (SOAs). http://www.statistics.gov.uk/geography/soa.asp.

ONS (**1991**) *Standard Occupational Classification.* Office for Population Censuses and Surveys, London.

Openshaw, S. (**1984**) *The Modifiable Areal Unit Problem.* GEO Books.

Parker, S., Peet, S., Jagger, C., et al. (**1998**) Measuring health status in older patients. The SF-36 in practice. *Age and Ageing*, **27**, 13–18.

Pawitan, Y. (**2001**) *In all likelihood:statistical modelling and inference using likelihood.* Oxford University Press, ISBN 0198507658.

Perenboom, R., Oudshoorn, K., van Herten, L., et al. (**2000**) *Life-expectancy in good mental health: establishing cut-offs for the MHI-5 and GHQ-12 (in Dutch).* Leiden: TNO-report.

Pickett, K., Pearl, M. (**2000**) Multilevel analyses of neighbourhood socioeconomic context and health outcomes: a critical review. *Journal of Epidemiology and Community Health*, **55**, 111–122.

Propper, C., Jones, K., Bolster, A., et al. (**2005**) Local neighbourhood and mental health: Evidence from the UK. *Social Science and Medicine*, **61**, 2065–2083.

R Development Core Team (**2006**) *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing, Vienna, Austria, ISBN 3-900051-07-0.

Rasbash, J., Browne, W., Healy, M. (**2003**) Mlwin beta version 2.0.

Reijneveld, A., Schene, A. (**1998**) Higher prevalence of mental disorders in socioeconomically deprived urban areas in the Netherlands: community or personal disadvantage. *Journal of Epidemiology and Community Health*, **52**, 2–7.

Reijneveld, S., Verhiej, R., de Bakker, D. (**2000**) The impact of area deprivation on differences in health: does the choice of the geographical classification matter? *Journal of Health and Social Behaviour*, **41**, 177–187.

Rice, N., Carr-Hill, R., Dixon, P., et al. (**1998**) The influence of households on drinking behaviour: a multilevel analysis. *Social Science and Medicine*, **46**, 971–979.

Roberts, R., Hemingway, H., Marmot, M. (**1997**) Psychometric and clinical validity of the SF-36 general health survey in the Whitehall ii study. *British Journal of Health Psychology*, **2**, 285–300.

Robinson, W. (1950) Ecological correlations and the behaviour of individuals. *American Sociological Review*, **15**, 351–357.

Ross, C. (2000) Neighborhood disadvantage and adult depression. *Journal of Health and Social Behaviour*, **41**, 177–187.

Rumpf, H., Meyer, C., Hapke, U., et al. (2001) Screening for mental health: validity of the MHI-5 using DSM-IV Axis 1 psychiatric disorders as gold standard. *Psychiatry Research*, **105**, 243–253.

Sampson, R., Morenoff, J., Gannon-Rowley, T. (2002) Assessing "Neighborhood Effects": Social processes and new directions in research. *Annual review of sociology*, **28**, 443–78.

Sawa, T. (1978) Information criteria for discriminating among alternative regression models. *Econometrica*, **46**, 1273–1291.

Schmitt, N. (1996) Uses and abuses of coefficient alpha. *Psychological Assessment*, **8**, 350–353.

Skapinakis, P., Lewis, G., Araya, R., et al. (2005) Mental health inequalities in Wales, UK: multi-level investigation of the effect of area deprivation. *British Journal of Psychiatry*, **186**, 417–422.

Snijders, T., Bosker, R. (1999) *Multilevel Analysis: An introduction to basic and advanced multilevel modelling.* Sage.

Snow, J. (1849) On the pathology and mode of communication of Cholera. *London Medical Gazette*, **9**, 745–53, 923–49.

Strand, B. H., Dalgard, O. S., Tambs, K., et al. (2003) Measuring the mental health status of the norwegian population: a comparison of the instruments SCL-25, SCL-10, SCL-5 and MHI-5 (SF-36). *Nordic Journal of Psychiatry*, **57**, 113–118.

Streiner, D., Norman, G. R. (2003) *Health measurement scales: A practical guide to their development and use.* Oxford University Press.

Subramanian, S., Jones, K., Duncan, C. (2003) Multilevel models for public health research. In *Neighborhoods and Health*, New York: Oxford University Press.

