UNIVERSITY OF SOUTHAMPTON

Design of a Census Coverage Survey and its Use in the Estimation and Adjustment of Census Underenumeration

A Contribution Towards Creating a One-Number Census in the UK in 2001

By James John Brown

Doctor of Philosophy

Department of Social Statistics
Faculty of Social Science

December 2000

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF SOCIAL SCIENCE

SOCIAL STATISTICS

Doctor of Philosophy

DESIGN OF A CENSUS COVERAGE SURVEY AND ITS USE IN THE ESTIMATION AND ADJUSTMENT OF CENSUS UNDERENUMERATION

By James John Brown

Many countries in the world conduct censuses of their populations. The UK is no exception, and every ten years there is a census undertaken in England and Wales, Scotland, and Northern Ireland. Although in theory it is three separate censuses, in practice these are planned and conducted as a joint project. The 1991 Censuses suffered from an increased level of underenumeration relative to the 1981 Censuses. This underenumeration was not detected by the 1991 follow-up survey that was designed to estimate this underenumeration.

The work presented in this thesis develops the design of a new follow-up survey for the 2001 Censuses that will be able to measure the level of census underenumeration. Much of the work presented in this thesis deals with the development of an effective estimation strategy for this follow-up survey that utilises all the available data. In addition the thesis contains an analysis that adjusts this design and estimation strategy to facilitate its implementation within Northern Ireland. The ultimate goal for the 2001 Censuses will be to create a 'One-Number Census'. This is a census database where the estimated underenumeration has been fully integrated into the output database so that all tabulations are consistent with the agreed national population adjusted for underenumeration. This thesis considers the development of an imputation system for this purpose. The system utilises a donor imputation approach but also makes use of the estimates already available to ensure that the resulting database is consistent with agreed population estimates.

Contents

Chapter One – Introduction

1.1) The Research Problem 1

1.2) Historical Context 2

1.3) Objectives of the Research 4

1.4) Organisation of the Thesis 5

Chapter Two – Review of the Literature

2.1) Introduction 6

2.2) History of the Estimation of Census Errors in the US 6

2.2.1) The 1950 Census 6

2.2.2) The 1960 Census 8

2.2.3) The 1970 Census 11

2.2.4) The 1980 Census 12

2.2.5) The 1990 Census 17

2.2.6) Plans for the 2000 Census 19

2.2.7) Underenumeration Measurement in US Censuses – Conclusions 20

2.3) Estimation of Underenumeration in the 1991 Censuses of the UK 21

2.4) Methods for Estimating Census Underenumeration 26

2.4.1) Dual-System Estimation and Capture-Recapture Methods 27

2.4.1.1) Properties of the Dual-System Estimator 30

2.4.1.2) Relaxing the Assumptions in Capture-Recapture Analysis 32

2.4.2) Other Survey-Based Methods 36

2.5) Conclusion 38

Appendix 2.1 – Re-expressing the Chapman estimator 40

Chapter Three – Census Coverage Survey Design

3.1) Introduction 41

3.2) The Basic Approach 44

3.3) One-Stage Design 45

3.3.1) Postcode Level Model 45

3.3.2) ED Level Model 47

3.4) Prototype Designs 48

3.4.1) National Hard to Count Index 49

3.4.2) Postcode Level Design 50

3.4.3) ED Level Design 51

3.5) Two-Stage Design 53

3.5.1) Comparisons of Different Postcode Selections 55

3.5.2) Multivariate Stratification 58

3.6) Conclusions 61

Appendix 3.1 – The variance of the estimated population total assuming a clustered super-population model 64

Chapter Four – Estimation of Census Underenumeration for Estimation Areas

4.1) Introduction 66

4.2) Population Estimation Using the 2001 Census Coverage Survey 68

4.2.1) Estimation Within the CCS Postcodes 68

4.2.2) Models for Population Estimation Using the CCS 70

4.2.2.1) Simple Approach 71

4.2.2.2) Ratio Model for Population Estimation 72

4.2.2.3) The Impact of Dual-System Estimation on the Ratio Model 74

4.2.2.4) Regression Model for Population Estimation 76

4.3) Simulation Study 77

4.3.1) Applying the CCS Design to the Simulation Population 78

4.3.2) Simulating a Census and its CCS 80

4.3.3) Population Estimation Results 82

4.3.4) Conclusions on the Simulation Study Results 88

4.4) A 'Robust' Estimation Strategy 89

4.4.1) A Model for Robust Estimation 90

4.4.2) Statistical Motivation for the Prediction in Non-Sampled Areas 93

4.4.3) Applying the Strategy 94

4.4.4) Results from the Robust Estimation Strategy 96

4.4.5) The Impact of Dependence 100

4.4.6) Variance Estimation 104

4.4.7) Implementing the Robust Estimation Strategy in 2001 107

4.5) Additional Considerations 108

4.5.1) Movers and the CCS 109

4.5.2) Overenumeration in the 2001 Census 111

4.5.3) Estimation for Individual LADs 112

4.6) Concluding Remarks 113

Appendix 4.1 – Creating census and CCS coverage probabilities for the simulation population 115

Appendix 4.2 – Census coverage by age-sex group for the 100 simulated censuses 117

Appendix 4.3 – Ratio estimators combined with weighted and unweighted dual-system estimation for females by age 118

Chapter Five – CCS Design and Estimation in Northern Ireland

5.1) Introduction 119

5.2) The CCS Design Applied in Northern Ireland 120

5.2.1) Classification Index for EDs 120

5.2.2) CCS Design for Northern Ireland 121

5.3) Northern Ireland Simulation Study 124

5.3.1) CCS Design for the Simulation Population 124

5.3.2) Running the Simulations 124

5.3.3) Results 125

5.3.3.1) Robust Strategy 133

5.4) Estimation for Other Variables 137

5.4.1) Household Size 139

5.4.2) Simulation Results 142

5.5) Other Issues Specific to Northern Ireland 148

5.5.1) Estimation of Population Counts for Local Government Districts (LGDs) 148

5.5.2) Selection of the Postcode Sample 149

5.6) Conclusions 150

Appendix 5.1 – The distribution of census errors for Northern Ireland by age, sex, and three levels of census coverage 152

Chapter Six – Adjusting for Census Coverage at the Household and Individual Level

6.1) Introduction 153

6.2) Development of the Framework 154

6.3) Controlled Imputation Methodology 156

6.3.1) Estimation of Household and Individual Coverage Weights 157

6.3.1.1) Derivation of Household Coverage Weights 157

6.3.1.2) Derivation of Individual Coverage Weights 158

6.3.2) Imputation of Households 160

6.3.3) Imputation of Individuals into Counted Households 160

6.3.4) Final Calibration ('pruning and grafting') 162

6.4) Simulation Study 162

6.4.1) The Household Coverage Model 163

6.4.2) The Individual Coverage Model 164

6.4.3) Simulation Results 165

6.5) Discussion 167

Chapter Seven – Conclusions

7.1) Introduction 169

7.2) The Census Coverage Survey Design 169

7.2.1) The Census Coverage Survey Design for Northern Ireland 171

7.3) The Estimation Strategy 172

7.3.1) Estimation of the Population by Other Characteristics 174

7.3.2) Estimation of Occupied Housing Units 174

7.4) Production of a One-Number Census 175

7.5) Concluding Remarks 175

References 177

List of Tables

2.1	Classification of enumeration status	27
2.2	Indicative example of household capture probabilities by household size	35
3.1	One-stage postcode level design using 1991 total population for size stratification	51
3.2	One-stage ED level design using 1991 total population for size stratification	52
3.3	Two-stage design using ED total population in 1991 and number of postcodes per ED for size stratification	57
3.4	Two-stage design using a multivariate approach for size stratification	60
4.1	Distribution of enumeration districts by HtC index for the total population of EDs and the ED sample	80
4.2	Performance of the population estimators based on weighted DSEs	84
4.3	Performance of the population estimators based on unweighted DSEs	84
4.4	Performance of the DSE at two levels for estimating the sample population	86
4.5	Performance of the 'robust' ratio estimators for the population total compared with simpler approaches	96
4.6	The impact of dependence in the simulation on the estimate of total population	102
4.7	Performance of variance estimators for the total population	105
5.1	Ranking of the ED classification index	121
5.2	Specification of the CCS for the Belfast estimation area	123
5.3	Results of the simulation study for estimates of the total population	126
5.4	Results of the simulation study for estimates of the total population by estimation area	127
5.5	Comparison of simulation results for the standard and robust estimators for the total population of Northern Ireland by estimation area	133
5.6	Comparison between the census and the robust estimation strategy for population shares by sex and estimation area	137
5.7	Proportions within the sample areas	140

5.8	Comparison between the census and the robust estimation strategy for population shares by religion and estimation area	144
5.9	Results of the simulation study for estimates of the total number of occupied households by estimation area	144
6.1	Relative average bias and relative root average mean square error across EDs for ten simulations: number of households by HtC index	166
6.2	Relative average bias and relative root average mean square error across EDs for ten simulations: number of individuals by HtC index	167

List of Figures

3.1	Comparison of the performance of different sized within ED samples for a fixed total postcode sample and a fixed ED stratification at both the total population and across the age-sex groups.	57
3.2	Comparison of the performance across the age-sex groups of ED stratification by total population and number of postcodes (*red*) with multivariate ED stratification (*blue*) for a within ED sample of five postcodes.	61
4.1	Ratio estimators combined with weighted and unweighted dual system estimation for males by age	87
4.2	Comparison of the robust ratio model using a postcode DSE constrained to a cluster DSE with the standard ratio model using cluster DSE and the US weighted DSE for males by age	97
4.3	Distributions of the errors for the standard and robust strategies	99
4.4	The impact of dependence in terms of relative bias on estimates of the total population by age and sex based on the robust strategy	103
4.5	Performance of the variance estimators for the individual age-sex estimates	106
5.1	Distribution of the errors for the estimator of the total population of Northern Ireland by age and sex for three levels of census coverage	128
5.2	Relative bias of the census data and the data adjusted using the standard estimator by age and sex for the three estimation areas	130

5.3	Relative RMSE of the census data and the data adjusted using the standard estimator by age and sex for the three estimation areas	131
5.4	Relative bias for counts adjusted using the standard and robust estimators by age and sex for the three estimation areas	135
5.5	Relative RMSE of the census counts compared to counts adjusted using the standard and robust estimators by age and sex for the three estimation areas	136
5.6	Performance of the census compared with the standard estimator by religion for the three estimation areas	143
5.7	Performance of the census compared with the standard estimator by household tenure for the three estimation areas	146
5.8	Performance of the census compared with the standard estimator by household size for the three estimation areas	147

Acknowledgements

The Economic and Social Research Council financially supported the work in this thesis through studentship R00429634268.

I would like to thank my supervisors, Ray Chambers and Ian Diamond, for their guidance and encouragement throughout my PhD studies and for their initial support without which I would never have undertaken a PhD. Ian was partly responsible for awakening my interest in Statistical Demography when I was a Mathematics undergraduate back in 1993 and without Ray's enthusiasm I would never have become interested in sampling.

I would like to thank all the One-Number Census team at ONS for their support and the provision of data without which many of the results presented in this thesis could not have been produced. I would particularly like to thank Owen Abbott, Lisa Buckner, and Marie Cruddas who have all given time to support and encourage my work, and Norma who has made sure that I have always been well looked after when visiting ONS.

A similar thank you to the team in Northern Ireland, particularly Maire Rodgers and Robert Beatty, for providing data and insights into Northern Ireland that were invaluable in designing the CCS for Northern Ireland.

I would also like to thank the other PhD students and members of the Department of Social Statistics who have made studying at Southampton a challenging and enjoyable experience, particularly colleagues from 141 University Road and the departmental secretaries.

A final thank you is due to my family, particularly my wife Melanie who has endured so much in the last few months while I have been completing this thesis and my parents who have always supported my studies over the many years prior to my PhD studies.

Chapter 1 – Introduction

1.1) The Research Problem

The Encyclopaedia Britannica defines a census in general terms as 'an enumeration of people, houses, firms, or other important items in a country or region at a particular time'. It extends this for the modern population census to 'a complete enumeration of all the people and their important characteristics for purposes of understanding the basic structure and trends of the society'. Every ten years such a population census is undertaken in each of the countries that constitute the United Kingdom (UK). In legal terms there are three independent censuses but in reality the planning and organisation are interlinked to such an extent that to the public it appears as one census for the whole UK. The main difference is a few country specific questions. The aim of each Census is to inform national and local government about the basic social and demographic characteristics of the nation by providing descriptive information at a very small geographic level of aggregation. It is for this reason that a census is necessary.

The theory of a census is straightforward, carrying it out is not. Steinberg *et al* (1962) in their paper discussing the accuracy of the 1960 United States (US) Census list population mobility, difficulty locating housing units, difficulty finding people at home, people with multiple homes, and inexperienced enumerators as just some of the problems faced by a census. The Office for National Statistics (ONS) in England and Wales is planning for many of these practical problems to be worse in 2001 than ever before (Jones, 1997). In other words, even when the census is compulsory, achieving a complete and accurate count with restricted resources is impossible, and so individuals and households get missed and the census 'underenumerates' the population by some amount. Overenumeration can also occur, with people being included more than once, or erroneously enumerate with people included once but in the wrong place. In general these problems are small relative to underenumeration and the net result is a census that does not count enough people.

The main problem with underenumeration in a census is typically not the total number of people missed. For example, the UK censuses still count about ninety eight per cent of the population. The real problem is the variation in underenumeration across different demographic groups and geographic areas. Therefore, underenumeration becomes a problem when comparing small areas or domains as one domain or area may have suffered a greater underenumeration than another.

1.2) Historical Context

The work in this thesis has a definite historical context. In particular, to understand the motivation underpinning this thesis it is necessary to consider briefly the 1991 Censuses of the UK. These Censuses encountered unforeseen problems, which affected their perceived success by the users of census data, with the level of net underenumeration in Great Britain rising from 0.45 per cent in 1981 to over two per cent in 1991. However, that fact alone is not sufficient to class the 1991 Censuses as a failure, since similar levels of net underenumeration had been observed in Canada, the US, Australia, and other comparable countries. So why did they produce the headline *1.8 million Britons 'disappear'*[1]?

There are probably at least three reasons for this. Firstly, it was a big increase in underenumeration from 1981. Secondly, the political climate at the time had been influenced by the poll tax, leading to further newspaper articles including *'Missing million indicates poll tax factor in census*[2]. While the Office for Population Censuses and Surveys (OPCS) was quoted in the article as denying any evidence for this, they did point to the third reason. The follow-up survey in 1991 had not found the missing people. The article reports that OPCS estimated that over 0.5 million people had been missed by both the Census and the follow-up survey. Ultimately it was the failure of the follow-up survey that caused the perception that the Censuses had 'failed' as OPCS had no estimate of the national population based on the 1991 Census adjusted for net underenumeration. OPCS also had no real evidence from the 1991 follow-up survey to place the missed people in local authority districts around the country. This will be discussed in greater detail in the following chapter.

[1] Article by Rosie Waterhouse in The Independent on Sunday, 13th September 1992, p1.

There is also an international context to this thesis. During the same inter-censal period of 1980 to 1990 the US Census Bureau had come under increasing pressure to 'adjust' its census count for estimated net underenumeration. The US Census Bureau already had a long history of using a follow-up survey with other demographic techniques to assess the coverage of the census. The first such survey had followed the 1950 Census. However, the role had always been to inform users of the census about the accuracy of the census count and not to actually correct the count by adjusting the census counts to reflect the estimated net underenumeration.

Following the 1980 US Census, the State of New York and New York City took the US Census Bureau to court to force their census counts to be adjusted. The plaintiffs argued that they were being caused 'injury' through the loss of federal funds as a direct consequence of the census count missing people. The Bureau's defence was not a denial of the existence of the underenumeration but a claim that there was no 'statistically defensible' method to adjust the 1980 Census. The ruling, reproduced in Werker (1981), went against the US Bureau of the Census, along with other defendants including the then President Jimmy Carter. This led to more hearings during the 1980s and the original ruling was subsequently overturned.

While this was happening the US Census Bureau was planning for the 1990 Census. Barbara Bailar[3] states that the 1986 pre-test showed that adjustment was technically feasible so the decision to not adjust, which had been taken by the Bureau for technical reasons in 1980, was taken for political reasons following the 1990 Census. This decision not to adjust the 1990 Census again led to court hearings and the setting up of National Academy panels to address the problem. The Bureau felt that a weakness of the 1990 Census had been the perception that the estimation of underenumeration was an 'add-on' to the actual Census rather than an integral part of it. They therefore started to plan for a 'One Number Census' in 2000. This is a census where the estimation of, and adjustment for, underenumeration is an integral part of the whole census process.

---

[2] Article by Rosie Waterhouse in The Independent, 17th October 1992, p8.

As the OPCS, and later the Office for National Statistics (ONS), started to plan for the 2001 Census there was a similar move towards the idea of a 'One Number Census' but without the difficult political climate of the US. The main users of the census outputs were already familiar with the idea of adjustment for net underenumeration. Following both the 1981 and 1991 Censuses the population counts by age and sex for each local authority district had been adjusted for net underenumeration estimated using the follow-up survey in 1981 and demographic methods in 1991. These counts formed the basis of the inter-censal series of mid-year population estimates. To extend this adjustment to all census outputs, integrate the follow-up survey as a major component of the whole census process, and plan the project well in advance was an obvious way to avoid the problems of 1991 and to deliver a 'better' product to the users. This thesis is a part of that project, leading to a One-Number Census throughout the UK in 2001.

1.3) Objectives of the Research

The research in this thesis does not cover every aspect of the processes associated with undertaking a One-Number Census in the UK. Its main objective is to present the statistical methods that have been proposed to enable a follow-up survey to deliver the data needed for the production of a One-Number Census. Consequently, the research described here does not deal with the many practical problems that arise such as fieldwork procedures to collect the data, the editing and coding of data, or matching information between the follow-up survey and the census. These are all extremely important for the success of a One-Number Census and are reported in the many working papers produced by ONS as part of the planning for the 2001 Census.

It is also important to realise that this research is based on the assumption that there will be underenumeration in the 2001 Censuses and therefore its extent must be estimated. However, it does not deal with the issue of what causes the underenumeration. It also does not deal with other census errors such as mis-reporting. Both of these issues are important issues understanding them for one census will inform the planning of the next census. However, they are beyond the scope of

[3] Ruminations on the Census. Article by Barbara Bailar in Amstat News, August-September 1997, p1.

this research. In general they have received less attention than the measurement of underenumeration although a recent paper by Iversen, Furstenberg Jr., and Belzer (1999) considers mis-reporting in the 1990 US Census.

1.4) Organisation of the Thesis

The next chapter reviews in more detail the estimation of underenumeration in the census and particularly looks at the statistical methods used for this purpose in the US. It also looks at what was done following the 1991 Censuses of the UK. Chapter three looks at the proposed design for the follow-up survey in 2001. Chapter four then deals with the estimation of underenumeration for large sub-national populations using data from the proposed follow-up survey. The estimation methods are assessed using a simulation study. Chapter five looks at the design of the follow-up survey and its use for estimation of census underenumeration with respect to the production of a One-Number Census database for Northern Ireland. Chapter six goes into detail on the proposed method for taking the results of chapter four and chapter five to produce a One-Number Census. Some of this is joint work done with Dr. Fiona Steele from LSE and where that is the case it will be made clear in the text. Finally, chapter seven draws some conclusions and points to where research is still needed to ensure that a One-Number Census for the UK can successfully be undertaken in 2001.

Chapter 2 – Review of the Literature

2.1) Introduction

The role of this chapter is to place the estimation of underenumeration in the 2001 Censuses of the UK in both a national and international context. The first section will look at the post-war censuses of the US up to and including the developments for the 2000 Census. The US is considered first due to its long history of estimating census underenumeration using a follow-up survey. The second section will look at the last two sets of censuses in the UK, particularly concentrating on the 1991 Censuses. This is important as it is the 1991 Censuses that provide the back-drop for the 2001 Censuses. The final section will review in more detail the methodology that has traditionally been used to measure census underenumeration, with particular emphasis given to dual system estimation.

2.2) History of the Estimation of Census Errors in the US

2.2.1) The 1950 Census

The 1950 Census was the first post-war census of the US population; it was also the first to make a serious attempt to measure census errors beyond estimates of net underenumeration at the national level using demographic techniques. Work on the 1948 Census of Agriculture suggested that it was possible to use sampling techniques to design a follow-up survey to measure all types of census errors. The basic principal was to re-enumerate a sample of areas after the traditional census. The logic behind this approach was that since the survey was on a much smaller scale than the census it would only use well trained and highly motivated enumerators who would be easier to manage. Therefore, the repeated count in the sample of areas would be of higher quality. In particular, the methodology used in 1950 required the repeat count to identify all missed households and individuals in the sample areas. By any standard this is and was an unrealistic expectation and recent work in Darga (1999) argues that on the ground, at least, any follow-up survey to a census will face all the same problems as the census and some of these may even be worse.

The view in 1950 was not so pessimistic. The survey had a standard population survey design. The US was stratified into different groups and within each stratum a sample of areas was drawn. Marks, Parker Mauldin, and Nisselson (1953) review in detail the design issues that were considered, in particular choices between clustering for cost efficiency verses sampling efficiency. The survey then attempted to 'correctly' re-enumerate the sampled areas. This re-enumeration was designed to:

a) identify households completely missed by the census and the individuals contained within them (*underenumeration*).
b) identify individuals missed by the census within households counted by the census (*underenumeration*).
c) identify households and individuals incorrectly included in the census (*overenumeration*).
d) identify errors in the answers given to questions regarding households and individuals correctly counted by the census (*reporting errors*).

The last aim required a 'dependent' re-enumeration in the sense that survey enumerators needed a record of exactly what the census had recorded. This enabled differences between answers given in the survey and census to be identified in the field and probes used to determine the 'correct' answer. The identification of overenumeration in the field also required the survey enumerator to know at the very least that the census had counted a particular person to be able to check whether their inclusion was correct. However, subsequent analysis of the 1950 Census by Coale (1955) and the US Census bureau (see Marks and Waksberg, 1966) agree that the survey failed at identifying underenumeration although there is disagreement about the extent. This appeared to be the case, particularly for persons missed in counted households where having the census record tended to mean the survey enumerator just repeated the census listing. It was also due to this requirement that the survey enumerators were 'perfect' at finding households. Any individuals in household missed by both the 1950 Census and the follow-up survey would remain as undetected underenumeration.

In their review of the initial results Hansen, Hurwitz, and Pritzker (1953) raised an important issue, the fact that both the original census results and any results based on the follow-up survey would be subject to error. They finished their review with a challenge to census users to consider the issue of errors in census data and what level is acceptable. The debate that has followed the 1990 Census, and is particularly raised by Darga (1999), suggests that this is unresolved. Following the 1950 Census, while there was discussion regarding the deficiency of the follow-up survey results, it also appears to have been accepted that the census results were at least equally deficient. This is reflected in attempts to get alternative estimates of census errors at the national level. Coale (1955) produced a set of adjusted population counts at the national level by age, sex and race. This was based on a combination of methods to build-up the 1950 population using the 1950 Census, the 1940 and 1930 Census, vital registration data, and information from the follow-up survey. The essence of the method was based on an assumption that census errors had been constant over time so cohorts could be followed through time, adjusting for deaths and migration to build the 1950 population. The paper acknowledges its own weaknesses (any method of estimating underenumeration requires some assumptions) and Coale rejects his method for the older ages and relies on the follow-up survey as do the 'minimum reasonable' estimates produced by the US Census Bureau. The final estimate of net underenumeration in Coale (1955) was 5.4 million persons, considerably higher than the US Census Bureau 'minimum reasonable' estimate of 3.7 million persons (Marks and Waksberg, 1966) produced by a combination of demographic methods and the follow-up survey estimates. However, both are higher than the estimate based on the follow-up survey alone of 2.1 million persons plus or minus 340,000 reported by Hansen *et al* (1953).

2.2.2) The 1960 Census

Estimation of net underenumeration in the 1960 Census built on the knowledge gained from 1950. Steinberg, Gurney, and Perkins (1962) highlight the much shorter period between the two counts as the single most important improvement. There was also a change to a more independent count. In 1960 the survey enumerators had no information about the number or characteristics of individuals found by the census in

an occupied housing unit. Differences were reconciled using a third visit. This resulted in Steinberg *et al* (1962) reporting the 1960 estimate of those missed from counted units as being roughly two million compared to less than a million in 1950. They conclude that the initial results from the re-interview approach suggest that work using this approach had reached 'maximum intensity' and any future improvements would be through improved processing and better questionnaire design rather than changes to the design and conduct of the survey.

The review by Taeuber and Hansen (1964) gives a preliminary evaluation of the entire 1960 Census including both coverage and quality of the census. They consider the impact of introducing sampling for some of the more detailed questions and therefore the use of a long-form for a sample and a short-form for everyone. The paper looks at general problems such as response bias and variance as well as more specific issues such as age heaping and comparisons with the 'Current Population Survey' to assess employment data in the 1960 Census. Their short section reviewing census coverage concludes that the improvements introduced between 1950 and 1960 resulted in more reasonable estimates of census net underenumeration, consistent with the independent demographic estimates presented by Akers (1962).

The review by Marks and Waksberg (1966) looks in more detail at the different evaluation programmes used after the 1960 Census to assess coverage. In particular, they consider the two surveys that generated the coverage results reported in both Steinberg *et al* (1962) and Taeuber and Hansen (1964). The first was a survey of areas, similar to 1950, to assess the coverage of housing units and those individuals missed in housing units. However, unlike 1950, this survey did not collect information on the individuals' characteristics and gave no estimate for people missed in enumerated units. This was assessed from a second sample of 15,000 census addresses that were then re-interviewed. It is this second sample where the independent re-enumeration reported by Steinberg *et al* (1962) is carried-out. The second sample also attempted to estimate coverage of housing units by checking whether the neighbour had been counted in the census. Both Marks and Waksberg (1966) and Steinberg *et al* (1962) considered these estimates of missed housing units generated by the second sample to be too low. This results in the 1960 Census estimates of coverage being a

combination of estimates due to complete housing units being incorrect (the area sample) and estimates due to enumerated housing units being partially incorrect (the list sample).

The US Census Bureau did not rely solely on re-interview surveys to generate estimates of coverage errors. Marks and Waksberg (1966) report the results of an evaluation using record checks. The approach involved the generation of an alternative population list to the 1960 Census. In the absence of a population register for the US this was generated from the 1950 Census adjusted for those known to have been missed using the 1950 follow-up survey, birth registration data, and data on registered aliens. Marks and Waksberg (1966) suggest that the coverage of this independent list was considered to be about 98 per cent. A sample of individuals was drawn from this list and the individuals were contacted. The aim was to estimate gross underenumeration in the 1960 Census by establishing whether the sample from the independent list had been counted in the census. The major problem with the approach was updating the address information for the sample, which meant that about 1,000 out of approximately 7,000 sampled units were not contacted. Therefore, to use these data, Marks and Waksberg (1966) make assumptions about the coverage in these remaining addresses and generate a minimum and maximum estimate of underenumeration. Both the demographic estimates and the estimates from the surveys lie within these extreme estimates.

Marks and Waksberg (1966) acknowledge that this approach has some deficiencies, not least the fact that up-to-date address information proved very difficult to obtain. It also only generated estimates of gross underenumeration whereas the re-enumeration surveys measured both overenumeration and underenumeration. However, its main advantage was that it could be considered independent of the current census. Even where the list was mostly generated from the previous census, Marks and Waksberg (1966) argue that this independence still holds as the considerable variation in census coverage by age means that there is little correlation between census errors and list omissions. Consequently, despite the problems associated with setting-up the lists, they argued that this approach warranted further research with much larger samples of records being drawn.

2.2.3) The 1970 Census

The review of the re-enumeration surveys of the 1960 Census in the previous section certainly suggests that they were reasonably successful at getting the net underenumeration correct (at the national level) in so far as they gave results that were consistent with the demographic analysis. However, Siegel (1974) calculated national estimates of underenumeration based on 'demographic techniques' and made the following statement criticising re-enumeration surveys.

> "The leading alternative methods for evaluating census data, namely case-by-case checking or matching techniques, involving a re-interview survey, a prior sample survey, or independent lists and records, have, in our experience, shown such serious limitations as devices for measuring the coverage of the total population and the accuracy of the counts by age, sex, and race that the principal reliance has been placed on the methods of demographic analysis for measuring coverage and accuracy in 1960 and 1970. These alternative methods either greatly understated the undercoverage rate or provide too broad a range of estimates in 1950 and 1960; the estimate obtained by demographic analysis proved to be much more reasonable. The case-by-case methods are handicapped by problems of matching, and the results are affected by sampling error."

This move away from using re-enumeration surveys to measure coverage is also reported in Kaplan (1970) and Waksberg and Perkins (1971). The latter argue that:

> "One plausible hypothesis for the failure of the re-interview method to provide reasonable estimates of undercoverage in the census is that the re-interview method is so closely patterned on the census enumeration procedure that errors in census coverage are highly correlated with coverage errors in the re-interview."

Both these statements seem at odds with the evaluation of the 1960 Census but are in line with the problems found in 1950. In fact, changes to the evaluation programme in 1960, highlighted in the previous section, were designed to overcome the problem of correlated errors and ensure more independence. They also ignore the strengths of the re-enumeration surveys in that they give estimates of both overenumeration and underenumeration as well as giving some insights into why people are missed (Marks and Waksberg, 1966).

The 1970 Census did use re-enumeration surveys as part of the complete evaluation programme. For example, Waksberg and Perkins (1971) highlight the use of re-enumeration samples to check the ability of the census enumerators to both find housing units and to then correctly classify them as vacant or non-vacant. A large re-enumeration sample was also used to look at content errors as in both 1950 and 1960. However, it only focused on a limited number of variables with much of the analysis coming from a match between 1970 Census data and the Current Population Survey.

2.2.4) The 1980 Census

The evaluation programme for the 1980 Census moved back to the explicit use of surveys to measure census coverage. However, the main tool used to measure underenumeration was not a special survey, as in 1950 and 1960, but the sample of individuals responding to the Current Population Survey for two months during 1980. In addition, a sample of 100,000 households that had been enumerated in the 1980 Census was selected and re-enumerated in order to estimate the different sources of overenumeration. The bringing together of these different sources is summarised in Bailar and Jones (1980). They also mention the use of dual system methodology (Sekar and Deming, 1949) with both the census and the Current Population Survey as well as with these sources and administrative lists for different population groups. Along with demographic methods, these methods were designed to give a set of estimates (as in 1960 and 1970) that could be combined and reviewed to get agreed estimates of underenumeration. The use of dual system methodology recognised the fact that any re-enumeration survey is also likely to suffer from underenumeration and therefore also miss people in the sampled areas.

However, the 1980 Census evaluation programme had to face problems that had not previously occurred. Bailar and Jones (1980) observed that in previous censuses estimates of underenumeration were mainly as a measure of the quality of the census. However, by 1980 there was an increasing use of the census counts to distribute federal funds, meaning that the underenumeration was now politically important. It was this climate that led to the ruling against the Census Bureau, although it was subsequently overturned on appeal. The full initial ruling is reported in Werker (1981) and states that while the methodology for the 1980 Census was reasonable it was not efficiently applied to New York State and specifically to New York City. Therefore, the high underenumeration would cause a loss of funds to the plaintiffs. For the first-time, the ruling states the need to use the estimates from evaluation programmes to 'adjust' the 1980 Census. The judgement acknowledges the difficulty of doing this but states that the issue is not a perfect adjustment but an adjustment that makes the imperfect census more closely reflect the true population.

In his paper addressing the issue of adjustment, Trussell (1981) discusses the issue of what would be acceptable. He points to the problem of not knowing what the truth is, an obvious but important point, and argues for an adjustment procedure that is robust to the assumptions that are needed for the procedure to work. He suggests a synthetic approach to producing adjusted counts; combing estimates from demographic methods that are considered very good at high levels of aggregation, with possibly poorer quality estimates at lower levels of aggregation obtained by matching to other lists. The higher the aggregation of the estimates of underenumeration the more extreme is the assumption of homogeneity required for this synthetic approach. Trussell (1981) completes his paper by supporting the view of the US Census Bureau that estimates of underenumeration generated from the 1980 Census are not sufficiently good quality to be assured that the adjusted data would be of 'better' quality. He concludes by stating that many politicians miss the point that it is the distribution of the population that is crucial for allocation purposes and not just getter closer to the true total.

Following the 1980 Census there was considerable interest shown by statisticians in the topic of adjustment, both from those in favour and those against. Ericksen and Kadane (1985) developed what they claimed to be a defensible method of adjustment after rejecting the 'complete-coverage' model that the US Census Bureau pursued in 1980. They argued that the latter was a futile exercise both on grounds of spiralling costs and the fact that procedures aimed to increase coverage often also increase erroneous inclusions. In particular, these authors developed procedures for combining all the available data from the follow-up studies and demographic methods to produce what they claimed to be statistically defensible estimates. The approach involves applying dual system estimation but using external population estimates for subgroups to allow for the situation when the data sources are not independent or none of the lists can be considered perfect. The estimates of the odds ratio of census coverage relative to coverage by other sources calculated for certain subgroups were then applied to other subgroups for which population estimates were not available. They also briefly discussed the use of many lists but rejected this approach for the 1980 Census as the US Census Bureau only had the Current Population Survey data augmented by a couple of additional administrative lists including IRS records.

The second stage of the procedure is the production of counts for much smaller areas. Ericksen and Kadane (1985) use a synthetic model based on age, sex, and race but then extend this to a regression model. This extends the ideas of Ericksen (1974), building a regression model that predicts the ratio between the 'truth' and the census in terms of 'symptomatic' variables that include factors such as age, sex, and race but may also include more qualitative measures of, for example, how the census went in a particular area. They also borrow ideas from Fay and Herriot (1979) amongst others to allow for the fact that the ratios are themselves survey estimates. Essentially, this type of regression model helps to smooth out the variability in estimates derived from the follow-up survey.

A critique of this approach is given in Freedman and Navidi (1986). They focus particularly on the use of the regression model and the statistical assumptions behind the model, for example the assumption that sampling variances are known without error, an assumption that cannot be true as they must be estimated from the sample

data. They also point out the fact that the estimates are unstable to changes in the independent variables in the model. This is contrary to a criterion that Trussell (1981) argues would be desirable, that is, stability to small changes. They also consider some of the practical problems with applying dual-system estimation, particularly the implications of matching two data sources where imputation has been used to fill-in missing data on the sources and also to fill in the match status when it cannot be determined. Ericksen and Kadane (1985) acknowledge problems with the matching and recommend that some cases should be excluded at the start when there is concern over the quality of the data rather than allowing possibly poor data to introduce bias into the estimates.

A detailed response to the criticisms made by Freedman and Navidi (1986) is given in Ericksen, Kadane, and Tukey (1989). In particular, they consider the concerns surrounding the use of the regression model and its assumptions, the choice of which series generated from the coverage evaluation programme to choose, model fitting, and the 'best' set of independent variables. They tackle these issues by repeating much of the analysis for several different sets of data and demonstrate that the results are usually 'robust' to the different choices. They argue strongly that, just because it is 'hard' to decide on the statistically 'best' approach to adjustment, this is not an excuse for not making an adjustment. In fact, they reference the work of Schirm and Preston (1987) which demonstrates that under plausible assumptions about the distribution of underenumeration by geography and socio-demographic variables even a very simple approach using synthetic estimation is an improvement over the unadjusted census in terms of the distribution at the state level. Another issue raised by Freedman and Navidi (1986) is whether it is sensible to take the model and apply it outside the sample. On this point Ericksen *et al* (1989) consider some strategies for dealing with the 1980 Census. However, the key point they raise is the need to design any follow-up survey for the 1990 Census to make the blocks used as the basis for the design as homogeneous as possible. This makes any extrapolation from the sample areas to the non-sample areas more defensible.

The main complaint against adjustment seems to be based on the practical issues surrounding matching between the census and the follow-up survey and which of the

several data sets, generated using different assumptions and two waves of the Current Population Survey, to use. Freedman and Navidi (1986) also challenge the statistical assumptions behind the model. However, Ericksen *et al* (1989) point out that their statements against the independence assumption in the regression model are somewhat far fetched. They also point out that the underestimation of standard errors claimed by Freedman and Navidi (1986) would not alter the outcome of the modelling and the adjustment would still be an improvement.

The issue of the regression approach is taken-up by Cressie (1989) who also challenges the appropriateness of some of the assumptions, particularly the assumption of a constant residual variance when modelling counts. He develops an Empirical Bayes estimator as an alternative. One major advantage of the approach is that it is what Cressie (1989) calls level consistent. In other words it easily allows aggregation of say North and South Dakota or the disaggregation of Los Angeles from California. The approach developed by Cressie (1989) also allows for more general models than the work by Ericksen and Kadane (1985), but the price is a greater level of complexity that makes explanation to census users more difficult. Again, the estimator relies on knowledge of the variances for the stratum means model, a model in which underenumeration is constant within specified strata, which is used to smooth out the estimates of underenumeration, and the sampling variances of the original dual-system estimators, as well as some well estimated population totals. Cressie (1989) acknowledges that the estimation of the variance parameters can be unstable and introduces an initial stage that smoothes the variances by collapsing across the strata in the model.

The approach in Cressie (1989) gives an efficient way to model underenumeration through "appropriate stratification and heteroscedastic modelling of variances". The paper applies the models to the 1980 Census data. Using a squared error loss function he shows that there is always a greater risk from using unadjusted census data. The models and assumptions in the paper were specifically designed for the 1980 Census but can easily be applied to any census. He also states that introducing spatially dependent variation would also be possible rather than making independence assumptions throughout the modelling.

2.2.5) The 1990 Census

The US Census Bureau claimed that there was no "statistically defensible" (Werker, 1981) method of adjustment for the 1980 Census, but this claim was based on the quality of the data that was available on underenumeration in 1980 rather than the availability of statistical methods to carry-out an adjustment. Therefore, the plans for the 1990 Census included a large-scale evaluation programme based on both demographic estimates and a post-enumeration survey (PES). An overview of the design of the PES and the practicalities, such as matching and producing estimates, is given in Hogan (1992). The design was very similar to the approach used in previous dedicated surveys of census underenumeration. A national sample of block clusters was selected after stratifying at the national level based on what was known about the distribution of census underenumeration in 1980. To estimate underenumeration the survey attempted to construct an independent list of housing units and enumerate all those individuals within the housing units who should have been enumerated as residents on census night. To estimate overenumeration all the census returns for the sampled block clusters were also checked to ensure that all the individuals and households were correctly included. (In reality, it is only necessary to check those individuals and households for whom there is a census return but no subsequent survey response.)

The estimation strategy required the formation of post-strata and the application of a weighted dual-system estimate within the post-strata. The original strategy used 1,392 post-strata defined using region, census division, race, place/size, and housing tenure as well as age and sex. There was also a special group for American Indians. The estimation strategy then proceeded along the lines of Cressie (1989) by estimating 'raw' adjustments using the dual-system estimates, using these to fit a regression model to predict adjustment factors, and then using a weighted average of the 'raw' adjustment and the predicted adjustment to get a final smoothed adjustment. As suggested by Cressie (1989) the sampling variances were pre-smoothed as their estimation was considered unstable for some of the post-strata with small samples.

This approach received some criticism both in respect to the formation of the post-strata and the variance assumptions in the smoothing. In response to the concerns Hogan (1993) presents results of a modified approach that used a new set of post-strata that were considered more homogeneous with respect to census underenumeration but which were also fewer in number. The matching was also re-visited and this resulted in some modifications. There were also corrections to errors that had occurred during the original processing. The results of the changes are "more stable estimates with a sharper distinction between groups" (Hogan, 1993). Another important conclusion made by Hogan (1993) is that operationally the PES succeeded and the processing of the data was achieved in the specified time-frame demonstrating the practical feasibility of adjustment.

As in 1980, the issue of the estimation of underenumeration was subject to considerable debate and litigation, once the Secretary of Commerce announced his decision not to adjust the 1990 Census on July 15th, 1991. As with the post-1980 debate, much of the criticisms focused on the application of dual-system estimation and the models for making the adjustment. A particularly interesting set of papers is published in Statistical Science (November, 1994). The paper by Breiman (1994) re-visits the concept of 'total error' in the proposed 1990 adjustments. This is based on work done by the US Census Bureau reported in Mulry and Spencer (1993). The key point of both papers is that there are significant sources of error in the proposed adjustments although the resulting conclusions are somewhat different. This point is picked-up by Belin and Wolf (1994), who criticise the conclusions of Breiman (1994). A clear message from this criticism is that any future attempts at estimation of underenumeration and a subsequent adjustment should carefully address the sources of error in both the statistical models and the processes involved with the collection and preparation of the data.

This point is at the core of the paper by Freedman and Wachter (1994) who address the issue of heterogeneity in the post-strata and its subsequent impact on an adjustment process. Diamond and Skinner (1994) in their discussion of the paper point out that care should be taken to not confuse this with biases in the dual-system estimator due to heterogeneity. The issue being assessed here is that an estimate for

the State of California is calculated by applying synthetic estimates to individuals in the different post-strata that appear in the State. However, the same post-strata may also apply to people in another State, New York for example, and this assumption of homogeneity between States may have an impact on the estimated State totals. The result is that, for the variables Freedman and Wachter (1994) analyse as proxies for underenumeration, the synthetic adjustments have failings. This is perhaps not surprising to anyone who has attempted to make small area estimates from survey data. This suggests that any subsequent PES needs to have a large enough sample to facilitate the production of direct estimates for state totals. This will have two positive outcomes, remove the issue of bias due to an untrue synthetic assumption, as well as make synthetic assumption for sub-state populations more plausible, or indeed allow the use of more complex small area models such as those extensively reviewed in Ghosh and Rao (1994).

2.2.6) Plans for the 2000 Census

The complexities of the court rulings surrounding the 1990 Census have impacted on the US Census Bureau's plans for the 2000 Census. The initial plan was to fully integrate coverage improvement into the census and produce a single set of numbers by the deadline of 31st December 2000. However, the controversy of the 1990 Census adjustment has not gone away and as recently as the 25th January 2000 the Supreme Court ruled against the use of adjusted census data for apportionment of the House of Representatives. The original plan for the 2000 Census included an element of sampling during the fieldwork follow-up of the actual census count. The advantage of this was a cost saving in terms of the number of temporary enumeration staff that would be required as well as requiring less time to achieve the coverage targets. As a consequence of finishing census fieldwork earlier, the independent coverage measurement would be more effective by getting it in the field more quickly. The original plans and resulting changes are outlined in Wright and Hogan (1999).

The plan that has been implemented involves a 'traditional' census count followed by an independent coverage measurement survey very similar to the 1990 Census. The key difference is that there has been extensive research by US Census Bureau in the

preceding ten years to plan for adjustment, and the follow-up survey will be considerably bigger, 300,000 housing units compared to 165,000 housing units in 1990. Hogan (2000) reports some of this research in respect to the application of dual-system estimation and references the more detailed work by Griffin (2000) on the calculation of the DSE and the treatment of movers, Griffin and Haines (2000) on the formation of post-strata, and Cantwell (2000) on the treatment of unresolved match status and on missing data procedures. The philosophy for 2000 has emphasised simplicity over complexity. Although the intention is to form more post-strata, the larger sample size means that variance smoothing is unnecessary, and this approach is certainly more 'transparent' to the users. In addition, the allocation of unresolved cases after matching in 1990 using a hierarchical logistic regression model, as reported by Belin *et al* (1993), was criticised by Breiman (1994) who argued that the results were not well supported by the error studies carried-out following the 1990 Census. While Belin and Wolf (1994) argue that the same studies show it to have been successful, the approach in 2000 will form classes by demographic and geographic characteristics, and simple cell probabilities will be used rather than predicted probabilities. As in 1990 synthetic estimation will be used to adjust the census down to the census block level. This will give adjusted block totals for individuals and households. The current proposal is then to release a two-number census, the unadjusted census database and a census database where imputation has been used to add the 'missing' housing units and the 'missing' individuals.

2.2.7) Underenumeration Measurement in US Censuses - Conclusions

Work done by the US Census Bureau demonstrates that estimation of census underenumeration is technically feasible with an adjustment to the census database for that estimated underenumeration. This is based on a large-scale re-enumeration of a sample of small groups of housing units that is undertaken independently of the census. This sample allows for the estimation not only of those missed by the census but those incorrectly counted by the census. Estimation also accounts for the fact that the survey will not count its areas perfectly and consequently some individuals will be missed by both the census and the survey. This is based on the experience of using follow-up surveys in 1950 and 1960 where, especially in 1950, there is evidence to

support the fact that the follow-up survey also missed people. The problems faced in 1980 and 1990 relate more to the practical issues of carrying out this survey and the subsequent estimation. In 1990 the issue of deciding when one approach is better than another also surfaces. What the 1990 experience in the US very clearly demonstrates is the need to ensure that whatever approach is adopted for the UK Censuses in 2001 it has been openly discussed with all parties and all parties are agreed. It is vital that this happens before interested groups can see the data and the possible impact of adjustment.

The US Census Bureau has also looked at other approaches and has relied heavily on demographic estimates at a national level to give the definitive results on net underenumeration. There has also been work following the 1960 Census and prior to the 1990 Census on the use of administrative data as alternative lists of the population to compare to the census. In both cases these approaches have been hampered by the lack of a single good list, such as the population register in Sweden (see Lyberg and Lundström, 1994) and have not been pursued further.

2.3) Estimation of Underenumeration in the 1991 Censuses of the UK

Since the 1971 Censuses of the UK, there has been a process of evaluating both the quality and coverage of the results. This evaluation has included the adjustment of population counts, used in the mid-year population estimates, for estimated net underenumeration. The intention in the 1991 Censuses of the UK was that the basis of the evaluation programme be a follow-up survey called the Census Validation Survey (CVS). This strategy was based on the successful 1981 Census evaluation programme. The 1981 Census evaluation programme was reviewed in Britton and Birch (1985), where the following three main objectives are stated:

(i) to check whether all persons present on census night in a private household had actually been correctly enumerated by the census;

(ii) to verify the classification by census enumerators of unoccupied residential accommodation;

(iii) to assess the quality of replies given to census questions, and hence the accuracy of the published 1981 Census results.

The original plan for the 1981 Censuses evaluation programme addressed assessment of the actual coverage of households via a separate survey. However, late in the planning stage it was decided all these objectives would be addressed by a single follow-up survey.

The basis of the design of this survey was a stratified multi-stage sample of blocks of enumeration districts (EDs) referred to as census districts. At the first stage the census districts were stratified by region and type of area (metropolitan, non-metropolitan, inner and outer London) and from the strata a sample of 300 census districts (including 29 from Scotland) was selected with probability proportional to an estimated population size that had been developed for use in planning the 1981 Census. In addition all EDs were graded according to the likelihood of census response based on the 1971 Census, using a classification suggested by Webber (1977). In particular, a census district containing EDs graded as difficult to enumerate was selected with probability proportional to twice its estimated size. The second stage then selected a cluster of four EDs per selected census district with probability proportional to its estimated number of households. Again, the size was doubled for graded EDs. The final survey can then essentially be thought of as several samples of households and individuals all drawn from the same sample of enumeration district (ED) clusters. Each sample of households from the selected ED cluster then assesses a particular component of the census fieldwork procedures. In 1981 this included a sample of properties identified as vacant in the 1981 Census, a sample of households that responded to the 1981 Census that was used to check the quality of census data as well as the coverage of individuals within counted households, and all households within the ED cluster identified as missing from the 1981 Census.

The above strategy was highly successful in 1981 in detecting the whole spectrum of census errors. The net level of underenumeration of 0.45 per cent reported in Britton and Birch (1985) for persons was, with the exception of estimates for 0 to 4 year olds, consistent with demographic estimates. Based on this success an integrated coverage and quality approach was again taken in 1991, and a very similar approach to the design of the Census Validation Survey (CVS) is reported in Heady, Smith, and Avery (1994). This involved a multi-stage strategy to select census districts and then clusters

of enumeration districts, with the enumerator then selecting samples of households to assess the different sources of census errors.

The above strategy was not successful in 1991. Demographic estimates in Heady *et al* (1994) suggest an underenumeration of about two per cent (p. 43) for England and Wales after adjusting for definitional differences compared to the 'best' CVS estimate of 0.5 per cent ± 0.22 per cent (p. 31). The decision to accept, in most cases, the demographic estimates in preference to the 1991 Census adjusted by the CVS is outlined in OPCS (Spring 1993). The decision was based partly on an analysis of the sex ratios in the unadjusted census, the adjusted census, and demographic estimate. The sex ratios in the adjusted census for young men were below one leaving only two possible explanations; mass undetected emigration between 1981 and 1991 by young men but not young women, or a differential underenumeration of young men that the CVS had not detected. Further work focused on other possibilities for the differences between the adjusted census and the demographic estimate. The work concluded that there was no evidence to support a major over estimate by the demographic method as this was based on birth registration, death registration, net migration estimated from the International Passenger Survey, with the 1981 Census as the base. Birth and death registration are considered more or less perfect. The article also rejects a problem with the 1981 Census or a problem with migration estimates as the main cause. However, the article lays the blame on additional underenumeration in the 1991 Census that the CVS had failed to measure.

Once it was accepted that the demographic estimate was the basis of the national population estimate for 1991, the problem was then how to allocate the additional people to the local authorities. This was achieved by adjusting the sex ratios in the 1991 Census for large groups of local authorities so that they were consistent with average sex ratios from the 1971 and 1981 Censuses, while at the same time making an overall adjustment upwards to meet the agreed national estimate. The effect of this was to make particularly large adjustments to young males in the inner city areas relative to other age groups and relative to females of the same age. Heady *et al* (1994) sets out the method in detail and gives the final underenumeration adjustment

factors that were used, along with the work in OPCS (Autumn 1993), to produce the 1991 local authority mid-year population estimates.