Taylor, M., Brice, J., Buck, N., et al. (2005) *British Household Panel Survey User Manual Volume A: Introduction, technical report and appendices.* University of Essex.

The National Assembly for Wales (1999) The National Assembly for Wales. Welsh Health Survey 1998. Results of the second Welsh Health Survey.

Thumboo, J., Fong, K. Y., Machin, D., et al. (2001) A community-based study of scaling assumptions and construct validity of the English (UK) and Chinese (HK) SF-36 in Singapore. *Quality of Life Research*, **10**, 175–188.

Townsend, P., Phillimore, P., Beattie, A. (1988) *Health and deprivation: inequality and the North.* London: Routledge.

University of Southampton (2000) Output Area Construction. http://www.geog.soton.ac.uk/research/oa2001/oaconst.htm

Üstün, T., Ayuso-Mateos, J., Chatterji, S., et al. (2004) Global burden of depressive disorders in the year 2000. *British Journal of Psychiatry*, **184**, 386–392.

Vickers, D., Rees, P. (2007) Creating the UK National Statistics 2001 output area classification. *Journal of the Royal Statistical Society A*, **170**, 379–403.

Wainwright, N., Surtees, P. (2003) Places, people, and their physical and mental functional health. *Journal of Epidemiology and Community Health*, **58**, 333–339.

Wallace, M., Denham, C. (1996) ONS classification of local and health authorities of Britain. Technical report, HMSO.

Ware, E., Gandek, B. (1998) Overview of the SF-36 survey and the international quality of life assessment(IQOLA). *Journal of Clinical Epidemiology*, **51**, 903–912.

Ware, J., Kosinksi, M., Dewey, J. (2000a) *How to score version 2 of the SF-36®Health Survey.* RI:Quality Metric Incorporated.

Ware, J., Kosinksi, M., Gandek, B. (2000b) *SF-36®Health Survey: Manual and Interpretation Guide.* RI:Quality Metric Incorporated.

Ware, J. E., Kosinski, M., Bayliss, M. S., et al. (1995) Comparison of methods for the scoring and statistical-analysis of SF-36 health profile and summary measures - summary of results from the Medical Outcomes Study. *Medical Care*, **33**, AS264–AS279.

Weich, S., Holt, G., Twigg, L., et al. (2003a) Geographic variation in the prevalence of common mental disorders in britain: A multilevel investigation. *American Journal of Epidemiology*, **157**, 730–737.

Weich, S., Lewis, G., Jenkins, S. (2001) Income inequality and the prevalence of common mental disorders in Britain. *British Journal of Psychiatry*, **178**, 222–227.

Weich, S., Slogett, A., Lewis, G. (1998) Social roles and gender difference in the prevalence of common mental disorders. *British Journal of Psychiatry*, **173**, 489–93.

Weich, S., Twigg, L., Holt, G., et al. (2003b) Contextual risk factors for the common mental disorders in britain: a multilevel investigation of the effects of place. *Journal of Epidemiology and Community Health*, **57**, 616–621.

Weich, S., Twigg, L., Lewis, G. (2006) Rural/non-rural differences in rates of common mental disorders in britain. *British Journal of Psychiatry*, **188**, 51–57.

Weich, S., Twigg, L., Lewis, G., et al. (2005) Geographic variation in rates of common mental disorders in britain:prospective cohort study. *British Journal of Psychiatry*, **187**, 29–34.

Weinstein, W., Berwick, D., Goldman, P., et al. (1989) A comparison of three psychiatric screening tests using reciever operating characteristics (ROC) analysis. *Medical Care*, **27**, 593–607.

Williams, P., Tarnopolsky, A., Hand, D., et al. (1986) Minor psychiatric morbidity and general practice consultations: the West London Survey. *Psychological Medicine*, **9**, 1–37.

Winston, M., Smith, J. (2000) A trans-cultural comparison of four psychiatric case-finding instruments in a Welsh community. *Social Psychiatry and Psychiatric Epidemiology*, **35**, 569–575.

Womble, W. (1951) Differential systematics. *Science*, **114**, 315–322.

World Health Organisation (1992) *The ICD-10 classification of mnetal and behavioural disorders: Clinical descriptions and diagnostic guidelines.* World Health Organisation.

Youden, W. (1950) An index for rating diagnostic tests. *Cancer*, **3**, 32–35.