The 1981 population estimates had included an element of adjustment for census underenumeration based on the 1981 follow-up survey. The problem in 1991 was that the level of the adjustment made for underenumeration in some of the local authority districts left local demographers unsure about the validity of census counts at ward and ED level, particularly for the allocation of resources (Simpson, 1994). In addition, the fact that the adjustments were not based on the CVS meant that it was difficult to get a feel for the characteristics, beyond age and sex, of the people missed. The sample design for the CVS had, by the stratification used, assumed that underenumeration would be homogeneous across reasonably broad groups of the population (ie the major metropolitan cities were all assumed to have the same level of underenumeration). Even if the CVS had made an acceptable estimate at the national level it is questionable whether this homogeneity assumption would have been sensible. The result of this was an attempt by the Economic and Social Research Council funded project 'Estimating with Confidence' to look at the issue of adjusting the census at levels lower than the local authority district. Simpson, Cossey, and Diamond (1997) describes a set of publicly available adjustments. The adjustments were the result of a consultation process and were based on a regression model that used unemployment and the level of census imputation to share out the underenumeration allocated to each local authority district. The adjustments also dealt with the definitional and timing differences between the 1991 Census and the 1991 mid-year population estimates. An example of this is the movement of students from their 'home' address (1991 Census) to their term address. Southampton, for example, had an adjustment of plus 10,400 of which 3,600 were students and 6,800 was census underenumeration.

There is a general perception that the underenumeration problem in 1991 was due operational problems with the 1991 Censuses, an easy conclusion to make as underenumeration increased by three or four times compared to the level in 1981. Therefore, one solution to underenumeration would be to ensure that those operational problems did not occur in 2001. Unfortunately, the situation is not quite so simple. It

was harder in 1991 than previous censuses for the census enumerators to contact households and individuals during the census period. This appears to have been associated with changes in society resulting in more single person households, more multi-occupancy, and more purpose built blocks of flats in new or converted buildings that utilise electronic entry systems. All these issues make the identification of and contact with households more difficult for census enumerators and it is likely that they will cause greater problems in 2001. This difficulty with contact has resulted in households being allowed to post back census forms in 2001 but this will not alleviate the problem with identification of households in the first place when the forms are delivered. In addition, the actual level of around two per cent underenumeration in 1991 for the national population was certainly in line with estimates for the United States 1990 Census and the 1996 Censuses of Australia, Canada, and New Zealand (Dunstan, Heyen, and Paice, 1999). In other words, compared to other countries, at the national level, the 1991 Census was not a particularly poor census. At the sub-national level, the 1991 Census encountered specific problems counting certain small (in national terms) but important special populations such as students and the armed forces. This is also an issue for the way the census conducts itself and in 2001, for example, the Census will count students at their term-time address to try and alleviate the problems of adjusting the location of students for the mid-year population estimates.

The more serious problem was not so much the existence of the underenumeration, but the inability of the CVS to measure the underenumeration. The 1991 CVS used a methodology that could be described as a dependent re-enumeration, as it started with the census and then checked each procedure (Diamond, 1994). The main problem with this approach is that in the sampled areas the CVS enumerators need to be essentially perfect. There is also the problem of correlated error, where a CVS that uses very similar methodology to the census, without any kind of maliciousness on the part of the target population, also misses those missed by the census procedures. In such a scenario a near perfect CVS would still fail to detect adequately census underenumeration. Heady *et al* (1994) finish their report on the CVS by suggesting possible improvements. They argue for an independent re-enumeration that would allow the use of capture-recapture methods to account for those missed by both the

census and any follow-up survey. This approach is also advocated by Diamond (1994) in a thorough review of the types of individuals missed by the CVS. It is also the approach used by the US Census Bureau in 1990 and the approach proposed for the 2000 US Census.

Other alternatives considered by Heady *et al* (1994) do not require the use of a follow-up survey. One such method would be to use another population list generated from some other administrative source such as health records. The advantage of such an approach is that large samples could be generated very cheaply and matched to census data. The problem is then matching large numbers of records that were never intended for such a purpose. Large overenumeration on the administrative data also becomes an issue. Another method mentioned by Heady *et al* (1994) is the 'reverse record check' used in Canada (see Belley *et al*, 1999). The major potential shortcoming here is the need to determine what has happened to people in the inter-censal period, and in Canada this requires a lot of time and effort using administrative data sources. This would be even more problematic in the UK where the inter-censal period is ten years, twice as long as Canada, and in general administrative data in the UK on population mobility is poor. Heady *et al* (1994) do not formally recommend any method for subsequent censuses but do seem to suggest that no method using administrative data is likely to be sufficiently good to replace some attempt at an independent follow-up survey.

2.4) Methods for Estimating Census Underenumeration

The previous sections of this review have looked at how the estimation of census underenumeration, and the moves to adjust censuses, have developed over time. This has included the use of dual-system estimation as a method of estimating census underenumeration and some of the practical issues associated with this particular methodology. In this section some of the theoretical aspects of this technique and other methods for the estimation of census underenumeration are briefly considered.

2.4.1) Dual-System Estimation and Capture-Recapture Methods

One of the problems with the approach used following the 1991 Censuses of the UK was that the survey needed to be perfect at finding those households and individuals missed by the 1991 Census. However, it seems sensible that this will not be the case and some individuals will possibly be missed by both the census and the follow-up survey. Under such a scenario one can use Dual-System Estimation to estimate the number of such individuals. This approach has been used extensively for the estimation of wildlife populations (see Seber, 1982) as well as estimation in human populations. An early example is Sekar and Deming (1949) who apply the approach to the estimation of total births using both a register and a survey and therefore obtains an estimate of under-registration of births. This was the approach used by the US Census Bureau following both the 1980 and 1990 US Censuses.

In general, suppose that shortly after the census a follow-up survey, often referred to in the literature as a Post-Enumeration Survey (PES), is used to obtain an independent re-count of the population in a sample of areas. After matching it is possible, within those areas in the PES sample, to produce Table 2.1.

TABLE 2.1

Classification of enumeration status

		PES		
		Counted	Missed	
Census	Counted	n_{11}	n_{10}	n_{1+}
	Missed	n_{01}	n_{00}	n_{0+}
		n_{+1}	n_{+0}	n_{++}

Individuals can be assigned to the cells in Table 2.1, and the counts n_{11}, n_{10}, and n_{01} observed. By definition, n_{00} and any margins that depend on it, including the overall population total n_{++}, cannot be observed. Therefore, the problem is to construct an estimate of n_{++} based on the observed data.

Assuming n_{++} is known, the observed counts for the cells in Table 2.1 can be thought of as realisations of random variables generated by an underlying multinomial process with parameters (n_{++}, p_{11}, p_{10}, p_{01}, $1-p_{11}-p_{10}-p_{01}$), where p_{ij} is the probability of being in cell ij of Table 2.1, and the four cell probabilities are constrained to sum to one. Assuming that the same multinomial model applies independently to each individual in the population, the expected value of those counted in both the census and the PES is

$$E[n_{11} \mid n_{++}] = n_{++} \times p_{11} \tag{2.1}$$

and an unbiased estimator for n_{11} would follow from (2.1) by replacing p_{11} with an unbiased estimator $\hat{p}_{11}$ to give $\hat{n}_{11} = n_{++} \times \hat{p}_{11}$. If the census and PES are independent of each other, the expected value of n_{11} can also be expressed as

$$E[n_{11} \mid n_{++}] = n_{++} \times p_{1+} \times p_{+1} \tag{2.2}$$

where p_{1+} is the probability of inclusion in the census and p_{+1} is the probability of inclusion in the PES. An estimator for n_{11} again follows by plugging-in unbiased estimators for the probabilities to give $\hat{n}_{11} = n_{++} \times \hat{p}_{1+} \times \hat{p}_{+1}$. (This estimator is not exactly unbiased as $E[\hat{p}_{1+} \times \hat{p}_{+1}]$ is only approximately equal to $E[\hat{p}_{1+}] \times E[\hat{p}_{+1}]$.) Therefore, replacing the probabilities in (2.2) with appropriate estimators gives

$$\hat{n}_{11} = n_{++} \times \frac{n_{1+}}{n_{++}} \times \frac{n_{+1}}{n_{++}} = \frac{n_{1+} \times n_{+1}}{n_{++}} \tag{2.3}$$

which is an approximately unbiased estimator of n_{11} under the independence assumption. After the census and PES a value for n_{11} is available but the true population total n_{++} is unknown. Therefore, re-arranging (2.3) yields the dual-system estimator (DSE) defined as

$$\hat{n}_{++} = \frac{n_{1+} \times n_{+1}}{n_{11}} \tag{2.4}$$

Although the theory used to motivate the DSE given in (2.4) is straightforward, its practical application creates problems relating both to the plausibility of the underlying multinomial model assumed above and the practical issues of getting a value for n_{11}. These are highlighted below.

a) The DSE assumes that in the target population the matched PES and census counts follow a multinomial distribution. That is, the probabilities of being counted by either or both the PES and the census are **homogeneous** across the population that dual-system estimation is applied to. This is unlikely for most populations.

b) Approximately unbiased estimation requires statistical **independence** between the census count and the PES count. This is impossible to guarantee.

c) It is necessary to **match** the two data sources to determine whether individuals on the lists were counted once or twice. Errors in matching become biases in the DSE.

In the 1990 Census the US Census Bureau tackled problem a) by splitting the population up into post-strata (Hogan, 1992 and Hogan, 1993) based on factors (e.g. race) which were thought to affect an individual's probability of being counted, a method originally proposed by Sekar and Deming (1949). Problem b) is typically handled by operational procedures that ensure the operational independence of the census and the PES. Problem c) is essentially unavoidable but it is absolutely essential to ensure that errors due to matching are minimised. This is highlighted in the earlier review of the literature criticising adjustment of both the 1980 Census and 1990 Census in the US.

If the PES samples the whole population (in other words is another census), the DSE defined by (2.4), under the assumptions already stated, would give an estimate of the total population in a particular post-stratum. However, only a sample of areas are included and therefore the simplest approach is to plug-in design unbiased Horvitz-Thompson estimators of the population quantities in (2.4) to give

$$\hat{N}_{++} = \frac{\sum_{i \in PES} n_{+1i} / \pi_i \times \sum_{i \in PES} n_{1+i} / \pi_i}{\sum_{i \in PES} n_{11i} / \pi_i} \qquad (2.5)$$

where n_{1+i} is the census count, n_{+1i} is the PES count, n_{11i} is the number of matched individuals and π_i is the probability of inclusion for sampled area i within the post-strata. However, the estimator (2.5) ignores the fact that the census provides auxiliary information in the shape of the population counts for all areas, and if these counts are correlated with the true population counts, this can be used in a ratio estimator to give

$$\hat{N}_{++}^{Ratio} = \frac{\hat{N}_{++}}{\sum_{i \in PES} n_{1+i} / \pi_i} \times N_{1+} = \frac{\sum_{i \in PES} n_{+1i} / \pi_i}{\sum_{i \in PES} n_{11i} / \pi_i} \times N_{1+} \qquad (2.6)$$

where N_{1+} is the total census count for a particular post-strata. Alternatively Wolter (1986) motivates estimator (2.6) directly from (2.4) by replacing the unknown population quantities in the DSE with estimates and develops approximations for the bias and variance of (2.6) accounting for the fact that they are generated from two sources; the underlying multinomial model that drives the DSE in (2.4) and the fact that population quantities in the DSE are unknown and estimated from a sample. Equation (2.6) forms the basis of the approach taken by the US Census Bureau. Further modifications are also applied to account for estimated overenumeration in the census due to erroneous enumerations and imputation at the processing stage. This essentially means that N_{1+} is not the actual census count but some adjusted count based on the census. A full discussion of this is in Hogan (1993).

2.4.1.1) Properties of the Dual-System Estimator

Wolter (1986) examines the properties of estimator (2.6) with respect to both the underlying multinomial model and the sampling process, which generates estimates of the unknown population quantities. Of particular interest are the properties of the estimator with respect to the underlying multinomial model outlined above, in other

words the properties of (2.4) ignoring the use of sampling to estimate parameters in the DSE. Using a second order taylor series expansion Wolter (1986) shows that the bias of (2.4) is approximately

$$\frac{(1-p_{1+})(1-p_{+1})}{p_{1+}\,p_{+1}} \tag{2.7}$$

where p_{1+} is the response probability of the first list (usually the census) and p_{+1} is response probability of the second list (usually the follow-up survey). He further shows that the variance of (2.4) can be approximated using a first order taylor series expansion, and ignoring covariance terms, this resolves to

$$\mathrm{Var}(\hat{n}_{++}) \cong n_{++}\frac{(1-p_{1+})(1-p_{+1})}{p_{1+}\,p_{+1}} \tag{2.8}$$

The key point from (2.8) is that the variance is proportional to the population size n_{++} while the first order bias given by (2.7) does not depend on the population size. Therefore, while the relative variance is high for small populations and decreases as the size increases, it is the bias that is particularly important when the population is small.

This property has long been recognised in the wildlife literature and an alternative estimator to (2.4) was suggested by Chapman (1951). In Seber (1982) this modified estimator is given by

$$\hat{n}^{C}_{++} = \frac{(n_{1+}+1)\times(n_{+1}+1)}{n_{11}+1} - 1 \tag{2.9}$$

With some manipulation, see Appendix 2.1, it is possible to re-express (2.9) into the following form

$$\hat{n}^{C}_{++} = \hat{n}_{++} - \frac{(1-p_{1+})\times(1-p_{+1})}{p_{1+}\times p_{+1}} + \text{terms of } O(n_{++}^{-1}) \text{ and smaller} \tag{2.10}$$

from which it is possible to see that the second term in (2.10) will always correct for the positive first order bias of the DSE given in (2.7) with the remaining terms tending to zero. Seber (1982) states that Chapman (1951) goes further to show that provided $n_{1+} + n_{+1}$ is greater than or equal to n_{++}, the unknown population total, the estimator given in (2.9) is exactly unbiased. In most scenarios when one capture is the census and the second is a high coverage follow-up survey this is likely to be true. However, as Wolter (1986) points out, when using dual-system estimation in the form given by (2.6) correcting for the bias in (2.7) is going to be unnecessary anyway as the population sizes involved will typically be 'large'. Therefore, such corrections have not been used by the US Census Bureau.

2.4.1.2) Relaxing the Assumptions in Capture-Recapture Analysis

All the estimators and their properties given in the previous section assume that the counts in Table 2.1 are generated from a closed population by an independent homogenous multinomial process. Wolter (1986) demonstrates that it is actually only necessary to have homogeneity for one of the capture probabilities. Therefore, as recommended by Sekar and Deming (1949), the approaches outlined above rely on the ability to post-stratify the data sufficiently to approximate this. An alternative to homogeneity would be complete heterogeneity of all capture probabilities across individuals but, as Wolter (1986) points out, while this may be the most plausible model, there are insufficient data to estimate all the required parameters. Work by Alho (1990) using a logistic regression model to estimate the capture probabilities is an extension of the post-stratification approach in that it allows formal modelling of the variables that explain the heterogeneity. This approach was applied to data from the 1990 US Census (Alho *et al*, 1993) and the properties of this estimation approach are further discussed in Alho (1994). However, this approach has yet to be used on a large-scale, although it is being further investigated in the 2000 US Census.

The second key assumption is that of independence between the first list that attempts to count the population (usually the census) and the second list (usually a follow-up survey). From the wildlife literature these attempts to list the population are referred

to as capture. When the assumption of independence between the captures fails the dual-system estimator will have a negative bias if the two captures are positively correlated and a positive bias if the two are negatively correlated. The argument is that as both the census and the follow-up survey tend to use similar fieldwork procedures a positive correlation is likely and therefore there will be a negative 'correlation bias'. Ericksen and Kadane (1985) recognised this as a potential problem as do Mulry and Spencer (1993) in their analysis of 'total error' in the proposed adjustments for the 1990 Census. They suggested estimating the dependence for groups of the population where very accurate administrative or demographic counts are available and use this knowledge of how the dependence behaves with respect to the different characteristics of the groups to make sensible 'guesses' for the rest of the population. Another possibility they suggest is the use of multiple lists, citing the work of Fienberg (1972). This approach uses a log-linear model to analyse the multi-way contingency table and allows for dependence between lists, although the highest order interaction is always missing.

The use of multiple lists was rejected as a possible approach for the 1980 US Census, due to the absence of a suitable third list, but was investigated as a possible approach for the 1990 Census. An administrative list was constructed for the 1990 Census Dress Rehearsal and a report of initial investigations using different triple-system models is given in Zaslavsky and Wolfgang (1990). Further developments are given in Darroch *et al* (1993). Despite the statistical advantages of such an approach it was not pursued as a possibility for the 1990 Census and is not part of the plans for the 2000 Census. The problem the US Census Bureau encountered was the construction of a third list which could only be accomplished by a very time-consuming process of combining data from several administrative sources to create a list. There were also concerns regarding how the public would view the confidentiality issues involved with linking census data to tax records, medicare data, school data, social security records, and so on.

In the UK there does exist a list of individuals held as part of the National Health Service records. The problem then becomes the quality of the data on the administrative list, in particular erroneously included individuals and those included

in the wrong place will positively bias any estimate of the population total. This is demonstrated by looking at the estimator given by Darroch *et al* (1993), which assumes independence across all three lists, as the numbers in the numerator are either individuals on all three lists (not subject to erroneous inclusions) or individuals found only on one list. In addition, the use of a third list adds to the complexity of matching the lists, especially when the data held on the administrative source are not designed to help match individuals to either the census or a follow-up survey. For these reasons this approach was also rejected, after initial investigations, by the ONS for use with the 2001 Censuses of the UK.

A final, but less obvious assumption is that the multinomial process is applied independently to each individual. This will not, in general, be the case when the data collection process is not a simple random sample of individuals. Cowan and Malec (1986) consider the problem when first listing households and then listing individuals within households generate the data from each capture. The problem occurs as missing individuals are 'clustered' within households that are missed. Cowan and Malec (1986) develop a model that allows for this, when the capture of households is independent across households and between lists, and within households the capture of individuals is independent across individuals within the household and between lists. They apply the EM algorithm to get an unbiased (first order approximation) estimate of the total number of individuals. Cowan and Malec (1986) also assess the performance of the DSE (2.4) under different scenarios. They demonstrate that ignoring the clustering will matter if the capture probabilities for households vary by size and the magnitude of the (first order) bias will depend on how the observed average household size differs from the true average household size. However, empirical results demonstrate that this is reasonably unimportant in relation to other possible causes of bias such as a failure in matching. To see this, consider the situation given in Table 2.2

TABLE 2.2

Indicative example of household capture probabilities by household size

Household Size	True Distribution	Census Coverage	PES Coverage
1	0.25	0.9	0.8
2	0.33	0.95	0.85
3	0.15	0.99	0.89
4	0.14	0.99	0.89
5	0.05	0.99	0.89
6	0.03	0.98	0.88
7	0.02	0.97	0.87
8	0.01	0.96	0.86
9	0.01	0.95	0.85
10	0.01	0.95	0.85

The data in Table 2.2 presents an indicative example, based on the experience of the 1991 Censuses reported in Heady *et al* (1994), of a population with a particular distribution of households by size and a varying pattern of both census and PES coverage by household size. The key point is that the variation in coverage will result in missed individuals being clustered within missed households and therefore violate one of the assumptions underpinning dual-system estimation. Using the bias formula in Cowan and Malec (1986) the standard dual-system estimator would under estimate the population total of individuals by 0.09 per cent because of the clustering of missed individuals within missed households. To put this in perspective, if the matching process failed to match 0.1 per cent of the individual records that should be matched this would result in a positive bias of the same magnitude when applying the standard DSE. Based on this empirical result it is perhaps not surprising that this approach was not applied by the US Census Bureau in 1990 and is not mentioned in the plans for the 2000 Census.

2.4.2) Other Survey-Based Methods

This review has concentrated heavily on the work of the US Census Bureau and the past experience in the UK. However, Statistics Canada, the Australian Bureau of Statistics, and Statistics New Zealand all carry out thorough evaluation programs. Of particular interest as an alternative method is the Reverse Record Check approach used by Statistics Canada. This has been used in Canada since 1966. An early review, based on the 1976 Census, can be found in Felligi (1980). The application to the 1996 Census is given in Belley *et al* (1999). At the heart of the method is the concept that a population frame, for the same population being counted by the census, can be created. This is done by combining the previous census database with other administrative sources such as births since the last census, deaths since the last census, those known to have been missed by the last census, and migration data.

A sample of individuals is selected from this combined frame and an intensive tracing exercise is carried out to contact the individuals at their current address. As a result of this interview and a subsequent matching exercise to the current census database, it is possible to evaluate whether the sampled person was correctly and uniquely counted by the census, missed by the census, or erroneously counted by the census. Additional studies also estimate overenumeration through extensive matching within the census database. The results are weighted for non-response and non-classification, combined with results from the additional overenumeration studies, and then used to estimate the gross underenumeration ratio and the net underenumeration ratio. The key assumption is that the population list created from the various sources includes the entire population or alternatively, those missing can be considered as missing at random. The first assumption is unlikely to be true in the UK context, especially as sampling from those missed by the 1991 Census would basically be impossible. Under the second assumption the estimator given in Belley *et al* (1999) is essentially a dual-system estimator adjusted for overenumeration. It is also important that the approach is independent of the census so that those who remain unclassified or non-contacted in the sample can also be considered as missing at random. If not independent of the census, it could be argued that those people the survey could not trace are also more

likely to be missed by the census, and hence introduce a bias into the estimate of underenumeration.

Like the follow-up survey approach used in the US, the reverse record check has some practical problems, especially with respect to tracing a current address for sampled individuals. In Canada using other administrative sources, such as health records, helps to update the addresses. The fact that the inter-censal period is only five years also helps. Issues relating to matching are also important for both the reverse record check and the other overenumeration studies. In the UK, where administrative data are of poorer quality and there are issues of data access, and with the inter-censal period being ten years, constructing a population list of sufficient quality and then tracing the sampled individuals would be extremely difficult.

The approach taken by the Australian Bureau of Statistics and, in 1996, Statistics New Zealand, is to use a follow-up survey. The strategy is outlined in Dunstan *et al* (1999). The survey design is very similar to the US, being a stratified multi-stage sample that selects small areas of housing units to be enumerated. The actual design used is based on the design of the labour force surveys in each country and professional interviewers are used to carry out the survey. An important difference is the fact that there is no special sample to collect information on erroneous census counts, this is all collected from the single survey. As with other methods this involves matching between the two databases. The estimation approach in Dunstan *et al* (1999) does not directly use dual-system estimation. However, the approach used is effectively (2.6) but directly stated as a ratio model and assuming that the survey achieves 100 per cent coverage in the sample areas or alternatively, lack of coverage by the survey is at random independent of the census. The survey is used to estimate the 'true' population after correcting for non-response in the survey and to estimate the census count. The estimated population counts are used to define a ratio that is applied to the actual census count to adjust it for net underenumeration in the census. The approach is also discussed by Steel (1994) with respect to the 1991 Census in Australia. This also includes a brief description of the process of reconciliation with demographic estimates and the use of synthetic estimates to adjust the census data for the use of population estimates in the five-year inter-censal period.

2.5) Conclusion

Lyberg and Lundström (1994) argue that the statistician should use all the available information to compile their estimates. In the case of the census this implies not ignoring the problem of differential underenumeration but instead attempting to correct for it using additional information from a follow-up survey, administrative sources, demographic estimates, and matching exercises. However, the debate in the US, for example Breiman (1994), highlights that the combining of data sources must be carried-out carefully or else errors introduced may swamp the error that is being corrected. The debate in the US also highlights the fact that statistics alone will not be able to say that the adjusted database is better as there will always be some 'loss function' at a low level of aggregation for which the unadjusted census is better. Belin and Rolph (1994) have the following quote from Citro and Cohen (1985):

> "It must be accepted that no adjustment procedure can be expected to simultaneously reduce error of all census information for every location."

As a consequence of this, Belin and Rolph (1994) argue for the need for consensus amongst the statisticians, politicians, and other census users regarding realistic goals for an adjusted database to achieve.

In his discussion of the 1991 Censuses of the UK Simpson (1994) further develops this theme of consensus. He argues that this consensus should be possible with careful and thorough planning and makes the following statement with respect to the 2001 Censuses of the UK.

> "The challenge is to gain widespread acceptance in advance of the next census for:
>
> - A *target level of accuracy* for estimates of non-response, for statistics for national and stated sub-national areas;

- A model to derive non-response estimates for smaller sub-populations;
- A timescale for publication of non-response estimates, that permits and better still requires their use in the main governmental applications of census data;

and to create the tools that can fulfil these targets."

The following chapters dealing with the design of a follow-up survey and subsequent estimation strategies represent a contribution towards achieving the challenge set by Simpson (1994) in the 2001 Census of the UK.

Appendix 2.1 – Re-expressing the Chapman estimator

The calculations here show how the Chapman estimator can be re-expressed as the standard DSE with a bias correction. Starting with the Chapman Estimator in the form of (2.9)

$$\hat{n}^{C}_{++} = \frac{(n_{1+}+1)\times(n_{+1}+1)}{n_{11}+1} - 1 = \frac{n_{1+}\times n_{+1} + n_{1+} + n_{+1} + 1 - n_{11} - 1}{n_{11}\left(1+\frac{1}{n_{11}}\right)}$$

This can now be expressed as a power series expansion ignoring second and higher order terms to give

$$\hat{n}^{C}_{++} \cong \frac{n_{1+}\times n_{+1} + n_{1+} + n_{+1} - n_{11}}{n_{11}}\left(1-\frac{1}{n_{11}}\right)$$

$$\hat{n}^{C}_{++} \cong \hat{n}_{++} + \frac{n_{1+}+n_{+1}-n_{11}}{n_{11}} - \frac{n_{1+}\times n_{+1}}{n_{11}^2} - \frac{n_{1+}+n_{+1}-n_{11}}{n_{11}^2}$$

Conditioning on the underlying multinomial model and the unknown population total n_{++} the Chapman estimator can be written as

$$\hat{n}^{C}_{++} = \hat{n}_{++} + \frac{n_{++}(p_{1+}+p_{+1}-p_{1+}\times p_{+1})}{n_{++}\times p_{1+}\times p_{+1}} - \frac{n_{++}^2\times p_{1+}\times p_{+1}}{n_{++}^2\times p_{1+}^2\times p_{+1}^2} - \frac{n_{++}(p_{1+}+p_{+1}-p_{1+}\times p_{+1})}{n_{++}^2\times p_{1+}^2\times p_{+1}^2}$$

Ignoring terms of $O(n_{++}^{-1})$ and smaller the above simplifies to

$$\hat{n}^{C}_{++} \cong \hat{n}_{++} + \frac{p_{1+}+p_{+1}-p_{1+}\times p_{+1}-1}{p_{1+}\times p_{+1}} = \hat{n}_{++} - \frac{(1-p_{1+})\times(1-p_{+1})}{p_{1+}\times p_{+1}}$$

which is the form of (2.10).

Chapter Three – Census Coverage Survey Design

3.1) Introduction

In response to the problems with underenumeration and its estimation for the 1991 Census, ONS initiated a census underenumeration research programme in 1996 to consider options available to be used with the 2001 Census. This was part of a wider programme considering all aspects of the conduct of the census, and innovations relating to the organisation, data collection, and processing of the 2001 Census can be found in the Government White Paper[1] presented to the UK Parliament. The obvious solution to census underenumeration is a complete census. However, the US experience with the 1990 Census demonstrates the impossibility of such an undertaking. Therefore, the goal in 2001 is to learn from 1991 by conducting a well-planned census but have in place the methods to estimate for any underenumeration. As called for by Simpson (1994), the aim was that these methods should be well researched, have general acceptance amongst the census user community, and involve a re-think of the approach used in any follow-up survey.

Early internal research within ONS focused on the need for some kind of independent follow-up survey to the census and the need to integrate estimation of census underenumeration into the census database. This built on the comments by Heady *et al* (1994) and the perceived quality and access problems associated with the use of administrative data in the UK. Users responded very positively to a discussion of these issues by Ian Diamond (University of Southampton) and Andy Teague (ONS) at the Royal Statistical Society Cathie Marsh Memorial Lecture in November 1997. They strongly expressed the desire for census underenumeration to be an integrated part of the census database. The response to this was the strategy for a One-Number Census (ONC) presented at the Leeds Conference for Census Users (May, 1998) and outlined in Brown *et al* (1999). The main aim of the strategy is to integrate estimates of census underenumeration, derived by combining the census with data from a follow-up survey and administrative sources, into the 2001 Census database.

[1] The Government White Paper entitled 'The 2001 Census of Population' was presented to Parliament in March 1999 by the Economic Secretary to the Treasury, the Secretary of State for Scotland, and the Secretary of State for Northern Ireland. Ref Cm 4253.

A key part of the strategy is the Census Coverage Survey (CCS), which is to be the main source of information on census underenumeration in 2001. In designing this survey, three important lessons learnt from the 1991 CVS need to be kept in mind:

i) the 1991 survey was not independent of the 1991 Census and, as a consequence of this, the methodology implicitly assumed that the survey would find everybody;

ii) the small sample size meant that estimates were only available for very large populations and geography was not preserved (Birmingham was in the same group as Newcastle-upon-Tyne and Durham was in the same group as Exeter);

iii) the combination of coverage and quality assessment meant a complex questionnaire needed to be used for the interviews.

In 1991 the third point was the main factor. The quality side of the survey required survey interviewers to know about the census forms for co-operating households to check answers, thereby compromising independence between the two counts. The US Census Bureau found, following the 1950 Census, that this approach would at the very least understate census underenumeration of individuals within counted households (Steinberg *et al*, 1962). One way to avoid this is the use of a third follow-up interview to reconcile differences between the survey and the census. The US Census Bureau used this approach in 1960 (Steinberg *et al*, 1962). However, the implications of a second survey are increased costs as well as an increased burden on the public. The use of a single survey in 1991 to assess coverage and quality also required a complex and time consuming interview using professional survey interviewers. Cost constraints consequently prevented the selection of a large sample size.

The combined approach of looking at census quality and underenumeration together had been successful in 1981, see Britton and Birch (1985), but the experience of 1991 suggested that in a climate of increased difficulty for the census to achieve a complete count, census underenumeration needs to be treated separately. This break immediately makes an independent re-enumeration possible. In addition, it is only necessary to collect the limited number of census variables, such as age, sex, ethnicity, and household tenure, which prior research has identified as being important for the modelling of census underenumeration. Therefore, simplifying the

questionnaire and data collection implies that, for a fixed cost, a larger sample size can be achieved.

Since the main aim of the survey is to check census coverage, its target population must be the same as that of the census, that is, all households and the individuals within them. In principle, this could be achieved by selecting a sample of households from a frame of all households. The postcode address file (PAF) is sometimes used to achieve this, an example being the current UK Labour Force Survey. However, as pointed out in Brown *et al* (1999), there are problems with using the PAF as a frame to check census coverage. The PAF is an electronic file that identifies all address points known to the post office. In many areas address points correspond to households. However, this is not the case in areas of multi-occupancy where one address point corresponds to an unknown number of households. There is also the travel cost considerations associated with drawing a sample of households that has no geographic clustering. The Labour Force Survey overcomes this by utilising telephone interviewing for those units sampled in a previous wave. Again, this is not an option for a survey of census coverage.

An alternative is to select small geographic areas and then re-enumerate all the households within the sampled area. This is the basis of the approach used in the US PES of 1990 and the planned approach for 2000. In the UK, one approach is to select a sample of postcodes[2]. Postcodes cover all private households and communal establishments in the UK and therefore constitute a possible sample frame. Such a frame naturally clusters households together and so sampling postcodes is cost efficient. In particular, interviewers will be given a map of each sampled postcode area and will then attempt to enumerate all households and individuals within the postcode who were usual residents on census night. The following sections in this chapter consider how best to select this sample of postcodes and expand the description of the design given in Brown *et al* (1999).

[2] A postcode is an identifier for a small collection of address points (on average 15) used by the Post Office to organise the delivery of mail in the UK.

3.2) The Basic Approach

Since the 1971 Censuses of the UK the mid-year population estimates for each local authority district (LAD) by age and sex have been based on census counts adjusted for underenumeration. The main aim of the CCS must be to deliver estimates of underenumeration by age and sex for each LAD to allow a similar adjustment to occur in 2001. This is of primary importance as the mid-year population estimates are used in the allocation of money from national to local government. Therefore, the aim of the CCS is to select a sample that can estimate the age-sex distribution for each LAD. In England and Wales there are about 400 LADs ranging in population size from less than 100,000 to over one million. (London consists of about 30 local authority districts, of varying population sizes and demographic make-up, that were grouped as Inner London and Outer London in 1991.) The ideal would be to select a sufficiently large sample to allow direct estimation for each LAD, but the sample size requirements to achieve that with the necessary precision would be prohibitive. Instead, the design aims to generate direct estimates of high precision for groups of contiguous LADs, called estimation areas, with approximately equal population sizes. This approach preserves geography and should deliver high quality estimates of the main variable of interest, the age-sex distribution, down to a low level of aggregation.

The problem can now be thought of as designing a survey such that the selected sample of postcodes will yield the age-sex distribution for each estimation area. Conceptually, the data available for designing this survey are the 1991 Census counts by age and sex for each postcode, represented by Z_{akelg} (age-sex group a, from postcode k, within 1991 enumeration district (ED) e, of LAD l and estimation area g) although, in practice accessing the data at the postcode level is difficult. These counts are used in the design as a proxy variable for Y_{akelg}, the true population count. In the subsequent analysis dropping the geographic subscripts (k for postcodes and e for EDs) will represent summing the counts across that geographic level. For example Z_{aelg} is the age-sex count summing across all the postcodes within ED e of LAD l and estimation area g. The estimation area can now be treated as a level of stratification and the aim is to draw an efficient sample of postcodes from within each estimation area stratum. In what follows a model-based approach to survey design will be taken,

with the same approach applied independently within each estimation area. (See Royall (1970) for a description and justification of the model-based approach.)

3.3) One-Stage Design

3.3.1) Postcode Level Model

A simple approach to the design problem would be to select a simple random sample of postcodes from within each estimation area. However, while simple to implement this would be a very inefficient strategy with little control over the types of postcodes within the selected sample. Such control is possible by allowing postcodes to be first stratified into different types (indexed d) and then by population size. A super-population model that reflects this structure is one where the postcode counts by age and sex satisfiy

$$\begin{aligned} &E\{Y_{kehd}\} = \mu_{hd} \\ &Var\{Y_{kehd}\} = \sigma^2_{hd} \\ &Cov\{Y_{kehd}, Y_{mfjc}\} = 0 \text{ for all } m \neq l, e \neq f, h \neq j, \text{ and } d \neq c \end{aligned} \qquad (3.1)$$

where the subscripts a, l, and g have been dropped as the same model applies independently to each age-sex group across all LADs within each estimation area. The model given by (3.1) implies that for a particular age-sex group, once the postcodes have been stratified by type and some measure of population size, the distribution of counts across the postcodes is generated by independent draws from a model with constant mean and variance. The best linear unbiased estimator of the total population for a particular age-sex group under (3.1) is

$$\hat{T} = \sum_{d=1}^{D} \sum_{h=1}^{H_d} \frac{N_{hd}}{n_{hd}} \sum_{k=1}^{n_{hd}} y_{kehd} \qquad (3.2)$$

where n_{hd} is the number of postcodes sampled from population size stratum h within type stratum d and N_{hd} is the corresponding population size. The prediction error variance for the estimator (3.2) is

$$\text{Var}\{\hat{T} - T\} = \sum_{d=1}^{D}\sum_{h=1}^{H_d} \frac{N_{hd}^2}{n_{hd}}\left(1 - \frac{n_{hd}}{N_{hd}}\right)\sigma_{hd}^2 \qquad (3.3)$$

This variance can be estimated from the sample by substituting an unbiased estimator for σ_{hd}^2 into (3.3).

The problem is now the efficient allocation of the sample to the pre-defined strata of postcodes. Under optimal allocation we need to minimise $\sum_{d=1}^{D}\sum_{h=1}^{H_d} \frac{N_{hd}^2}{n_{hd}}\sigma_{hd}^2$, the only term in (3.3) that depends on the sample size n_{hd}, subject to a fixed sample size constraint $n = \sum_{d=1}^{D}\sum_{h=1}^{H_d} n_{hd}$. There is a fixed sample size constraint assuming that cost constraints are not postcode specific but just define a total number of postcodes that can be sampled. This leads to an allocation of the sample given by

$$n_{hd} = n \times \frac{N_{hd}\sigma_{hd}}{\sum_{d=1}^{D}\sum_{h=1}^{H_d} N_{hd}\sigma_{hd}} \qquad (3.4)$$

By substituting (3.4) back in the variance (3.3) the problem is now simply the choice of the total sample size n. Therefore, assuming optimal allocation the variance formula given by (3.3) simplifies to

$$\text{Var}\{\hat{T} - T\} = \frac{1}{n}\left(\sum_{d=1}^{D}\sum_{h=1}^{H_d} N_{hd}\sigma_{hd}\right)^2 - \sum_{d=1}^{D}\sum_{h=1}^{H_d} N_{hd}\sigma_{hd}^2 \qquad (3.5)$$

The form of the variance given by (3.5) can now be used to specify a value for n, to satisfy a condition that the relative standard error (RSE) or coefficient of variation of the estimator equals a fixed percentage α. That is, we require

$$\alpha = \frac{\sqrt{\text{Var}\{\hat{T} - T\}}}{T} \times 100 \qquad (3.6)$$

and substituting the form of the variance given by (3.5) leads to a total sample size specified as

$$n = \frac{\left(\sum_{d=1}^{D}\sum_{h=1}^{H_d} N_{hd}\sigma_{hd}\right)^2}{\frac{\alpha^2 T^2}{100^2} + \sum_{d=1}^{D}\sum_{h=1}^{H_d} N_{hd}\sigma_{hd}^2} \qquad (3.7)$$

Often the $\sum_{d=1}^{D}\sum_{h=1}^{H_d} N_{hd}\sigma_{hd}^2$ term in (3.7) is dropped from the formula as this is of lower order compared to the term in T^2.

The problem with (3.7) is it requires knowledge of the population total and variance of the variable of interest, in this case Y_{akehd}, the postcode counts for each age-sex group. These are unknown and therefore, to get an approximate value for n, a design variable is used and in this case any one of the Z_{akehd}'s is a possible choice or some combination of them such as the total count for the postcode in the 1991 Census. This means that the sample size should satisfy the RSE constraint for the design variable and implies an 'expected' RSE of the same value will be achieved for the actual variables of interest.

3.3.2) ED Level Model

The design described in section 3.3.1 should be efficient in statistical terms but may actually be costly as there is no control within an estimation area of the geographical spread of the selected postcodes. Such a design will either involve the recruitment and training of a large number of interviewers who will each do very few interviews, or wasted time and travel costs while interviewers travel between postcodes. A highly clustered, but equally straightforward, alternative is to select EDs and then to re-enumerate all the postcodes within the selected ED. The problem is now the selection of a sample of EDs and this can be tackled using the same approach as (3.1) but specifying the model at the ED rather than postcode level. Such an approach has high cost advantages, as the postcodes that the interviewer has to cover are all close

together. An ED level approach also has data advantages, as the population counts for 1991 EDs are readily available. Conversely, the postcode level information from the 1991 Census is not easy to access, and there is the additional problem of changes to postcodes between 1991 and 2001.

3.4) Prototype Designs

To demonstrate the two approaches outlined in Section 3.3, 1991 Census data for a large LAD, with a 1991 Census population of around half a million individuals, is utilised. For this LAD, data are available at the postcode level for 1991 postcodes. For the purposes of comparing the two approaches, the assumption is that postcodes have not changed since 1991. A practical consideration is that in 1991 (and in 2001) some postcodes cross ED boundaries. One possible solution is to allocate postcodes to EDs based on the ED that contains the postcode centroid or alternatively the ED that contains the most households from the ED. However, for the purpose of this study it was considered unnecessary and postcodes split by EDs are treated as separate postcodes. The consequence of this is an increase in the number of postcodes within the LAD by approximately 2,000 and a decrease in the average number of households per postcode from 15.5 to 13. The RSE chosen is consistent with the estimated RSE for the estimated population total in 1981 with a national (England and Wales) level RSE of approximately 0.06%. Assuming the national estimate $\hat{T}$ is the sum of one hundred *similar* independent estimation area estimates, $\hat{T}_g$,

$$\alpha = \frac{\sqrt{\sum_{g=1}^{100} \mathrm{Var}\{\hat{T}_g - T_g\}}}{\sum_{g=1}^{100} T_g} \times 100 = 0.06\% \Rightarrow \alpha_g = \frac{\sqrt{\mathrm{Var}\{\hat{T}_g - T_g\}}}{T_g} \times 100 \cong 0.6\% \quad (3.8)$$

The approximation in (3.8) gives an RSE at the estimation area level. The remainder of this section considers the practical aspects of applying the two design models to these data.

3.4.1) National Hard to Count Index

In the design of the surveys that followed both the 1981 and 1991 Censuses, the EDs were graded based on their expected difficulty to count and this was used to over-sample areas that were expected to be hard to count. This is not necessarily efficient, especially if the graded EDs or postcodes are all at least as homogeneous as the non-graded EDs. The designs described in Section 3.3 introduce this idea of graded EDs through two levels of stratification, the first level being the type of ED. The aim is to stratify the EDs into different types based on how difficult they were to count in 1991. The US Census Bureau used a similar approach when designing the PES for the 2000 Census (see Hogan, 2000).

Standard indexes for classifying small population areas are not necessarily appropriate for this purpose as they concentrate on health and deprivation. While census underenumeration may be linked to areas of deprivation it is not by any means an exact mapping. What is needed is an index that utilises the variables associated with census underenumeration, such as high levels of multi-occupancy but not necessarily factors that create more work for the census enumerator such as a large geographic area. The work presented in Brown *et al* (1999) used a prototype hard to count (HtC) index, which is also used here. The index is constructed from a score calculated for all EDs in the 1991 Census. The prototype index ranks the EDs using each of the following variables:

- percentage of heads of household who experienced language difficulty;
- percentage of young people who migrated into the ED in the last year;
- percentage of imputed residents for the enumeration district;
- percentage of households in multiply-occupied buildings; and
- percentage of households which were privately rented;

assigns normal scores based on the ranks and then sums across the variables to get an overall score for the ED. This is split into quintiles to create a five level index. All postcodes within EDs are assigned the same HtC index as the ED.

Since this prototype index was proposed, extensive work has been done in England and Wales to create the 'best' index based on analysis of underenumeration in the

1999 Census Rehearsal and analysis of the 1991 Census (see Simpson *et al*, 1997). This work is summarised in ONS (2000a). Similar work has been done in Scotland and the approach for Northern Ireland is described in chapter five.

For the purposes of comparing design strategies it does not matter if the prototype differs from the index used by ONS in the final CCS design. This is because the same prototype is being used in all the designs being compared and in addition, it is not expected that the final index will be dramatically different.

3.4.2) Postcode Level Design

In applying the model given in (3.1) to the data, a second level of stratification based on some measure of population size is proposed to improve efficiency. Any one of the age-sex counts for the postcodes could be used but in an attempt to get some efficiency gain across all the age-sex groups, the total population is used as a design variable for the purposes of defining size strata. The Dalenius-Hodges rule for specifying stratum boundaries is used, see Cochran (1977), with additional boundaries formed for the largest and smallest postcodes in each level of the HtC index. Using (3.7) it is possible to calculate the required sample size, which is then allocated to the specified strata using (3.4). The final design and stratum allocations are given in Table 3.1.

The design in Table 3.1 specifies a sample of 531 postcodes, a sampling fraction of just less than four per cent. Sample design is not an exact science and the specification of the size strata involved some trial and error. In the final design, eight size strata are defined within each HtC category based on the Dalenius-Hodges rule. The rounding-up of sample sizes means that the expected RSE for the estimated 1991 population total is 0.59 per cent.

TABLE 3.1

One-stage postcode level design using 1991 total population for size stratification

HtC Index	Number of Size Strata	Number of Postcodes	Sample Size
1	8	2258	87
2	8	3379	121
3	8	3041	108
4	8	2881	109
5	8	2202	106
All	40	13761	531

The design in Table 3.1 uses the total population as the design variable to define strata, specify the sample size, and allocate that sample to the strata. However, the real interest is in how well the design in Table 3.1 might be expected to perform for estimating the population distribution by age and sex. Using five-year age groups for males and females, with the last category 85+, and collapsing together the age groups between 45 and 79 as there was little evidence of underenumeration across these ages in 1991 (see Heady *et al*, 1994), generates 24 age-sex groups. Applying the design in Table 3.1 to each of the age-sex groups gives a median RSE of 3.88 per cent across these age-sex groups, while the maximum RSE is 14.49 per cent for males in the oldest age group. (This high RSE is not necessarily a problem as this particular age-sex group represents less than 0.5 per cent of the population in the estimation area.) Therefore, while the stratification is efficient in terms of the total population, for the individual age-sex groups the achieved efficiency depends both on the population size of the particular group and more importantly, whether the distribution of the total population is a good proxy for the distribution of the particular age-sex group. This issue of choosing a design variable that is a good proxy across the age-sex groups is considered in section 3.5.2.

3.4.3) ED Level Design

To generate an ED design the same approach is used to stratification, but at the ED level rather than postcode level. As with the postcode model, the EDs are stratified by the HtC index and then size, defined as the total population of the ED in 1991. As

before, when allocating the sample all non-integer samples have been rounded-up to give the final sample sizes. The final design is given in Table 3.2 and specifies a sample of 109 EDs, a sampling fraction of over ten per cent. On average, this would translate into a sample of postcodes that would be three times higher than the sample of postcodes required for approximately the same degree of accuracy using a postcode level model. The rounding-up of sample sizes means that the expected RSE of the survey estimate of the total 1991 population is 0.56 per cent.

TABLE 3.2

One-stage ED level design using 1991 total population for size stratification

HtC Index	Number of EDs	Number of Size Strata	Sample Size (EDs)	Sample Size[a] (Postcodes)
1	144	6	16	249
2	210	6	22	354
3	186	6	22	356
4	193	6	24	358
5	197	6	25	261
All	930	36	109	1578

a. Based on the expected number of postcodes per ED within each stratum.

As with the postcode level design, the design in Table 3.2 uses the total population as the design variable to define strata, specify the sample size, and allocate that sample to the strata. However, it is of interest to see how well the design will perform across the age-sex distribution. Using the same 24 age-sex groups as before, and applying the design in Table 3.2 to each group, gives a median RSE across the age-sex groups of 3.16 per cent. As with the postcode model, the maximum RSE of 9.35 per cent is for males in the oldest age group. Therefore, while at the total population level the design in Table 3.2 requires a much larger sample of postcodes to achieve the same accuracy, the increased sample size improves estimates across the age-sex groups. This suggests that at the ED level, the distribution of the total population is a better proxy for the distributions of each age-sex group. This makes sense as, at the ED level there is generally a better spread across all the age-sex groups while at the postcode level the population often does not include all age-sex groups. Therefore, at the ED

level a high total population is likely to represent high counts across all the age-sex groups but this will be less true at the postcode level. The second advantage of the ED level model is based on the assumption that one interviewer can re-enumerate more than one postcode. This leads to cost and time advantages from using the ED level design, as each interviewer does not need to spend time or money travelling between their sample of postcodes. The third, as already mentioned, relates to the fact that population counts are readily available at the ED level.

3.5) Two-Stage Design

Clustering the selected postcodes together within EDs has cost advantages but this must be traded against a loss in efficiency evident by the difference in the estimated number of sampled postcodes compared to the design at the postcode level outlined in Table 3.1. However, the design at the ED level outlined in Table 3.2 is in effect over-clustered if all the postcodes within one ED are more than the workload for a single interviewer. Therefore, there is a natural compromise between the two extremes; a two-stage approach of sampling EDs and then a fixed sample of postcodes per ED that represents the workload for an interviewer. A super-population model that allows for postcode counts within EDs to be correlated but still uncorrelated between EDs can be used to represent such a design. Again, dropping the age-sex subscript, but applying the same model to each age-sex group gives

$$
\begin{aligned}
&E\{Y_{kehd}\} = \mu_{hd} \\
&Var\{Y_{kehd}\} = \sigma^2_{hd} \\
&Cov\{Y_{kehd}, Y_{mfjc}\} = \rho\sigma^2_{hd} \text{ for all } k \neq m, e = f, h = j, \text{and } d = c \\
&\qquad\qquad = 0 \text{ otherwise}
\end{aligned}
\tag{3.9}
$$

The model given by (3.9) represents a design that stratifies the postcode counts by HtC index and population size and allows for the selected postcodes being clustered within EDs. In (3.9) the within ED correlation is constant across all the strata but in practice it could vary. Under model (3.9) the optimal predictor of the within stratum population total (but for simplicity not including the stratum subscripts) is given by

$$E[T \mid S] = \sum_{e \in S} \sum_{k \in S_e} Y_{ke} + \sum_{e \in S} \sum_{k \in R_e} E[Y_{ke} \mid S_e] + \sum_{e \in R} \sum_{k \in e} E[Y_{ke}] \qquad (3.10)$$

where S represents the sample of EDs from within the particular stratum and S_e represents the sample of postcodes from within ED e while R and R_e represent the corresponding non-sampled EDs and postcodes. The overall total population is then given by summing across the strata. The first term in (3.10) is the sum of the sample postcode counts, the second term is predicting counts for non-sampled postcodes in sampled EDs, and the third term predicts counts for the rest of the non-sampled postcodes from non-sampled EDs. From (3.9) the expectation in the third term is μ while the expectation in the second term can be modelled as

$$E[Y_{ke} \mid S_e] = (1 - \alpha_e)\mu + \alpha_e \bar{y}_{S_e} \qquad (3.11)$$

which is a weighted average of the mean for the postcodes in the population as a whole and the sample average for the observed data from within the particular ED. Under an assumption of normality for the distribution of the postcode counts it is possible to derive 'optimal' weights, but a 'safe' alternative is to set α_e equal to one and use sample data to estimate the mean within the sampled EDs. The within stratum predictor in (3.10) can now be written as

$$E[T \mid S] = \sum_{e \in S} m_e \bar{y}_{S_e} + \sum_{e \in S} (M_e - m_e)\bar{y}_{S_e} + \mu \left[N - \sum_{e \in S} M_e \right] \qquad (3.12)$$

where M_e is the total number of postcodes in ED e, m_e is the number sampled from ED e, and N is the total number in the population. From (3.12), a linear unbiased estimator follows by plugging-in a linear unbiased estimator for the model parameter μ, defined in general terms as $\sum_{e \in S} w_e \bar{y}_{S_e}$, to give

$$\hat{T} = \sum_{e \in S} m_e \bar{y}_{S_e} + \sum_{e \in S} (M_e - m_e)\bar{y}_{S_e} + \sum_{e \in S} w_e \bar{y}_{S_e} \left[N - \sum_{f \in S} M_f \right] \qquad (3.13)$$

The form of the estimator (3.13) has intuitive appeal. The first term is the sum of the sampled postcodes, the second term estimates for the non-sampled postcodes within a sampled ED using data from the sampled postcodes within the ED, while the third term estimates for the postcodes from non-sampled EDs using all the sample data.

To estimate the variance of (3.13) it is easier to re-write the estimator as

$$\hat{T} = \sum_{e \in S} m_e \bar{y}_{S_e} + \sum_{e \in S} m_e u_e \bar{y}_{S_e} \qquad u_e = \frac{1}{m_e}\left[(M_e - m_e) + w_e\{N - \sum_{f \in S} M_f\}\right] \qquad (3.14)$$

so that the first term in (3.14) is the sum over the sampled postcodes and the second term is a weighted sum of the sampled postcodes that estimates for the non-sampled postcodes. After some simplification, see Appendix 3.1 for the details, the variance of (3.14) is given by

$$\mathrm{Var}[\hat{T} - T] = \left[\begin{array}{l} \sum_{e \in S}\left\{(1-\rho)\left(m_e u_e^2 + (M_e - m_e)\right) + \rho\left(m_e u_e - (M_e - m_e)\right)^2\right\} \\ + \sum_{e \in R} M_e(1 - \rho - \rho M_e) \end{array}\right]\sigma^2 \qquad (3.15)$$

The estimator of total given by (3.13) and its variance given by (3.15) can be further simplified when a constant sample size is used for the within ED sample, in other words m_e equals m across the sampled EDs. In particular, $\sum_{e \in S} w_e \bar{y}_{S_e}$, the sample based linear estimator for the model parameter μ, is then just the equally weighted average of the sample means $\bar{y}_{S_e}$ as the sample of postcodes from each sampled ED contains the same amount of information about the overall mean μ. The option of a constant second stage sample also has attractions in terms of survey management, as it tends to make the allocation of interviewer workloads more straightforward.

3.5.1) Comparison of Different Postcode Selections

Optimal design based on (3.9) is a complex process as it depends on the correlation of postcode counts within EDs, as well as the distribution of the number of postcodes per

ED, and will typically lead to variable sample sizes of postcodes within EDs. In addition, as stated earlier, postcode counts from the 1991 Census are not readily available across the country. However, the approach to design specified by (3.9) allows for stratification of the sample of postcodes using ED information but then spreading the postcode sample across more EDs than the design in Table 3.2. While an 'optimal' solution may not be possible, the efficiency of this approach can be assessed for different values of the within ED postcode sample and different levels of correlation making use of the same data as analysed in section 3.4.

To apply the model given by (3.9) the EDs were again stratified by the HtC index and then by the size of the ED population. However, as the model specifies homogeneity of postcode counts within strata, a third stratification using number of postcodes within the ED was also added to distinguish between EDs that have the same total population with quite different numbers of postcodes. The final design is given in Table 3.3. To allocate the sample, optimal allocation using the ED total population specifies the within stratum ED sample, as with the ED level design. The final postcode sample is then specified by selecting a fixed number of postcodes per selected ED. Table 3.3 gives the postcode samples under the assumption of three postcodes sampled per ED, four postcodes sampled per ED, and five postcodes sampled per ED.

The samples specified in Table 3.3 are the same size, in terms of the total postcode sample, as the design in Table 3.2. The difference is the number of EDs, which vary from 526 EDs for three postcodes down to 316 EDs for five postcodes, compared to 109 EDs in Table 3.2. The efficiency of the approach used in Table 3.3 will depend the level of the intra ED correlation of the postcode counts and the choice of the number of postcodes per ED. To assess this, RSEs were calculated using the variance given by (3.15) for different levels of correlation and the three sample sizes in Table 3.3. (The variance also depends on exactly what EDs are selected so the calculations were based on applying the three possible designs in Table 3.3 to the data and in each case selecting a specific sample of EDs.) The results are presented in Figure 3.1 for both the RSE for the estimate of the total population, and the median RSE across the individual age-sex groups.

TABLE 3.3

Two-stage design using ED total population in 1991 and number of postcodes per ED for size stratification

HtC Index	Number of 'Size' Strata	Number of Sampled Postcodes: Three Postcodes Per ED	Four Postcodes Per ED	Five Postcodes Per ED
1	24	207	228	220
2	24	342	340	345
3	24	345	328	340
4	24	330	320	325
5	24	354	364	350
All	120	1578	1580	1580

FIGURE 3.1: Comparison of the performance of different sized within ED samples for a fixed total postcode sample and a fixed ED stratification at both the total population and across the age-sex groups.

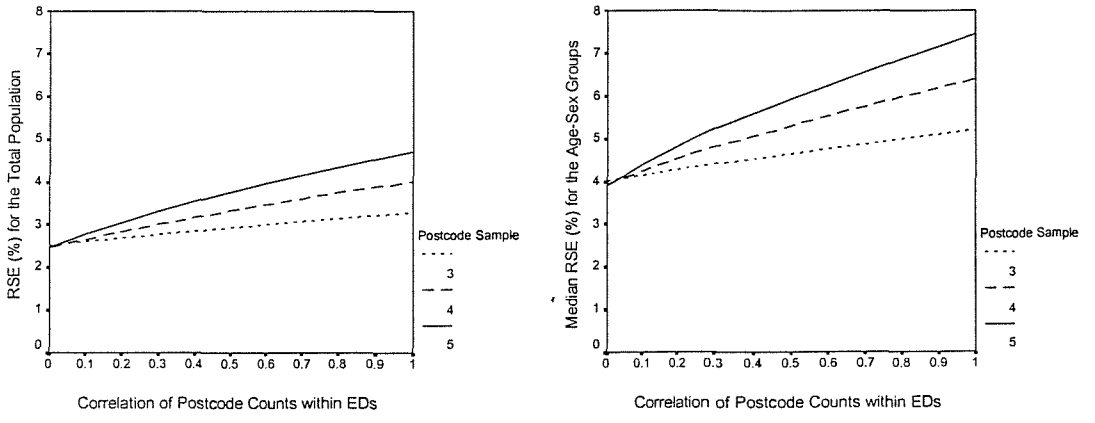


Figure 3.1 demonstrates two important points. Firstly, provided the level of correlation is low or moderate, there is little to choose in efficiency terms between the three different within ED postcode samples. In the data used for this example, once age and sex are controlled for, there is effectively zero correlation of postcode counts within EDs. This was confirmed by fitting a two level random intercepts model, postcode counts by age and sex clustered within EDs, and estimating the intra-cluster correlation. While the correlation may not be zero everywhere this suggests that the assumption of low correlation is reasonable. Therefore, this fact leads to the choice of

the within ED postcode sample being made on management grounds and early fieldwork tests conducted by ONS in Brent, following the 1997 Census Test, suggested that on average five postcodes per ED would be an acceptable workload for an interviewer.

The second point shown by Figure 3.1 is the fact that across the age-sex groups, for a low level of correlation, the median RSE is in line with the result from the ED level model where the same sized postcode sample generated a median of 3.16 per cent. However, at the total population level the RSE is significantly higher, around 2.5 per cent in Figure 3.1 compared to less than 0.6 per cent for the ED level model. This is a result of the fact that the model given by (3.9) assumes efficient stratification at the postcode level, while what has been achieved by the design in Table 3.3 is a proxy for this using ED level information. This contrasts with the ED level model where the stratification is very efficient for estimating the total population as it uses the total population to define the size strata in the design.

The question then is why use the two-stage approach? One potential weakness of the ED level design is that it concentrates the sample in a relatively small number of EDs based on out-of-date information while the two-stage design spreads the sample over many more EDs. It is intuitive that this second approach of spreading the sample across more EDs will be more robust to changes in the population between the censuses compared to the sample that concentrates the sample of postcodes within fewer EDs. The two-stage approach also has the appeal of covering more of the population in geographic terms, again a property that has intuitive appeal.

3.5.2) Multivariate Stratification

The aim of the designs outlined in the previous sections is to facilitate efficient estimation across the age-sex distribution. A key component of this has been stratification and so far the approach has been to use the total population of the ED or postcode as a design variable or proxy for the 24 possible variables generated by the age-sex groups. An alternative approach based on multivariate methods is outlined in Brown *et al* (1999). The aim is to construct a design variable and set of strata within the HtC index that better reflect the variability across the age-sex distribution than

using total population. This has several stages. The first stage uses principal component analysis to summarise the age-sex counts. Taking the first three components, cluster analysis using Ward linkage (SAS Institute, 1990) is applied to form strata that minimise the within stratum variance. Using more components makes the cluster analysis unstable. A proxy variable W_e on which to base the design is then constructed for each ED from the first three principal components using

$$W_e = \frac{|\Omega| \sum_{j=1}^{3} P_{je}}{\sqrt{\sum_{j=1}^{3} \mathrm{Var}(P_{je})}} \qquad (3.16)$$

where P_{je} represents the j^{th} principal component score for the e^{th} ED and Ω is the variance-covariance matrix of the original age-sex counts calculated across all the EDs. The motivation for (3.16) comes from the desire to construct a design variable that has a variance similar to the original data. Assuming that the determinant of the variance-covariance represents a summary of that original variability, (3.16) achieves this as the variance of W_e is $|\Omega|^2$.

Initial work found that using all 24 age-sex counts in (3.16) was rather cumbersome and therefore the work in Brown *et al* (1999) simply concentrated on six age-sex groups; males aged 0 to 4, females aged 0 to 4, males aged 20 to 24, males aged 25 to 29, males aged 30 to 34, and females aged 85 and over. The age-sex groups were chosen based on the identification of them as being highly associated with underenumeration in the 1991 Census (see Heady *et al*, 1994). In addition, an initial stage selects a small number of EDs that due to the size of one, or several, of the six counts is a 'large' outlier and would therefore unduly influence the principal component analysis and subsequent cluster analysis. The same approach is taken here in applying multivariate stratification to the data used for the three previous designs. Only two EDs were identified as outliers at the initial stage, with both allocated to 'take all' stratum. The remaining sample was allocated using multivariate stratification and optimal allocation based on a design variable calculated using

(3.16). The within ED sample of postcodes was set at five postcodes per sampled ED. The resulting design is given in Table 3.4.

TABLE 3.4

Two-stage design using a multivariate approach for size stratification

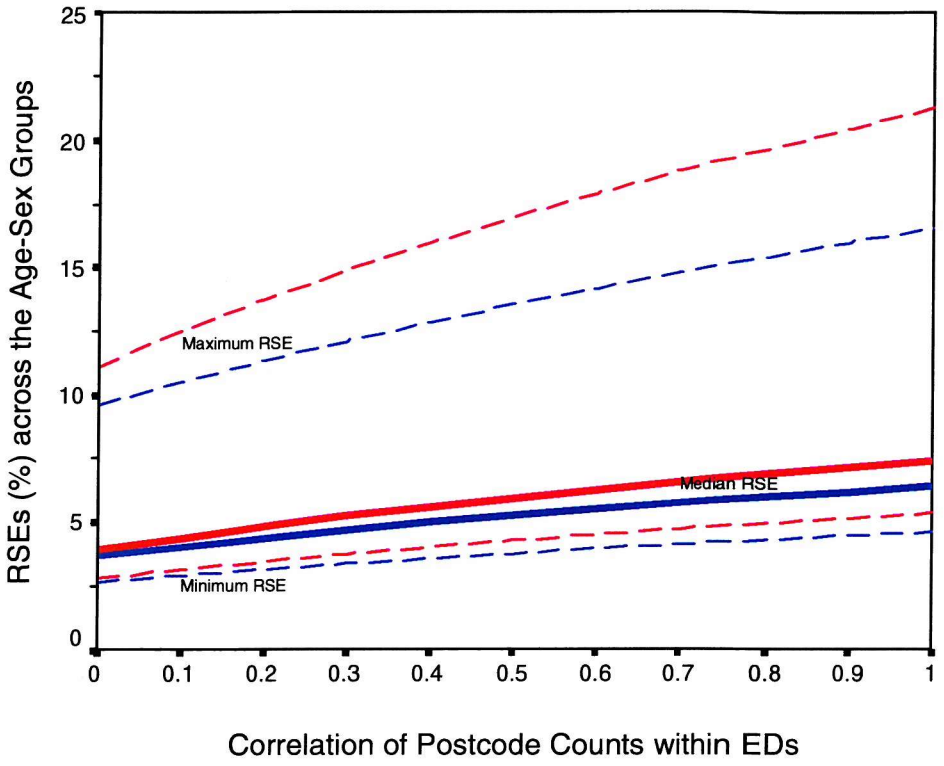
HtC Index	Number of EDs	Number of 'Size' Strata	Sample Size (EDs)
1	144	10	49
2	210	16	66
3	186	12	64
4	193	16	69
5	197	16	68
All	930	70	316

The design represents the same total postcode sample (1580) as the design in Table 3.3 assuming a sample of five postcodes per ED. As with the design in Table 3.3, it is possible to assess the efficiency of this approach in terms of the RSE, and compare it with the previous approach, by selecting a sample and calculating the RSE using the variance formula given by (3.15) for different levels of correlation of postcode counts within EDs. For estimating the total population, the two approaches are very similar with an RSE of 2.46 per cent for the design in Table 3.3 compared to 2.44 per cent for the design in Table 3.4, assuming the correlation of postcode counts within EDs is zero. Both increase as the correlation increases but the gap between the two also widens so that at a correlation of one the gain from using multivariate stratification is 0.5 per cent. Figure 3.2 compares the distribution of the RSEs generated by the age-sex groups for the two designs.

Figure 3.2 more clearly demonstrates an advantage from using the multivariate approach. The minimum RSE generated by the age-sex groups is slightly lower and, as with the RSE for the total population, the gap increases as the correlation increases. The same is true for the median RSE. This drops from 3.9 per cent using the design in Table 3.3 to 3.7 per cent using multivariate stratification with a correlation of zero. The biggest gain is in reducing the maximum RSE, generated for the males aged 85

and over, from 11.1 per cent to 9.6 per cent. Therefore, the multivariate stratification proposed in Brown *et al* (1999) achieves its aim as the resulting sample performs better across the age-sex groups both in terms of the median RSE and the spread of the RSEs.

FIGURE 3.2: Comparison of the performance across the age-sex groups of ED stratification by total population and number of postcodes (*red*) with multivariate ED stratification (*blue*) for a within ED sample of five postcodes.



3.6) Conclusions

The work presented here has described the CCS design and shows that the proposed design strategy, while not 'optimal', is efficient in the sense that it best makes use of the available data, is simple in terms of the management of interviewer workloads, as well as having cost advantages. This conclusion is based on a comparison with other approaches to the design. The advantage of the two-stage approach is its ability to use the readily available ED data while still generating a sample of postcodes that is spread across the estimation area. The analysis in section 3.5 additionally

demonstrates that the application of multivariate techniques for stratification of the population leads to a design that captures the variability across the age-sex distribution better than a design based solely on the total population.

The final design in Table 3.4 achieves comparable performance across the age-sex groups to the ED level design in Table 3.2 for the same sized sample of postcodes assuming a sensible level of correlation between postcode counts within EDs. However, it has the already stated advantage of spreading the sample over a wider geographic area. The most efficient approach in terms of sample size is the postcode level design in Table 3.1. The postcode sample for the design in Table 3.4 is three times the size of the postcode level design. However, as the final design clusters the sample of postcodes within EDs to form interviewer workloads this would suggest that on average the strategy outlined in section 3.5.2 will be cheaper than the postcode level sample for comparable RSEs across the age-sex distribution. It also does not have the data problems associated with designing at the postcode level.

The final issue that needs consideration is the formation of the estimation areas. The data in this chapter have assumed an estimation area of approximately 0.5 million people which in the case of the data used here represents a single large LAD in 1991. Ideally, the CCS would treat each local authority district as its own estimation area as the age-sex distribution is required for each local authority district adjusted for census underenumeration. Unfortunately, the national sample size to support such a design with direct estimates of sufficient quality would be prohibitive. However, the estimation areas do want to be as small as possible so that it is not necessary to group large numbers of local authority districts together. The simulations in ONS (1998a) suggest that 0.5 million is a good compromise between the two extremes, a decision endorsed by the One Number Census Steering Committee. ONS has now formed the estimation areas for England and Wales and the final groupings are presented in ONS (2000b). Similar work has been undertaken in Scotland and the formation of estimation areas for Northern Ireland is covered in chapter five of this thesis.

As a final comment, the sample size of 1,580 postcodes used in this chapter is based on achieving an RSE that would be approximately in line with the achieved accuracy in 1981. If this were interpolated to a sample size at the national level, the design in

Table 3.4 would represent a sample of 170,000 postcodes (over 2.5 million households) for England and Wales, a rather daunting proposition. The key point is that the design strategy, due to a lack of suitable information from 1991, cannot take account of efficiency gains from using auxiliary information at the estimation stage. Work presented in ONS (1998a), using an extensive simulation study, demonstrates that an RSE of approximately 0.05 per cent is achievable for the national level population total with a sample of approximately 20,000 postcodes or 300,000 households. The efficient use of auxiliary information at the estimation stage is considered in detail in the following chapter.

Appendix 3.1 – The variance of the estimated population total assuming a clustered super-population model

For the model given by (3.9) the population total can be estimated as

$$\hat{T} = \sum_{e \in S} m_e \bar{y}_{S_e} + \sum_{e \in S} m_e u_e \bar{y}_{S_e} \qquad u_e = \frac{1}{m_e}\left[(M_e - m_e) + w_e \{N - \sum_{f \in S} M_f\} \right]$$

and therefore the variance can be written as

$$\mathrm{Var}(\hat{T} - T) = \mathrm{Var}\left[\sum_{e \in S} m_e \bar{y}_{S_e} + \sum_{e \in S} m_e u_e \bar{y}_{S_e} - \sum_{e \in S} \sum_{k \in S_e} Y_{ke} - \sum_{e \in S} \sum_{k \in R_e} Y_{ke} - \sum_{e \in R} \sum_{k \in R_e} Y_{ke} \right]$$

$$\mathrm{Var}(\hat{T} - T) = \mathrm{Var}\left[\sum_{e \in S} m_e u_e \bar{y}_{S_e} - \sum_{e \in S} \sum_{k \in R_e} Y_{ke} - \sum_{e \in R} \sum_{k \in R_e} Y_{ke} \right]$$

Remembering that postcodes in sampled EDs are independent of postcodes in non-sampled EDs due to independence between EDs gives

$$\mathrm{Var}(\hat{T} - T) = \mathrm{Var}\left[\sum_{e \in S} m_e u_e \bar{y}_{S_e} - \sum_{e \in S} (M_e - m_e) \bar{y}_{R_e} \right] + \mathrm{Var}\left[\sum_{e \in R} \sum_{k \in R_e} Y_{ke} \right]$$

Again, remembering that EDs are independent but the postcodes within the same ED are not

$$\mathrm{Var}(\hat{T} - T) = \sum_{e \in S} m_e^2 u_e^2 \, \mathrm{Var}(\bar{y}_{S_e}) + \sum_{e \in S} (M_e - m_e)^2 \, \mathrm{Var}(\bar{y}_{R_e}) - 2 \sum_{e \in S} m_e u_e (M_e - m_e) \, \mathrm{Cov}(\bar{y}_{S_e}, \bar{y}_{R_e}) + \sum_{e \in R} M_e^2 \, \mathrm{Var}(\bar{y}_e)$$

Using the model each quantity can now be expressed in terms of model parameters so that

$$\mathrm{Var}(\bar{y}_e) = \frac{(1 - \rho + \rho M_e)}{M_e} \sigma^2$$

$$\mathrm{Var}(\bar{y}_{S_e}) = \frac{(1 - \rho + \rho m_e)}{m_e} \sigma^2 \qquad \mathrm{Var}(\bar{y}_{R_e}) = \frac{(1 - \rho + \rho (M_e - m_e))}{(M_e - m_e)} \sigma^2$$

$$\mathrm{Cov}(\bar{y}_{S_e}, \bar{y}_{R_e}) = \rho \sigma^2$$

Substituting back into the variance formula gives

$$\mathrm{Var}(\hat{T} - T) = \sigma^2 \left[\begin{array}{l} \sum_{e \in S} (1-\rho)\left(m_e u_e^2 + (M_e - m_e)\right) \\ + \sum_{e \in S} \rho\left(m_e^2 u_e^2 + (M_e - m_e)^2 - 2 m_e u_e (M_e - m_e)\right) \\ + \sum_{e \in R} M_e (1 - \rho + \rho M_e) \end{array} \right]$$

The variance can be further simplified to give

$$\mathrm{Var}[\hat{T} - T] = \left[\begin{array}{l} \sum_{e \in S} \left\{(1-\rho)\left(m_e u_e^2 + (M_e - m_e)\right) + \rho\left(m_e u_e - (M_e - m_e)\right)^2\right\} \\ + \sum_{e \in R} M_e (1 - \rho - \rho M_e) \end{array} \right] \sigma^2$$

which is the form given in (3.15).

Chapter Four – Estimation of Census Underenumeration for Estimation Areas

4.1) Introduction

The previous chapter considered the design of a Census Coverage Survey (CCS) and the expectation for the 2001 Census is that the CCS will be the main source of data on census underenumeration and provide the necessary data for creating a One-Number Census (ONC). A key component in that creation is a strategy for estimating the age-sex distribution of the population within each estimation area that utilises data from both the census and the CCS. It is also essential that estimates of underenumeration be available for individual LADs by age and sex to facilitate the adjustment of the 2001 Census as a basis for the mid-year population estimates. By necessity, the CCS design is based on data from the 1991 Census. However, when CCS estimates are produced, the available data will include the 2001 Census counts for all postcodes as well as the CCS counts for the sampled postcodes. The purpose of this chapter is to develop an age-sex specific estimation strategy for the CCS that makes efficient use of this information. Consequently, in what follows it is assumed that; the 2001 Census has been carried out in all postcodes and the data are available at the postcode level; the CCS has been carried out independently of the 2001 Census in a sample of postcodes as per the design in chapter three; and the ONC matching strategy, outlined in ONS (1998b) and ONS (2000c), has successfully matched the two data sources.

The US Census Bureau approach to estimation, as outlined in Hogan (1993), is to construct post-strata based on age, sex, race, tenure, and geography within which it is assumed that the multinomial model outlined in section 2.4 is an appropriate representation of the relationship between the census and the follow-up survey counts. Dual-system estimation, as outlined in Wolter (1986), is then used to estimate the population total within each post-stratum. In other words, the US Census Bureau use a single dual-system estimator (DSE) per post-stratum, with estimates for the population quantities required for calculating that DSE computed from the survey data.

The proposed approach to the estimation problem in the context of the CCS and the 2001 Censuses of the UK is slightly different. If the CCS was a 'perfect' survey in the sense that within the sampled postcodes the CCS obtained complete coverage, there are standard estimation techniques that would allow the estimation of the population total from the data collected in the sampled postcodes. These techniques would utilise the 2001 Census data in the non-sampled postcodes as an auxiliary variable to improve precision. Examples of such techniques are ratio and regression estimation, and the application of these methods would produce a set of age-sex estimates for the estimation area. However, the reality is that the CCS will also miss people and the issue then becomes one of dealing with non-perfect CCS counts. Under the assumptions required for dual-system estimation, a DSE can be calculated in each sampled area to 'correct' the CCS count for underenumeration. This leads to a set of counts for the CCS sampled areas, corrected for underenumeration, and standard estimation techniques can then be applied to the corrected counts to estimate the true population total and its distribution by age and sex. In Brown *et al* (1999) dual-system estimation was combined with regression estimation techniques to achieve this goal.

Brown *et al* (1999) used a simulation model to test the robustness of this estimation strategy to departures from the assumptions underpinning dual-system estimation. In particular, the issue of dependence between the census and CCS is considered. Other alternatives to correcting the CCS count, such as just combining the census and CCS and assuming no individuals are missed by both, are also compared to dual-system estimation. The work in this chapter expands Brown *et al* (1999). In particular, the use of ratio estimation is developed as an alternative to regression estimation. A refined version of the simulation model used in Brown *et al* (1999) is developed to compare the performance of several alternative strategies. A robust alternative to the standard ratio estimator is also developed and tested using the same simulation model. Finally, variance estimation for the proposed estimation strategy is also considered.

4.2) Population Estimation Using the 2001 Census Coverage Survey

As already stated, the work in this chapter assumes that the CCS has been designed as outlined in chapter three. It is further assumed that both the census and CCS have been carried out and that the two databases have been successfully matched using computer and computer-assisted probability matching. It follows that, for individuals counted in postcodes from the CCS sample it is possible to determine whether they were counted in both the census and the CCS, the census only, or the CCS only. There will also be some individuals who are missed by both the census and CCS. The estimation strategy then consists of two parts. The first is to estimate the true population in the CCS postcodes; the second part then builds on the first part to produce an estimate for the whole estimation area. Section 4.2.1 considers the first part and Section 4.2.2 considers different approaches to the second part.

4.2.1) Estimation Within the CCS Postcodes

Dual-system estimation was reviewed as a method of estimation for an unknown population total in section 2.4, particularly its use in the estimation of census underenumeration. The dual-system estimator is defined as

$$\hat{n}_{++} = \frac{n_{1+} \times n_{+1}}{n_{11}} \qquad (4.1)$$

where in this application n_{1+} is the total number of people counted by the census, n_{+1} is the total number counted in the CCS, and n_{11} is the total number counted in both. As pointed out in chapter two, for dual-system estimation to be applicable the following assumptions need to be plausible.

a) The DSE assumes that in the target population the matched CCS and census counts follow a multinomial distribution. That is, the probabilities of being counted by either or both the CCS and the census are **homogeneous** across the population of interest. This is unlikely for most populations.

b) Approximately unbiased estimation requires statistical **independence** between the census count and the CCS count. While impossible to guarantee this can be approximated.

For estimation following the 1990 Census in the US, assumption a) was approximated by forming post-strata defined by variables thought to be associated with census underenumeration. Assumption b) was approximated by carefully planning the follow-up survey to be independent of the census. See Hogan (1993) and the discussion in chapter two. The formation of the post-strata is crucial as the population parameter estimated under this strategy is the value that would be taken by a dual-system estimator (DSE) if the post-strata had been completely enumerated in their PES. This is achieved by replacing the resulting population quantities n_{+1} and n_{11} in (4.1) with survey estimates. A failure of the homogeneity assumption within the post-strata will therefore lead to a biased estimate of the population total. Unlike the bias of the DSE under the multinomial model discussed in section 2.4, which is unimportant as the population size increases, the bias due to heterogeneity can be considerable and does not decrease as the population size increases. In fact it can be argued that as the population size increases the likelihood of heterogeneity bias also increases.

The motivation of the work by Alho (1990) is the desire to get away from broad post-strata where heterogeneity can cause assumption (a) above to fail. This is achieved using a logistic model. Underlying the approach is the concept that the joint census / PES response of each individual can be modelled as a multinomial outcome, with response probabilities that depend on the characteristics of the individual. However, as noted in chapter two, this approach has yet to be applied on a large scale. The strategy for the UK takes a slightly different approach but the ethos is the same as in Alho (1990), in the sense that the aim is to not use dual-system estimation for large post-strata. Instead, dual-system estimation is thought of as a way of adjusting the sample counts generated by the CCS to account for those missed by the CCS. These adjusted counts are then treated as 'observed' sample data and used to estimate the total population. Assumption a) is approximated by splitting the population into groups by age and sex within the sampled postcodes of each HtC category, resulting in the DSE being calculated at a very low level of aggregation. Along with operational

independence, this also helps ensure that assumption b) is well approximated. In addition, work presented in Brown *et al* (1999) shows that some dependence between the census and the CCS has only a limited effect on the overall approach. This issue will be considered in more detail in subsequent sections of this chapter.

As the proposal is to carry out dual-system estimation at a low level of aggregation, the standard DSE defined by (4.1) is corrected for its small sample bias to give

$$\hat{n}_{++} = \frac{(n_{1+}+1)\times(n_{+1}+1)}{(n_{11}+1)} - 1 \qquad (4.2)$$

The estimator defined in (4.2) was proposed by Chapman (1951) for use in wildlife populations and its use is discussed by Seber (1982). It is discussed in more detail in section 2.4 but the key point is the correction gives an exactly unbiased estimator provided $n_{1+} + n_{+1}$ is greater than n_{++}, which is a reasonable assumption in most situations.

4.2.2) Models for Population Estimation Using the CCS

After the CCS has been carried out, there will be two population counts for each postcode in the CCS sample. One approach would be to assume that the CCS count is equal to the population count in the sampled postcodes and that, therefore, there is no underenumeration in the CCS. However, it is more sensible to assume that there will be underenumeration in both census and CCS and hence for each sampled postcode the two counts generated by these sources will contain non-response. Under the assumptions of homogeneity and independence previously outlined (4.2), the DSE with Chapman correction can be used to estimate the true population counts, Y_{aked}, for age-sex group a in postcode k from enumeration district e in HtC stratum d. The problem is then how to estimate the overall population total in the estimation area, T_a, for age-sex group a using this information together with the corresponding 2001 Census counts defined as X_{aked}.

4.2.2.1) Simple Approach

The simplest way to construct population estimates is to assume that information is only available for the census and the CCS in the sample areas. In this situation Alho (1994) proposes an adaptation of the Horvitz-Thompson estimator (Horvitz and Thompson, 1952) for T_a defined by

$$\hat{T}_a = \sum_{k,e \in CCS} \hat{Y}_{aked} / \pi_{ked} \qquad (4.3)$$

where π_{ked} is the probability of inclusion in the sample for postcode k from enumeration district e of HtC category d and $\hat{Y}_{aked}$ is the corresponding DSE estimate with Chapman correction for age-sex group a. Note that (4.3) involves calculation of the DSE at postcode level for each age-sex group. Since sample sizes for this can be extremely small, an alternative is to use the clustered nature of the sample, five postcodes per enumeration district (ED), and to compute a single DSE for the cluster of five postcodes giving

$$\hat{T}_a = \sum_{e \in CCS} \tilde{Y}_{aed} / \theta_{ed} \qquad (4.4)$$

where θ_{ed} is the probability of inclusion in the sample for the cluster of postcodes selected from enumeration district e of HtC category d and $\tilde{Y}_{aed}$ is the corresponding DSE estimate for age-sex group a. (It is important to note that the DSE $\tilde{Y}_{aed}$ only estimates the total population in the cluster of sampled postcodes, since sample weights are not introduced into the calculation of the DSE.) Within the CCS design, the postcode sample from within the sampled enumeration districts is a simple random sample and so π_{ked} does not depend on k and equals θ_{ed}. Therefore, the difference between (4.3) and (4.4) is whether you apply the DSE at the postcode level and then sum across the five postcodes sampled within the ED or alternatively sum across the postcodes and then apply the DSE. In general, (4.3) and (4.4) will not lead to the same estimate unless Y_{aked} is known without error. One would expect (4.4) to

be more 'stable' than (4.3) due to the use of larger counts in the DSE. However, it will be more susceptible to 'correlation bias' caused by the violation of assumptions a) and b).

4.2.2.2) Ratio Model for Population Estimation

In reality, census counts are available for all postcodes and can be used as auxiliary information to improve on the Horvitz-Thompson estimator. The simplest way to introduce these auxiliary data is to assume that the true count is approximately proportional to the census count. This leads to the classical ratio model for each age-sex group. Dropping the age-sex group indicator a, and representing the census count in postcode k from ED e of HtC stratum d by X_{ked}, this model can be written as

$$\begin{aligned} &E\{Y_{ked} \mid X_{ked}\} = R_d X_{ked} \\ &Var\{Y_{ked} \mid X_{ked}\} = \sigma_d^2 X_{ked} \\ &Cov\{Y_{ked}, Y_{jfg} \mid X_{jed}, X_{jfg}\} = 0 \text{ for all } k \neq j \end{aligned} \qquad (4.5)$$

where R_d and σ_d^2 are unknown model parameters to be estimated from the data. Under (4.5) it is straightforward to show (Royall, 1970) that the best linear unbiased estimator for the true population total T of an age-sex group is the stratified ratio estimator

$$\hat{T}_{RAT} = \sum_{d=1}^{5} \hat{R}_d \sum_{e=1}^{N_d} \sum_{k=1}^{M_e} X_{ked} \qquad (4.6)$$

where N_d is the total number of enumeration districts in HtC stratum d, M_e is the total number of postcodes in ED e, and $\hat{R}_d$ is the least squares estimate of the population ratio of true counts to census counts. Strictly speaking the assumption in (4.5) of zero covariance between postcodes counts is violated, as the design of the CCS has postcodes clustered within enumeration districts. However, this is not a serious problem for estimation of the population total, as (4.6) remains unbiased when this assumption is violated with only a small loss of efficiency (Scott and Holt, 1982).

Typically $\hat{R}_d = \frac{\sum_{e=1}^{n_d}\sum_{k=1}^{5} Y_{ked}}{\sum_{e=1}^{n_d}\sum_{k=1}^{5} X_{ked}}$ where n_d is the number of enumeration districts sampled in HtC index d and there are five postcodes sampled from enumeration district e. This is the best linear unbiased estimator of the ratio when (4.5) is a true representation of the population. However, (4.5) ignores the differential sampling from the size strata in the design, as well as the clustering of postcodes within EDs. Therefore, an alternative estimator of the ratio includes the sample weights to account for this. The estimator including sample weights would not be as efficient when (4.5) holds but may be more robust to model failure.

In practice, of course, the Y_{ked} are unknown and replaced by their corresponding DSEs within each postcode, given by $\hat{Y}_{ked}$, and this leads to (4.6) being expressed as

$$\hat{T}_{RAT} = \sum_{d=1}^{5} \frac{\sum_{e=1}^{n_d}\sum_{k=1}^{5} \hat{Y}_{ked}}{\sum_{e=1}^{n_d}\sum_{k=1}^{5} X_{ked}} X_d \qquad (4.7)$$

where X_d is the census count across all postcodes in HtC category d for the age-sex group being estimated. As with the Horvitz-Thompson approach the estimator of R_d can also be adapted to allow for calculating the DSE at different levels. If the DSE for each cluster of five postcodes is represented by $\tilde{Y}_{ed}$, an alternative version of (4.6) is

$$\hat{T}_{RAT} = \sum_{d=1}^{5} \frac{\sum_{e=1}^{n_d} \tilde{Y}_{ed}}{\sum_{e=1}^{n_d}\sum_{k=1}^{5} X_{ked}} X_d \qquad (4.8)$$

where again the difference is whether you sum postcode DSEs within the ED or sum postcode counts within the ED and then calculate the DSE. A third alternative is to compute one DSE across all the CCS sample postcodes within a HtC stratum, then

'ratio' this total up to a population estimate for that stratum. If the single DSE is represented by $\dot{Y}_d$, then this third alternative version of (4.6) is

$$\hat{T}_{RAT} = \sum_{d=1}^{5} \frac{\dot{Y}_d}{\sum_{e=1}^{n_d} \sum_{k=1}^{5} X_{ked}} X_d \qquad (4.9)$$

This is analogous to treating the HtC stratum as a post-stratum in the US Census context and applying the ratio estimator in the form proposed by Alho (1994). If the true counts are known in the sampled areas (ie the CCS has no non-response) all three estimators given by (4.7), (4.8), and (4.9) are equivalent and reduce to the same estimator for T_a.

A reasonable expectation is that the approaches based on (4.8) and (4.9) will have lower variances due to the larger counts contributing to the DSE but be increasingly subject to correlation bias due to heterogeneity of capture probabilities within each HtC stratum and possible dependence. Defining the HtC strata after the census can reduce this correlation bias as is done by the US Census Bureau. However, it appears unlikely that all the necessary data for such a post-stratification will be available in time for such an exercise to be carried out on the UK data after the 2001 Census.

4.2.2.3) The Impact of Dual-System Estimation on the Ratio Model

The ratio model defined by (4.5) assumes that from the survey the true population counts are known in the sampled postcodes. The reality is that the true count for postcode k is replaced with its DSE $\hat{Y}_{ked}$, where

$$\begin{aligned} E[\hat{Y}_{ked} \mid Y_{ked}] &= Y_{ked} \\ Var[\hat{Y}_{ked} \mid Y_{ked}] &= Y_{ked}\sigma^2_{rked} \end{aligned} \qquad (4.10)$$

The variance term in (4.10) uses the fact that the variance of the DSE is proportional to the population size being estimated multiplied by a term that depends on the

coverage probabilities of the census and the CCS (see equation 2.8). In general, this term will vary from postcode to postcode as the homogeneity assumption for the DSE only needs to hold independently within each postcode. In reality, it is feasible that these coverage probabilities will not vary tremendously across postcodes within the same ED or same category of the HtC index. Seber (1982) demonstrates that for the DSE with the Chapman correction, the variance is also proportional to the population size being estimated.

Replacing the true population count by the DSE in (4.5) requires an expression for

$$E[\hat{Y}_{ked} \mid X_{ked}] = E[E[\hat{Y}_{ked} \mid Y_{ked}, X_{ked}] \mid X_{ked}] \tag{4.11}$$

Under the assumption that the DSE is independent of the census count conditional on the truth, (4.11) can be re-expressed as

$$E[\hat{Y}_{ked} \mid X_{ked}] = E[E[\hat{Y}_{ked} \mid Y_{ked}] \mid X_{ked}] \tag{4.12}$$

The assumption of independence is crucial but unfortunately it can not be justified in general and so the following analysis is mainly illustrative. However, given the independence assumption and (4.10), the expected value for the DSE conditional on the truth, (4.12) becomes

$$E[\hat{Y}_{ked} \mid X_{ked}] = E[Y_{ked} \mid X_{ked}] = R_d X_{ked} \tag{4.13}$$

so that provided the DSE is an unbiased estimator of the true count (4.13) demonstrates that replacing the true count by the DSE in (4.5) will not affect the expectation. The variance follows as

$$\begin{aligned} Var(\hat{Y}_{ked} \mid X_{ked}) &= E[Var(\hat{Y}_{ked} \mid Y_{ked}) \mid X_{ked}] + Var(E[\hat{Y}_{ked} \mid Y_{ked}] \mid X_{ked}) \\ Var(\hat{Y}_{ked} \mid X_{ked}) &= E[\sigma^2_{rked} Y_{ked} \mid X_{ked}] + Var(Y_{ked} \mid X_{ked}) \\ Var(\hat{Y}_{ked} \mid X_{ked}) &= \sigma^2_{rked} R_d X_{ked} + \sigma^2_{ed} X_{ked} = X_{ked}\left(\sigma^2_{rked} R_d + \sigma^2_{ed}\right) \end{aligned} \tag{4.14}$$

by again using the relationships in (4.10) and (4.5). Although (4.13) implies that replacing the true count with the DSE in (4.5) will not affect the expected value, (4.14) shows it will impact on the variance, the relationship being more variable when the DSE is used. The best linear unbiased estimator for R_d is now the weighted least squares estimator given by

$$\hat{R}_d = \frac{\sum_{k \in S_d} Y_{ked} \Big/ \left(\sigma^2_{rked} R_d + \sigma^2_{\varepsilon d}\right)}{\sum_{k \in S_d} X_{ked} \Big/ \left(\sigma^2_{rked} R_d + \sigma^2_{\varepsilon d}\right)} \qquad (4.15)$$

Using (4.15) does not lead to the same estimator of the total given by (4.7) unless the capture probabilities are constant across the sampled postcodes implying $\sigma^2_{rked} = \sigma^2_{rd}$. However, even when this does not hold, ignoring the weights in (4.15) will have little impact provided $\sigma^2_{rked} R_d$ is 'small' compared to $\sigma^2_{\varepsilon d}$. This is a reasonable assumption based on (2.8). In conclusion, ignoring the fact that the true count is estimated by the DSE and estimating the total using (4.7) will in general be approximately unbiased and exactly unbiased when the capture probabilities are homogeneous across all sampled postcodes in the HtC category.

4.2.2.4) Regression Model for Population Estimation

The model (4.5) specifies a proportional relationship between the census and true counts. However, such a relationship does not hold in the case where census counts are zero, but the CCS counts individuals. Therefore, Brown *et al* (1999) proposed the use of a simple regression model to explicitly allow for such situations. This model is given by

$$\begin{aligned} &E\{Y_{ked} \mid X_{ked}\} = \alpha_d + \beta_d X_{ked} \\ &Var\{Y_{ked} \mid X_{ked}\} = \sigma^2_d \\ &Cov\{Y_{ked}, Y_{jfg} \mid X_{ked}, X_{jfg}\} = 0 \text{ for all } k \neq j \end{aligned} \qquad (4.16)$$

Under (4.16) it is straightforward to show (Royall, 1970) that the best linear unbiased estimator for the true population total T of an age-sex group is then the stratified regression estimator

$$\hat{T}_{REG} = \sum_{d=1}^{5} \sum_{e=1}^{N_d} \sum_{k=1}^{M_e} (\hat{\alpha}_d + \hat{\beta}_d X_{ked}) \qquad (4.17)$$

where $\hat{\alpha}_d$ and $\hat{\beta}_d$ are the OLS estimates of the model parameters α_d and β_d in (4.17). Like the ratio estimator defined in (4.6), (4.17) is robust to the intra ED correlation of postcodes due to the sample design (Scott and Holt, 1982).

There are two problems with the regression model (4.16). The first is the appropriateness of the constant variance assumption when using count data. This was a criticism made by Cressie (1989) regarding the use of regression models by Ericksen and Kadane (1985). Secondly, it is not robust to a large number of zero census / CCS counts, since a large number of sample postcodes with the same census / CCS counts can significantly influence the fitted regression line. However, no statistical model is perfect and considering the regression model provides an alternative with which to compare to the ratio model. As with the ratio model, actual estimation involves replacing the unknown true postcode counts with their DSEs.

4.3) Simulation Study

The work in section 4.2 presented a two-stage strategy for estimating the population by age and sex for an estimation area using data from the 2001 Census and the CCS. The first stage acknowledges that both the 2001 Census count for an age-sex group within a postcode, and the corresponding CCS count, will be subject to underenumeration. Under the assumptions specified above, dual-system estimation, corrected for its small sample bias, can be used to combine the two counts and estimate the true counts by age and sex for all postcodes in the CCS sample. The second stage then considers the problem of estimating a population total from a sample.

Section 4.2.2 outlines three approaches to the second stage. The first ignores the existence of the 2001 Census counts for non-sampled postcodes. This is essentially the estimation approach underpinning the design of the CCS and, as already pointed out in chapter three, is inefficient. The second and third strategies use the 2001 Census counts as an auxiliary variable, assuming either a proportional (ratio model) or linear (regression model) relationship between the true postcode counts and the 2001 Census counts.

In this section we develop and implement a simulation model based on data from the 1991 Census in order to assess all three strategies as well as the basis of the approach taken by the US Census Bureau. Underpinning this is a model of census response patterns and likely CCS response patterns constructed from the evidence available on the patterns of underenumeration in the 1991 Census. The model is applied to a set of anonymous individual records for a single local authority district (LAD) from the 1991 Census, augmented by the prototype HtC index described in section 3.4, and these form the basis for the simulation. The population is the same data as used in chapter three. As the LAD contains approximately half a million individuals, in 930 EDs, it is treated as a single estimation area.

4.3.1) Applying the CCS Design to the Simulation Population

One of the problems with designing the CCS is that the only detailed information available is the previous census, in this case the 1991 Census, and this information is out of date. This means that the design is based on a certain population structured that will almost definitely no longer exist in the current population. If the population used in the simulation study were also the exact population that the CCS design was based on this source of variability would not be captured by the simulation study.

To overcome this problem the ED data for the simulation population was 'aged' backwards using data from the mid-year population estimates, produced by ONS, for LADs in England and Wales between 1991 and 1996. For each LAD, a simple exponential growth curve was fitted over the six years for each of 24 age-sex groups

defined as; males aged 0 to 4, ..., males aged 40 to 44, males aged 45 to 79, males aged 80 to 84, males aged 85+, and the corresponding 12 age groups for females. (These correspond to the same 24 groups specified in chapter three.) This defined a population of about 400 different sets of growth curves. Taking a simple random sample with replacement, each of the 930 EDs in the simulation population was assigned one set of growth curves. The growth curves were then used to predict the ED population by age and sex for a point ten years in the past. This approach was used, rather than just applying the actual growth curve for the LAD to all EDs within the simulation population, to capture the fact that while at the LAD level the population may be quite stable over the ten years, changes at the ED level can vary quite dramatically.

The design using multivariate stratification outlined in section 3.5 was then applied the set of 'aged' ED populations using the same prototype HtC index as applied in chapter three. This is an issue where the simulation is not completely realistic. The HtC classification used at the design stage, and at the estimation stage, is also used as a variable that defines census response. However, in reality the design is based on a HtC classification constructed from the previous census and while the variables related to census underenumeration will probably change little, mobility of individuals will mean that the actual population in the ED in 2001 may not exactly reflect the HtC classification made from the 1991 Census. The US Census Bureau tackles this problem through post-stratification. The variables used are chosen prior to the census but then membership of the post-strata is defined after the census based on where individuals are actually found. This is of particular importance in the US context as the homogeneity assumption, underlying the dual-system estimator (DSE), must be approximated at the post-strata level. However, in the context of the estimation strategy outlined in section 4.2, postcodes incorrectly categorised with respect to the HtC index will simply increase the variance in either the ratio or regression model rather than lead to heterogeneity bias in the DSE. In extreme cases, postcodes may need to be treated as outliers when estimating the model parameters.

The sampling fraction chosen represents a sample of 20,000 postcodes (4,000 EDs) for England and Wales as specified in ONS (1998a). This implies a total sample of

approximately 35 EDs with a simple random sample of five postcodes per selected ED (or less if the ED does not contain five postcodes). Table 4.1 shows the distribution of EDs by the HtC index in the estimation area and the number of EDs selected in each stratum.

TABLE 4.1

Distribution of enumeration districts by HtC index for the total population of EDs and the ED sample

HtC Index Value	Number of Enumeration Districts	Sample of Enumeration Districts
Very Easy	144	6
Easy	210	7
Medium	186	6
Hard	193	7
Very Hard	197	9
TOTAL	930	35

4.3.2) Simulating a Census and its CCS

The first stage of simulating census underenumeration was to build a model for census underenumeration based on the experiences of 1991. The starting point for this was the set of ED adjustment factors by age and sex, calculated as part of the 'Estimating with Confidence (EwC) Project'. These are available for academic research via the Manchester computing facilities. Each ED on the EwC data set was assigned its HtC category based on the prototype index. Then, using sampling with replacement, each individual was assigned at random an adjustment factor from the EwC data based on their age, sex, and the HtC category of their ED. The inverse of the adjustment factor assigned to the individual forms the basis of their probability of being counted in a census. Using this approach has four advantages.

a) The EwC data represents the 'best' estimate available of the patterns of underenumeration in the 1991 Census by age and sex at the ED level.

b) The EwC data is consistent with higher-level estimates of census underenumeration and so at the estimation area level the censuses generated by the simulation should have plausible underenumeration patterns.

c) The EwC data does not just adjust for census underenumeration; it also corrects the location of students to account for definitional differences between the 1991 Census and the mid-year population estimates. The advantage of this is that some ED adjustment factors for young adults are large, greater than ten, and correspondingly, in the simulation population, some young adults have a very low probability of being included in the census. This does also mean that some adjustment factors are less than one. For those individuals the probability of coverage by the census is set equal to one.

d) By assigning a different ED adjustment factor at random to the individuals means that within a postcode, two individuals with the same age and sex will have a similar propensity to go missing from the census but not the same probability. This is important so that the census underenumeration model does not exactly satisfy the DSE assumption of homogeneity.

While age and sex dominate the patterns of census underenumeration, they are not the only factors. Based on anecdotal evidence from the 1991 Census and further research by the 'Estimating With Confidence Project' (Simpson *et al*, 1997), the probabilities assigned to each individual were also adjusted based on the census variable 'Primary Activity Last Week' so that the categories representing unemployed had lower census coverage than those representing paid employment. A small ED effect was also introduced to represent local factors that may affect all the individuals within an ED, such as a poorly motivated enumerator. Within the simulation data, each household was also assigned a probability of being counted in a census. The starting point for this was the average of the probabilities across the individuals within the household. This was then adjusted so that those households with 'Tenure' categories of private rented had lower probabilities than those households with homeowner categories. In addition, households containing only one individual were also given lower census coverage probabilities than larger households. The household and individual probabilities remain fixed throughout the simulation study.

Each individual and household is also assigned a value that defines the differential nature of response in the CCS. These mirror the same pattern as the census coverage probabilities but the differentials are much less extreme. This extends the simulation study in Brown *et al* (1999) so that there is some heterogeneity in both the census and the CCS for age-sex groups at the postcode level. Further details of the assigning of census and CCS coverage probabilities are given in Appendix 4.1.

To generate a census and its corresponding CCS, independent Bernoulli trials are used to determine first whether the household is counted and second whether the individuals within a counted household are counted. There is also a check that converts a counted household to a missed household if all the adults in the household are subsequently missed at the individual stage. In these simulations the census and CCS outcome for households and individuals are independent. This assumption can be investigated by specifying the odds ratio between the two outcomes to be different from one, see Brown *et al* (1999). Two levels of coverage are used in the CCS. First a perfect CCS is simulated and then coverage in the CCS is set at approximately 90 per cent for households with 98 per cent of individuals within those households being counted. For each census ten CCS postcode samples are selected based on the design in Table 4.1. The estimators described in section 4.2.2 are then applied to each age-sex group and population totals are calculated. The whole process is repeated for 100 independent censuses.

4.3.3) Population Estimation Results

For the simulation of 100 censuses the average census coverage for the total population is 94.90 per cent, which drops to 85 per cent for males aged 20-29 and females aged 85+. Details of the census coverage for each age-sex group are given in Appendix 4.2. The coverage is rather less than observed nationally in 1991 where it was around 98 per cent. However, the simulation aims to assess the robustness of the procedure to the levels of underenumeration that were observed in certain areas such as areas of inner London and the major cities such as Birmingham and Manchester. For such areas, the census response patterns for the simulation population are in line with the census adjustment factors reported in Heady *et al* (1994).

In section 4.2.2 three strategies for estimating from the CCS postcodes to the total population are discussed. However, there are actually eight estimators being evaluated and they are:

a) Weighted DSE – **(2.5)**
b) US Census Bureau weighted DSE – **(2.6)**
c) The Horvitz-Thompson (HT) estimator with the DSE at the postcode level – **(4.3)**
d) The Horvitz-Thompson (HT) estimator with the DSE at the cluster level **(4.4)**
e) The ratio estimator with the DSE at the postcode level **(4.7)**
f) The ratio estimator with the DSE at the cluster level **(4.8)**
g) The ratio estimator with the DSE at the HtC index level **(4.9)**
h) The regression estimator with the DSE at the postcode level **(4.17)**

where the number refers to the equation that defines the estimator. In estimators c to h the Chapman correction is applied to the unweighted DSEs.

As this is a simulation study the true population is known. Therefore, the relative root mean square errors (RRMSE) and the relative biases, calculated from the empirical distributions for the estimators based on the simulation study, can be used to assess the performance of the estimators relative to each other (and the census) over the 1000 CCSs. For each estimator the RRMSE is defined as:

$$\text{RRMSE} = \frac{1}{\text{truth}} \times \sqrt{\frac{1}{1000} \times \sum_{j=1}^{1000} \left(\text{observed}_j - \text{truth}\right)^2} \times 100 \qquad (4.18)$$

and can be considered as a measure of the total error due to bias and variance. Relative bias is defined as:

$$\text{Relative Bias} = \frac{1}{\text{truth}} \times \frac{1}{1000} \times \sum_{j=1}^{1000} \left(\text{observed}_j - \text{truth}\right) \times 100 \qquad (4.19)$$

The usual aim is to opt for an estimation strategy that will give unbiased estimation, as bias is not easily estimated from the sample. (If the bias of an estimator can be estimated with precision that would imply that an efficient unbiased estimator is also available.) However, it can be better overall to adopt a slightly biased estimator if its total error is small. In addition to estimating the bias of a particular estimation strategy using (4.19), the standard error for the estimated bias can also be estimated by assuming independent iterations and ignoring the fact that ten independent CCS simulations were carried out for each of 100 independent census simulations. As a consequence of this the standard errors are likely to be slightly under-estimated and some care may be needed when using them in inference. However, any standard error can still be used to calculate an indicative Z-value and assuming normality, this can be used to indicate whether the bias is significantly different from zero. This allows for the fact that any estimated bias might be due to Monte Carlo variation where the simulation has not run for sufficient iterations.

TABLE 4.2

Performance of the population estimators based on weighted DSEs

Type of Estimator	Relative Bias (%)	Relative RMSE (%)	Z-value for Bias
Weighted DSE	0.26	6.82	1.23
US Weighted DSE	**0.23**	**0.61**	**13.31**

TABLE 4.3

Performance of the population estimators based on unweighted DSEs

Type of Estimator	Relative Bias (%)	Relative RMSE (%)	Z-value for Bias
Horvitz-Thompson			
Perfect CCS	0.24	6.83	1.11
Postcode DSE	0.11	6.82	0.50
Cluster DSE	0.22	6.82	1.03
Ratio Estimator			
Perfect CCS	***0.24***	***0.52***	***15.92***
Postcode DSE	***0.10***	***0.49***	***6.70***
Cluster DSE	***0.22***	***0.53***	***13.95***
Index DSE	***0.23***	***0.54***	***15.30***
Regression Estimator			
Perfect CCS	0.37	0.64	22.77
Postcode CCS	0.23	0.57	13.97

Table 4.2 and Table 4.3 summarise the results for the estimation of the total population derived by summing the individual age-sex estimates. Across the different types of estimator the relative bias for each is similar although the bias estimated from the simulations for the simple weighted DSE in Table 4.2 and the Horvitz-Thompson estimators in Table 4.3 are not statistically significant at the 95% level. This can be seen from the Z-values for the bias, which are less than two. However, when you look at the total error measured by the relative RMSE those estimators are much less efficient. This is exactly what you would expect as the Horvitz-Thompson estimators with unweighted DSEs and the simple weighted DSE make no use of extra information available from the 2001 Census for postcodes not in the CCS sample.

Based on the results in Table 4.3, estimators based on the ratio model are generally better than the regression estimator both in terms of bias and total error. The bias for the ratio model with a perfect CCS, which based on the Z-value of 15.92 cannot be due to Monte Carlo error, will primarily come from two sources. The first is the fact that with respect to repeated sampling, ratio and regression type estimators are not exactly unbiased, whereas Horvitz-Thompson estimation is. Both regression and ratio estimators have a bias that tends to zero as the sample size increases. However, with the CCS as the ratio and regression estimators are applied independently within each HtC category the sample sizes are only between 30 and 45 postcodes. The second source is likely to be a failure of the appropriateness of the estimation model for certain censuses and CCSs.

Considering the relative biases and relative RMSEs across all the estimators in Table 4.2 and Table 4.3, the postcode DSE with the ratio model looks 'best', as the relative bias of this estimator is less than for a perfect CCS. Based on the simple analysis of combining the DSE with the ratio model in section 4.2.2.3, this should not be the case, introducing the DSE should have negligible impact on the bias of the estimator with some increase in variance. (In fact, introducing the DSE at the postcode level does increase the RSE from 0.46 per cent to 0.48 per cent but in terms of the total error the drop in the bias hides this.) In addition, there is a similar reduction in bias when the DSE component is introduced at the postcode level for the regression model and the

Horvitz-Thompson approach. Caution is required; it is not the case that the postcode level DSE 'fixes' the problems causing the bias for a perfect CCS but that the unweighted DSE at the postcode level introduces an additional negative bias. The results in Table 4.3 suggest this additional bias is not introduced when the DSE is used at either the cluster or index level. This is further confirmed by the results in Table 4.4.

TABLE 4.4

Performance of the DSE at two levels for estimating the sample population

	Relative Bias (%)	Z-value for Bias
DSE at Postcode Level	-0.10	-22.01
DSE at Cluster (5 Postcode) Level	0.0086	1.82

Table 4.4 presents results for just estimating the population in the sample postcodes (by simply summing the unweighted DSEs) over the 1,000 iterations of the simulation. The Z-values show that at the postcode level the DSE has a highly significant negative bias. Looking more closely at the simulation population suggests that one cause is the very small counts in the individual age-sex groups leading to zero cells. The consequence of this is that the multinomial model on which the DSE is based will not always be appropriate at the postcode level. In addition, the Chapman correction requires $n_{1+} + n_{+1} > n_{++}$ for exact unbiasedness and this will also be more likely to fail when observed counts get close to or equal zero. Presented in Seber (1982), the work of Robson and Regier (1964) gives the following approximation for the bias of the DSE with the Chapman correction when the above condition for unbiasedness fails.

$$E[\hat{n}^{C}_{++} \mid n_{1+}, n_{+1}] = N - Nb \text{ where } b = \exp\left\{-\frac{(n_{1+}+1)(n_{+1}+1)}{n_{++}}\right\} \qquad (4.20)$$

This shows that when the estimator is biased, unlike the standard DSE the bias given by (4.20) can be non-negligible; especially when the number of individuals counted in both the census and CCS is small relative to the true population. However, the results in Table 4.4 confirm that the DSE at the cluster level is essentially unbiased over the

simulation. Therefore, the cluster level DSE with the ratio model looks the 'best' option as a compromise between the unstable DSE at the postcode level due to very small (zero) counts and the increased risk of correlation bias with the index level DSE.

Comparing the US weighted DSE in Table 4.2, essentially a design-based ratio estimator, with the model-based ratio estimators in Table 4.3 highlights two additional points. Excluding the postcode level DSE ratio estimator for the reasons above, all give approximately the same bias due to 'model failure' and 'ratio bias' over repeated sampling. The second point is a gain in terms of total error, implying reduced variance for the model-based approach which, when the model is appropriate, is what would be expected.

Looking at the total population can hide problems with the estimation of the individual age-sex population estimates. Figure 4.1 presents the results for the male age groups using the ratio estimator with both weighted and unweighted DSEs. The results for females are in Appendix 4.3.

Figure 4.1: Ratio estimators combined with weighted and unweighted dual system estimation for males by age

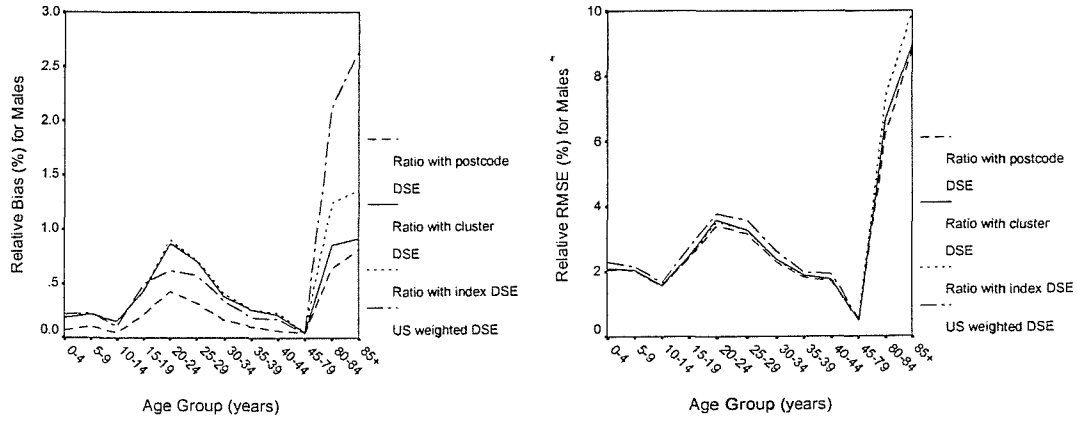


Figure 4.1 demonstrates that across the age groups for males the four estimators are very similar in terms of RRMSE with the exception of men aged 85 years and over where there are more noticeable differences. In particular, the results for the US weighted DSE are not displayed as these are 12.85 per cent for males age 80-84, and

24.88 per cent for males age 85+. This is not a problem in the context of estimation after the US Census as the post-stratification is for much broader age-sex groups. However, in the UK context the current decision is to produce estimates by sex for five-year agegroups and therefore this would require consideration. (The use here of a single group for 45-79 reflects the fact that there is no variation by age in underenumeration for this group in the simulation.)

With respect to bias Figure 4.1 shows that for the three estimators based on the ratio model, with unweighted DSEs, the postcode DSE has a consistently lower bias. This reflects the negative bias in the DSE at this level demonstrated by Table 4.4 and discussed previously. In general it is not good practice to rely on biases cancelling each other to get a 'better' estimator and in terms of total error there is very little impact confirming that any reductions in bias are balanced by increased variance. Comparing the US weighted DSE with the cluster level and index level unweighted DSEs, the US approach has a smaller bias for the young males. This is a crucial age group for the estimation of total underenumeration. However, as seen in Table 4.2 and Table 4.3, this does not translate into a gain in terms of bias for the total population as the US weighted DSE has a much higher bias for the oldest agegroups.

4.3.4) Conclusions on the Simulation Study Results

Taking the results of Tables 4.2, 4.3, and 4.4 with Figure 4.1, the estimators based on the ratio model with unweighted DSEs are best overall with little to choose between the cluster level DSE and the index level DSE. However, in practice the estimator using the index level DSE needs to be treated with care as it relies heavily on the HtC index defining homogeneous strata with respect to the response rate in either the census or the CCS. This is the condition for the bias term in Wolter (1986) due to heterogeneity to equal zero. In the simulation, although not exactly satisfied, once age and sex are controlled for, this is approximately true for the CCS response rate.

In 2001 assumptions of homogeneity across all postcodes in each HtC category will be shakier when the index has been defined for postcodes based on their 1991 characteristics and there will certainly be postcodes that will have changed in ten

years. This will cause the DSE calculated at the HtC stratum level to be biased. For the postcode-based estimator this will not impact on the individual DSEs to cause bias but it will increase the variance as the relationship between the census and CCS counts within each stratum will not be as strong. However, as demonstrated by the simulation results, at this level of aggregation the DSE is unstable. Therefore, estimators with a cluster level DSE are, as already stated, a good compromise between the two 'extremes' of potential heterogeneity bias and problems with small population counts.

4.4) A 'Robust' Estimation Strategy

The results from the simulation presented in section 4.3.3 demonstrate the existence of problems with both the ratio and the regression model as the census count gets small causing model failure and potentially bias. As stated in Section 4.2.2.4, the regression model will fit well when census counts are approaching zero and the CCS is finding extra people but it will not be robust to a large number of postcodes where both the census and CCS counts are zero. As the postcode is a very small geographic area the count for a particular age-sex group will often be zero. While such postcodes do not affect the ratio model, as it is constrained to pass through the origin, postcodes where the census count is zero and the CCS is greater than zero do. These happen in a few postcodes for all the age-sex groups.

There is a second issue that impacts on the estimation. The ratio estimator essentially uses the ratio estimated from the sample to predict the count in the non-sample areas. This becomes a problem when there exist census counts in the non-sample postcodes that are greater than those in the sample postcodes. In such situations, the danger is making a prediction where there is no sample to support such a prediction and a few outlying census counts can have a considerable impact on the final estimate.

There is a third situation that can occasionally occur. It happens when there are large numbers of census counts in the sample areas that are zero, some with a non-zero CCS count, combining with an extreme form of problem two where the non-zero census counts in the sample areas are all close to zero. This results in the situation

where zero census non-zero CCS counts have a large impact on the estimated ratio which is then used to predict for counts well outside the range of the sample data. The oldest age groups are particularly vulnerable to this happening leading to the positive bias observed in Figure 4.1.

4.4.1) A Model for Robust Estimation

The previous section highlights three problems with the ratio model that are causing model mis-specification bias. This section takes each problem in turn and proposes a strategy to deal with it. The first problem is a zero census count with a non-zero CCS count. Initial work considered a mixture type model to cope with this problem but simulations showed that the sample provided insufficient data to facilitate estimation of all the necessary parameters. Therefore, a simpler approach is proposed below.

If $X_{ked} > 0$;

$$\begin{aligned} &E\{Y_{ked} \mid X_{ked}\} = R_d X_{ked} \\ &Var\{Y_{ked} \mid X_{ked}\} = \sigma_d^2 X_{ked} \\ &Cov\{Y_{ked}, Y_{jfg} \mid X_{ked}, Xg_{jfg}\} = 0 \text{ for all } k \neq j \end{aligned} \tag{4.21}$$

If $X_{ked} = 0$;

$$\begin{aligned} &E\{Y_{ked}\} = \mu_d \\ &Var\{Y_{ked}\} = \gamma_d^2 \\ &Cov\{Y_{ked}, Y_{jfg}\} = 0 \text{ for all } k \neq j \end{aligned} \tag{4.22}$$

The approach outlined in (4.21) and (4.22) works by splitting the estimation into two parts, one for postcodes with a zero census count (4.22) and one for postcodes with a non-zero census count (4.21). Here, the model proposed for the zero census counts is just the simple stratified homogeneous model and for the non-zero counts the standard ratio model.

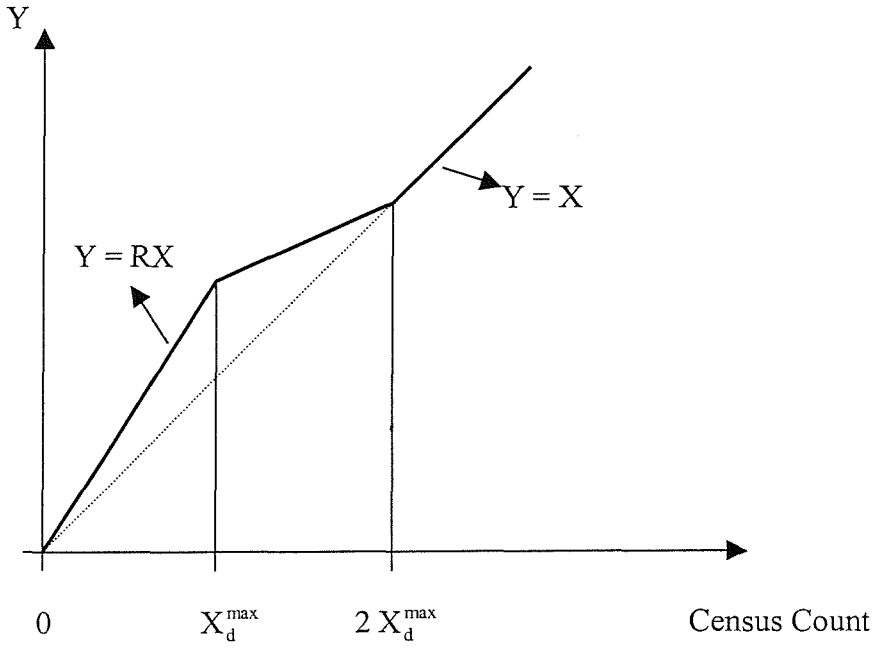
The second problem is the prediction outside the range of the sample data. Empirical evidence from the simulations suggests that this is important and causes large 'over-

estimates'. To reduce the bias caused by this the ratio part of the estimator generated by (4.21) and (4.22) is modified such that the overall estimator is given by

$$\hat{T}_d = \sum_{X_{ked}=0} \hat{\mu}_d + \sum_{0<X_{ked}\le x_d^{max}} \hat{R}_d X_{ked} + \sum_{x_d^{max}<X_{ked}\le 2x_d^{max}} \left\{\left(2-\hat{R}_d\right)X_{ked} + 2x_d^{max}\left(\hat{R}_d - 1\right)\right\} + \sum_{X_{ked}>2x_d^{max}} X_{ked} \qquad (4.23)$$

$$\hat{T} = \sum_{d=1}^{5} \hat{T}_d$$

where x_d^{max} is the largest census count for a CCS sample postcode in a particular age-sex HtC combination and $\hat{\mu}_d$ is the unweighted mean of the Y_{ked}'s in the sample with $X_{ked} = 0$. Graphically, these modifications to the ratio part of the estimator (4.21) can be represented as



where the aim is to reduce the influence of outliers in the census on the final estimate of $\hat{T}$. The choice of $2X_d^{max}$ is arbitrary. It is chosen to reflect a point beyond which it is felt that no adjustment can be made to census counts for postcodes not in the sample based on the postcodes in the sample. The justification for this is that these large census postcodes are not part of the general population of postcode counts. An example of this would be where the census has enumerated a small communal

establishment within a postcode, such as an old people's home, as a set of households, thus creating a very high count of old people.

The final problem does not occur often but when it does the result is usually a dramatic over-estimate. The older ages are most vulnerable but it can occur for any age-sex group within a HtC category for a given sample. To combat this, in the situations where there are not three distinct non-zero census counts in the CCS sample, the estimation strategy outlined by (4.21)-(4.23) is not used for that particular age-sex HtC combination. Instead, an alternative model is used given by

$$\begin{aligned} &E\{Y_{ked} \mid X_{ked}\} = X_{ked} + \delta_d \\ &Var\{Y_{ked} \mid X_{ked}\} = \Omega_d^2 \\ &Cov\{Y_{ked}, Y_{jfg} \mid X_{ked}, Xg_{jfg}\} = 0 \text{ for all } k \neq j \end{aligned} \qquad (4.24)$$

which is just a regression model where β_d is constrained to one. The resulting estimator of the total for the particular age-sex HtC combination is then given by

$$\hat{T}_d = \sum_{e=1}^{N_d} \sum_{i=1}^{M_e} \left(X_{ked} + \hat{\delta}_d \right) \qquad (4.25)$$

where N_d is the number of EDs in stratum d, M_e is the number of postcodes in ED e, and $\hat{\delta}_d = \bar{y}_d - \bar{x}_d$ where $\bar{y}_d$ and $\bar{x}_d$ are the unweighted sample means. The justification of (4.24) in this situation is that the model does not attempt to estimate a slope parameter from very little information. However, it does utilise the fact that the CCS has identified some extra people and combines this with the fact that census counts are available for all postcodes.

The standard ratio estimator is level consistent in the sense of Cressie (1989). In other words, the ratio is applied uniformly to census counts and the overall adjustment would not change if a postcode were split in half. However, this robust strategy does not lead to such an estimator as a 'large' postcode may get no adjustment while two smaller postcodes, formed from splitting this large postcode, could each get some

adjustment. However, in the context on the 2001 Census this is not a problem as the postcode geography will be fixed just prior to the 2001 Census day and all subsequent census outputs will be based on this fixed set of postcodes.

4.4.2) Statistical Motivation for the Prediction in Non-Sampled Areas

The robust approach to prediction of postcode counts in non-sampled areas, outlined in section 4.4.1, is motivated by considering scenarios that can occur based on the output of the simulation study and then applying common sense. An alternative is to consider it in the context of the philosophy behind M-estimation, used in robust estimation of model parameters, extended to the problem of prediction from a model. The approach used in M-estimation is to allow observations to have full influence on the estimation of model parameters over a certain range and then reduce that influence as the residual associated with an observation increases. It is then a case of choosing a sensible function that defines how this happens. There are three basic ideas (see Andrews *et al*, 1972):

a) Trim observations so once residuals exceed some value they have no influence.

b) Decrease the influence of observations once residuals exceed some value such that eventually very large residuals mean the observation has no influence (Hampel type influence functions).

c) Keep the influence of observations at a constant level once residuals exceed some value (Huber type influence functions).

In the context of the CCS and (4.23), the parameter being estimated is the total underenumeration in the population. This is achieved by predicting the true count for all the non-sampled postcodes based on the estimated ratio model, and each postcode's contribution to the estimate of underenumeration is the difference between this predicted count and the observed census count. In this problem the 'residual' is how far the census count for a non-sampled postcode exceeds the census counts in the sampled postcodes and the influence exerted on the total is through applying a ratio to

the observed census count to get a predicted true count. Therefore, the approach to prediction for non-sampled postcodes presented in section 4.4.1 can be considered as an approach with the same ethos as (b). Within the range of the sample data, each postcode exerts full influence on the estimate of underenumeration. Beyond the range of the sample data, reducing the ratio applied to the census counts reduces the influence. Beyond some '*extreme*' point, the ratio applied is one meaning the postcode exerts no influence on the estimation of underenumeration.

4.4.3) Applying the Strategy

The same simulation as used in section 4.3 can now be applied to the robust estimation strategy. For a perfect CCS the above strategy can be applied directly to the sample of postcode counts generated by the simulation. However, in reality Y_{ked} will not be known but will be estimated using dual-system estimation. It has already been shown that at the postcode level the DSE is unsatisfactory but for prediction purposes the models and estimators proposed in section 4.4.1 are at the postcode and not cluster level. Therefore, the postcode level DSE is used but scaled to the cluster level DSE so that they do sum to an unbiased estimate of the population in the sample postcodes. In other words, if F_{ked} is the raw CCS count and B_{ked} is the matched count for postcode k of the cluster of postcodes selected from ED e of HtC category d then

$$Y_{ked}^{DSE} = \frac{F_{ked} \times X_{ked}}{B_{ked}} \text{ and } Y_{ed}^{DSE} = \frac{\sum_{k=1}^{5} F_{ked} \times \sum_{k=1}^{5} X_{ked}}{\sum_{k=1}^{5} B_{ked}} \Rightarrow \delta_{ed} = \frac{Y_{ed}^{DSE}}{Y_{ked}^{DSE}} \text{ giving} \qquad (4.26)$$

$$\hat{Y}_{ked} = \delta_{ed} \times Y_{ked}^{DSE}$$

where Y_{ked}^{DSE} is the postcode level DSE, Y_{ed}^{DSE} is the cluster level DSE, and δ_{ed} scales Y_{ked}^{DSE} so that $\hat{Y}_{ked}$, the count used for estimation, is consistent with Y_{ed}^{DSE}.

In addition to the strategy outlined in section 4.4.1, postcodes are treated as outliers if when estimating the overall ratio R_d, the ratio defined for that postcode exceeds pre-specified bounds when using the model given in (4.21) in conjunction with the

estimator defined by (4.23). The bounds are; greater than or equal to three for HtC categories one and two, greater than or equal to four for HtC categories three and four, and greater than or equal to five for HtC category five. Currently, these bounds have been chosen based on examining output generated by the simulation. There is the possibility of further work to refine the bounds and make them age-sex and HtC specific. If a postcode is defined as an outlier it is removed from the estimation process and then simply added on to the estimate of T at the end.

Having some method for dealing with extreme counts, generated after dual-system estimation has been applied, is necessary. Occasionally, purely due to random variation, the DSE will simply be an unrealistic estimate of the population total. For example, consider a postcode where the census counts six individuals, the CCS counts two, and one person is in both. The DSE, with the Chapman correction, would estimate the total population as 9.5, and the adjustment factor for underenumeration would be 1.58. This may be a little high but hardly extreme. However, consider the same postcode but this time the census counts two and the CCS counts six. The estimated adjustment factor for underenumeration would now be 4.25. For the specific postcode it may well be the correct representation, but unless the 2001 Census has been subject to extreme levels of underenumeration in a particular area, is unlikely to represent a large number of postcodes. In addition, as the HtC index is based on the 1991 distribution of the variables that constitute the index, it is likely that a few postcodes will have dramatically changed from, for example, HtC category two in 1991 to HtC category five in 2001. This small number of postcodes may produce estimates of underenumeration that are inconsistent with the rest of the sample and likely to be unrepresentative of the population as a whole. The existence of cases caused by either of the above should be rare. Therefore, in the context of section 4.4.2 and m-estimation, the influence of postcodes is being trimmed, approach (a), beyond the pre-specified cut-off points.

4.4.4) Results from the Robust Estimation Strategy

The simulation study was repeated using the methodology described in section 4.3 with the same census and CCS coverage rates. In this case there is an additional estimator being evaluated:

i) The 'robust' ratio estimator with the DSE at the postcode level constrained to the DSE at the cluster level.

where 'robust' ratio estimator refers to the whole strategy outlined in section 4.4.1 and applied as per section 4.4.3. Table 4.5 summarises the results for the estimation of the total population derived by summing the individual age-sex estimates and compares this with estimators (2.6) and (4.8) that were analysed in section 4.3.

TABLE 4.5

Performance of the 'robust' ratio estimators for the population total compared with simpler approaches

Type of Estimator	Relative Bias (%)	Relative RMSE (%)	Z-value for Bias
US Weighted DSE	0.23	0.61	13.31
Ratio Estimator			
Perfect CCS	0.24	0.52	15.92
Cluster DSE	0.22	0.53	13.95
Robust Ratio Estimator			
Perfect CCS	***-0.02***	***0.47***	***-1.30***
Postcode DSE Constrained to Cluster DSE	***-0.07***	***0.48***	***-4.43***

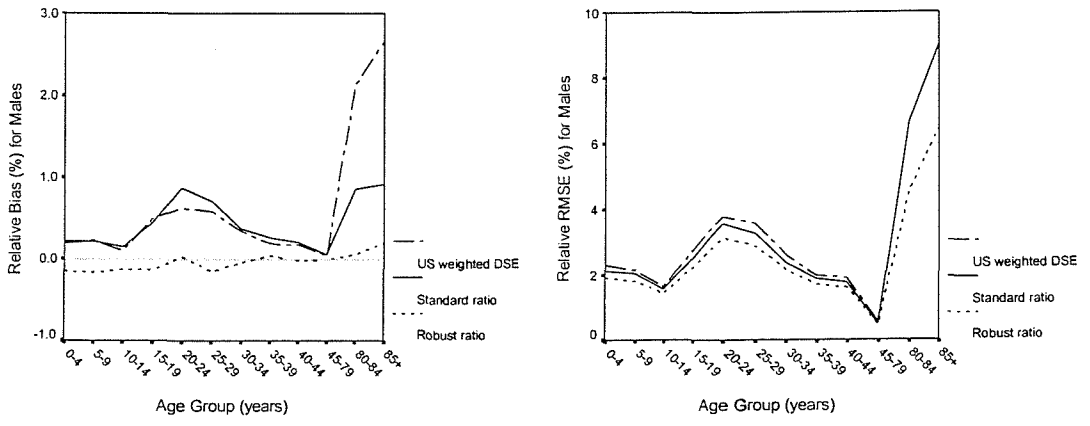
Table 4.5 demonstrates that for a perfect CCS, the adjustments to the estimation strategy are working as expected to reduce bias. This is demonstrated by comparing the bias of -0.02 per cent in Table 4.5 for the robust strategy with 0.24 per cent for the standard ratio estimator. In addition, the bias for the robust ratio estimator is no longer significantly different from zero. Once the DSE is introduced into the estimation strategy, the negative bias does increase slightly and becomes significant. As expected, introducing the DSE also leads to a slight increase in the variance.

However, the constraining of the unweighted postcode DSEs to the unweighted cluster DSEs is reducing the impact of the bias seen in Table 4.4; a consequence of applying the DSE at low levels of aggregation.

The bias of -0.07 per cent in Table 4.5, for the robust strategy combined with dual-system estimation, can be compared to the bias of 0.22 per cent for the standard ratio estimator with cluster level DSE and the bias of 0.23 per cent for the US weighted DSE. This suggests that there is an 'over-correction' for the bias, and the z-value confirms that it is significant. However, the aim of applying these adjustments is not to produce an 'unbiased' estimator with respect to repeated sampling but to make the strategy more robust and produce a better estimate for a given set of data. The results in Table 4.5 suggest that this is being achieved, as there is also a decrease in the total error to 0.48 compared to 0.53 and 0.61. This reduction not only represents the reduction in absolute bias but also a slight reduction in variance by reducing the impact of a few extreme iterations from the simulations on the overall performance.

As before, looking at the total can hide what is happening across the age-sex groups. Figure 4.2 presents results for males that compares the robust approach using the postcode level DSE constrained to the cluster level DSE with the standard ratio model using cluster level DSE and the US weighted DSE.

Figure 4.2: Comparison of the robust ratio model using a postcode DSE constrained to a cluster DSE with the standard ratio model using cluster DSE and the US weighted DSE for males by age

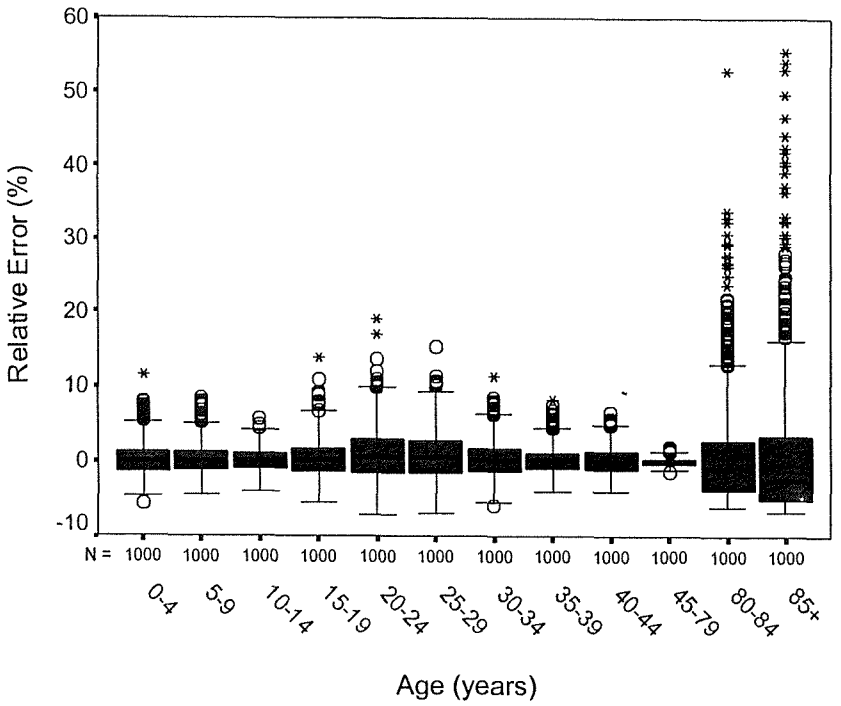


The graph in Figure 4.2 for relative RMSE shows a gain at all ages in terms of total error from using the robust approach. The graph for the bias shows that the robust procedures do introduce some negative bias into the estimator, particularly at the youngest age groups. However, it also shows that the procedures are doing well to correct the large positive bias for males in the age groups 20-34 that is present with the standard estimators. The robust procedures also work better for the oldest age groups in terms of bias and total error where the standard ratio model is particularly unsatisfactory due to very low population counts in many postcodes.

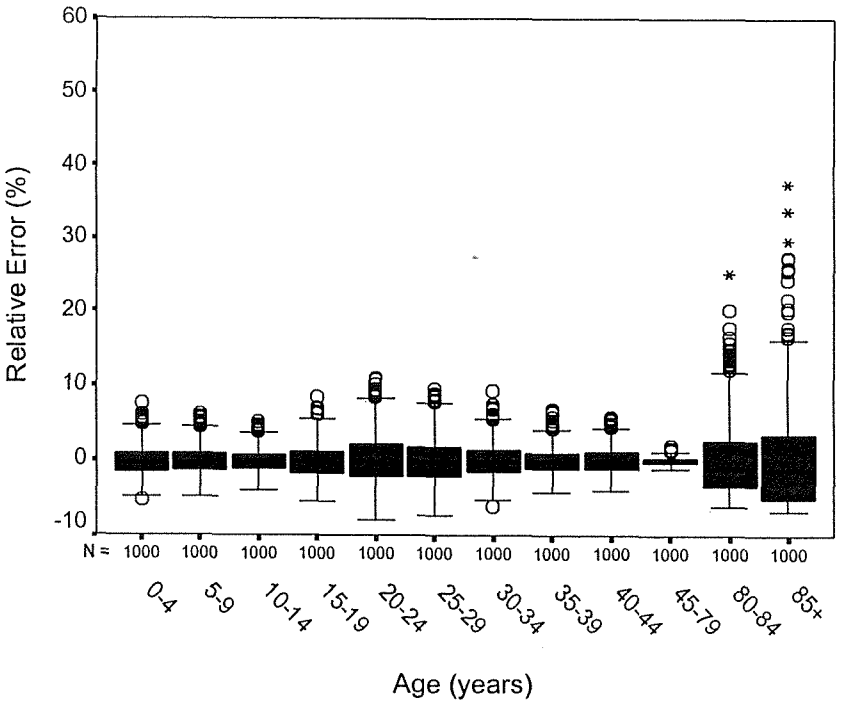
By looking at the empirical distribution of the estimators generated from the simulation study, Figure 4.3 demonstrates more clearly exactly what the robust procedures are doing. The introduction of a negative bias into the estimation strategy has been achieved by preventing the estimator producing extreme over-estimates of the population through reducing the influence of outlying points in the estimation procedures. Therefore, the reduction in total error seen in Figure 4.2 reflects not only a reduction in absolute bias but also a reduction in variance and Figure 4.3 demonstrates that this can be attributed to the application of the robust strategy.

Figure 4.3: Distributions of the errors for the standard and robust strategies

Errors from a standard ratio estimator with a cluster level DSE for males by age



Errors from a robust ratio estimator with a constrained postcode level DSE for males by age



4.4.5) The Impact of Dependence

The simulation results presented so far for both the standard estimators in section 4.3, and the robust approach, were based on the assumption that the census and the CCS are independent of each other. In other words, the probability that an individual is counted in both the census and CCS is the product of the probability that they are counted in the census with the probability that they are counted in the CCS. A failure of this assumption can occur for two reasons. The first is a failure of the homogeneity assumption. Suppose the DSE is applied to a group of individuals where the responses for 80 per cent of the group are generated from a particular multinomial model but 20 per cent are generated by a different multinomial model. The census and CCS can be independent of each other in both models but at the level the DSE is applied the data will not follow the independence model. An element of this is already in the simulations as there is an element of heterogeneity in both the census and CCS response probabilities at the level of aggregation the DSE is applied to.

The second reason refers to a failure to keep the census and CCS operationally independent. In other words, the response of an individual to the CCS starts to 'depend' on their response to the census. An example of this would be using a census enumerator in the CCS and sending them to postcodes sampled from the ED they had enumerated in the census. If the enumerator missed a housing unit when carrying out the census they will probably miss it in the CCS. Therefore, for individuals in the housing unit their response to the CCS is 'dependent' on the fact that they had been missed by the census. In the plans for the 2000 Census in the US, the pre-listing of areas in the follow-up survey prior to the 2000 Census is noted as a possible source of dependence if those individuals contacted in the pre-listing confuse this with the actual census.

In the UK in 2001, dependence between the census count and the CCS count could occur due to the use of address data on a computer system called Address Point in planning for both the 2001 Censuses and the CCS. In the 2001 Censuses of the UK, enumerators will be given an initial list of addresses in their ED based on address point. If a housing unit is missing from this list the Census enumerator should add it to

the list but it seems reasonable that such housing units will have higher census underenumeration. The dependence can then occur in the construction of postcode maps for the CCS enumerators. Postcode boundaries do not 'exist' but have to be constructed from Address Point based on the location of addresses within the postcode. The inaccuracy in Address Point that caused a housing unit to be missing from the census enumerator's list may also cause the postcode boundary constructed from Address Point to exclude it; and hence the housing unit's non-response in the CCS is not independent of its non-response in the census. To prevent such dependence between the census and the CCS, CCS enumerators are required to check whether housing units at the boundaries of their sampled postcodes are in or out of the postcodes.

The simulations in Brown *et al* (1999) introduced the concept of dependence by changing the odds ratio between the probabilities that generate whether an individual is counted in both the census and the CCS (p_{11}), just the census (p_{10}), just the CCS (p_{01}), or missed by both (p_{00}). The odds ratio is defined as

$$\gamma = \frac{p_{11}p_{00}}{p_{10}p_{01}} \qquad (4.27)$$

and this equals one when the census and CCS are independent. For each individual in the simulation data, a value is specified for p_{1+} (their overall census coverage probability) and p_{+1} (their overall CCS coverage probability). When the census and CCS are independent, p_{11} is simply the product of these two probabilities and the remaining probabilities follow. For a general odds ratio, the simulation programme solves a quadratic equation to get p_{11} and then the remaining probabilities follow. The same approach as in Brown *et al* (1999) has been used to introduce dependence in the simulations presented in this section.

The outcome of a failure to preserve independence between the census and CCS will be bias in the DSEs for the sampled postcodes. The actual value and nature of the bias will depend on the level of dependence, the coverage probability in the census, and the corresponding coverage probability for the CCS. For example, in the simulations

already presented the average census coverage probability is 95 per cent and the corresponding CCS coverage probability is approximately 88 per cent. Under dependence the relative bias of the DSE could range from –5 per cent (the CCS finds no new people – large odds ratio) through zero per cent (the CCS is independent – odds ratio = 1) to 0.7 per cent (the CCS finds all the missed people – odds ratio = 0). If the CCS coverage probability increased to 90 per cent, the range of the bias would then be -5 per cent to 0.6 per cent. If census coverage also improves to 98 per cent the range becomes –2 per cent to 0.2 per cent.

In the estimation strategy the relationship is not quite so simple. To test the actual impact of dependence, odds ratios of eight and 0.125 are applied to the individuals and households in the simulation data. (The choice of 8 and 0.125 is rather arbitrary but they do relate to $\log_e$ odds of 2 and –2, which would generally be considered extreme). For the overall coverage probabilities in the simulation population, an odds ratio of eight would generate a bias in the DSE of –2 per cent. In the simulation, the actual impact of the dependence at the estimation stage will vary from postcode to postcode and by age-sex group as the coverage probabilities also vary by these factors. Therefore, an overall odds ratio of eight will mean a negative bias in the estimation, but not necessarily –2 per cent, and an odds ratio of 0.125 will mean a positive bias. The actual impact of dependence on the estimated population total, using the robust estimation strategy, is presented in Table 4.6.

TABLE 4.6

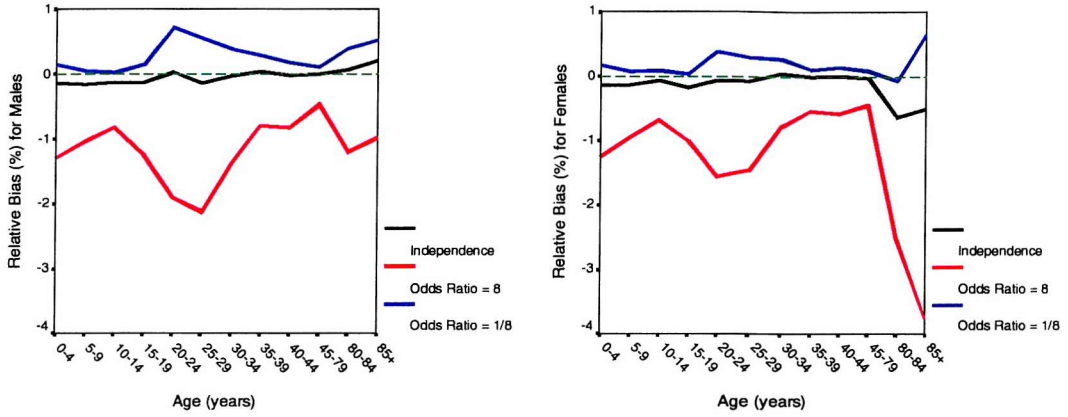
The impact of dependence in the simulation on the estimate of total population

	Relative Bias (%)	Relative RMSE (%)	Z-value for Bias
Independence	-0.07	0.48	-4.43
Odds Ratio = 8	-0.96	1.05	-74.32
Odds Ratio = 1/8	0.18	0.53	11.69

As expected, Table 4.6 shows that for an odds ratio of eight, representing the scenario where those counted by the census have a higher coverage in the CCS than those missed by the census, there is a negative bias in the estimation. However, this is less than the two per cent that may have been expected based on the overall coverage of

both the census and the CCS. Conversely, Table 4.6 shows that for an odds ratio of an eighth, representing the scenario where those counted by the census have a lower coverage in the CCS than those missed by the census, there is a positive bias in the estimation. In both cases, the dependence results in an increase in total error. Figure 4.4 gives the relative bias for the individual age-sex estimates.

Figure 4.4: The impact of dependence in terms of relative bias on estimates of the total population by age and sex based on the robust strategy



The results in Figure 4.4 demonstrate the varying impact of dependence. For young adult males, the low census coverage of these age groups means and odds ratio of eight has a much greater impact than say for females aged 45 to 79. The positive bias generated when the odds ratio is an eighth also reflects the pattern of varying census and CCS coverage across the different age-sex groups.

The results in Table 4.6 and Figure 4.4 are not startling in the sense that they simply confirm that the estimation strategy behaves as one would expect under dependence, and therefore highlight the need to preserve operational independence as much as possible. Caution is needed when interpreting the actual level of the impact. As already explained, the relationship between bias and dependence is not simple and the complexity of the estimation strategy further compounds this. In addition, and perhaps most importantly, the simulation takes an extreme level of dependence that may occur in a few isolated areas, and applies that everywhere. In other words, it really does represent a worst-case scenario.

4.4.6) Variance Estimation

So far this chapter has concentrated on the strategy to get population estimates by age and sex for each estimation area. However, estimated variances for these estimates of the population are essential to allow quality assurance with other external estimates of the population in 2001, such as those produced nationally using demographic methods. Work in Brown *et al* (1999) used the ultimate cluster variance estimator. This has a general form for the variance of an estimator $\hat{\theta}$ given by

$$\hat{\mathrm{Var}}(\hat{\theta}) = \frac{1}{n(n-1)} \sum_{g=1}^{n} \left(\hat{\theta}_g - \hat{\theta}\right)^2 \qquad (4.28)$$

where n is the number of PSUs in the sample, and $\hat{\theta}_g$ is an estimator based only on the data from PSU g.

Another common, and related, variance estimator is the jackknife estimator. This has a general form for the variance of an estimator $\hat{\theta}$ given by

$$\mathrm{Var}(\hat{\theta}) = \frac{1}{n(n-1)} \sum_{g=1}^{n} \left(\left\{n\hat{\theta} - (n-1)\hat{\theta}^{(g)}\right\} - \hat{\theta}\right)^2 \qquad (4.29)$$

where n is the number of PSUs and $\hat{\theta}^{(g)}$ is an estimate based on all the data excluding PSU g. In both (4.28) and (4.29) a finite population correction can be included but for large population sizes it will make very little difference. Both can also be generalised to allow the estimation of covariances between the estimates of the different age-sex groups. These are needed to estimate a variance on the total population estimate when it is derived from summing the estimates for the age-sex groups.

The two variance estimation techniques were applied to the simulation data for the robust ratio estimator (4.23). To do this the ED is treated as the PSU. For the ultimate cluster variance estimator this implies that $\hat{\theta}_g$ is based on data from the five postcodes in sampled ED g while for the jackknife estimator $\hat{\theta}^{(g)}$ is based on all the EDs except

the five postcodes in sampled ED g. Variances are estimated across the size strata but within the HtC strata. The number of EDs sampled from each size stratum is too small, usually only one or two EDs, to allow variance estimation within each size stratum. In addition, the estimation strategy based on (4.5) assumes the same value for R_d across all size strata within a given HtC stratum. Based on these two points it makes sense to apply variance estimation within the HtC but across the size strata. Therefore, n in both (4.28) and (4.29) refers to the number of EDs sampled within a specific HtC stratum and the overall variance is computed by summing the variances across the independent HtC strata.

As already stated estimates of the complete variance-covariance matrix are required to allow the computation of a variance for the estimate of the total population. From the simulation, the empirical variance for the estimator of the total population, which represents the true variance of the estimator of the total population, is 4,478,323. Table 4.7 gives the corresponding mean value and coverage over the simulation for the two variance estimators given by (4.28) and (4.29).

TABLE 4.7

Performance of variance estimators for the total population

Type of Estimator	Mean Value	Coverage[1]
Ultimate cluster	7,936,568	0.963
Jackknife	5,065,500	0.944

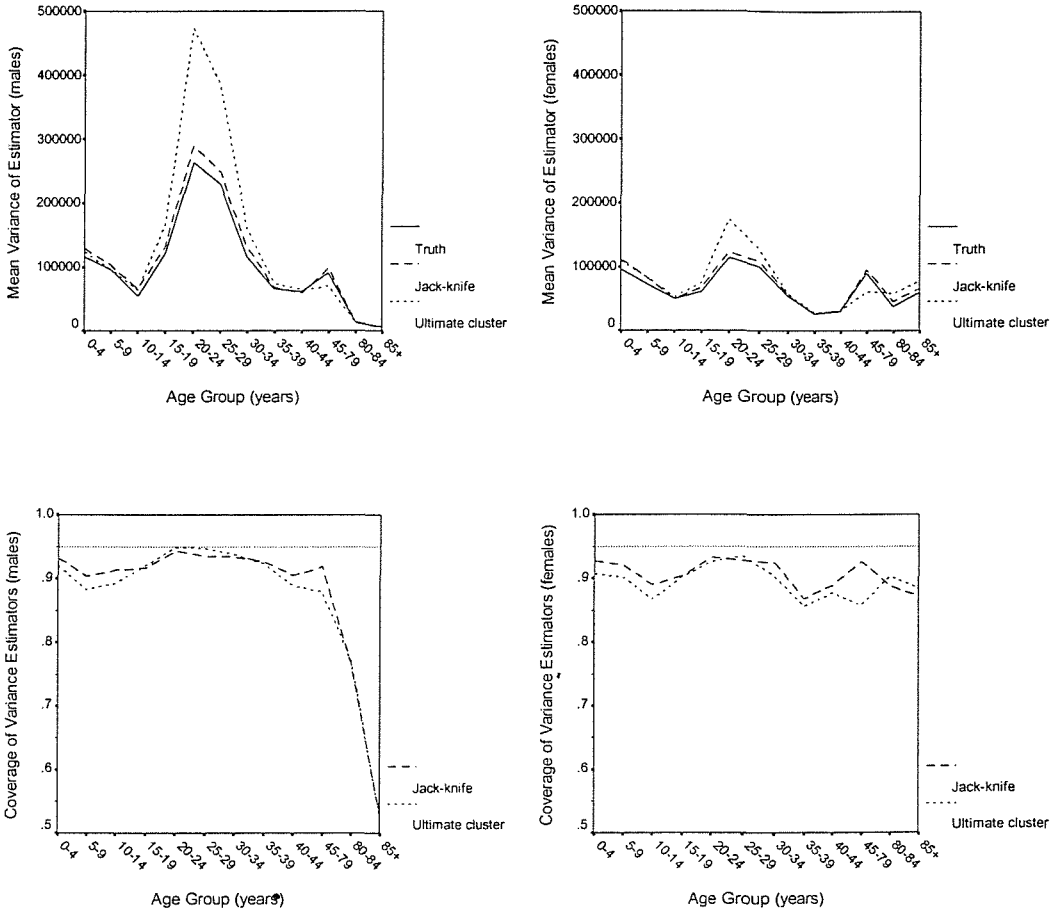
1. Based on estimated 95 per cent confidence intervals

Table 4.7 demonstrates that both estimators are conservative (positively biased) when compared to the empirical variance, and this is particularly true for the ultimate cluster variance estimator. However, in both cases the coverage for a 95% confidence interval is approximately correct, the ultimate cluster variance estimator being one per cent over and the jackknife being 0.5 per cent under. Figure 4.5 presents the same results but for the individual age-sex estimates.

Figure 4.5 shows that the jackknife variance estimator tracks the true variance more closely than the ultimate cluster variance estimator. This is particularly true for young

adult males and, to a lesser extent, the same age groups for females. For these age groups, particularly males, there is considerable underenumeration and it is intuitive that the complex estimator (4.23) will be more stable when estimated with more data. Therefore, although the two approaches are related, the ultimate cluster variance estimator, which relies on estimates based on single PSUs, appears to be more unstable and over states the variability of the estimator for certain age-sex groups.

Figure 4.5: Performance of the variance estimators for the individual age-sex estimates



Another consideration is the fact that the simple ultimate cluster variance estimator given by (4.28) is only appropriate if each cluster has the same expected value under the model used to estimate the population total. This is true for linear estimators and while the regression estimator reported in Brown *et al* (1999) and the standard ratio estimator are both linear estimators the robust ratio strategy does not, in general, lead to a linear estimator. For age-sex groups where levels of underenumeration are high,

the adjustments for outliers will in particular lead to a non-linear estimator causing problems for variance estimation based on the ultimate cluster variance estimator. One solution is to use a taylor series to approximate the non-linear estimator with a linear estimator and apply ultimate cluster techniques to the linear estimator. However, in the case of the robust strategy it is hard to see how taylor series linearisation could be applied as this would require differentiating the robust estimator.

It can still be desirable to use a slightly conservative estimator if it has better coverage. However, Figure 4.5 also shows that in fact both estimators have coverage problems in that they do not give 95% coverage for 95% confidence intervals across the age-sex groups. This is a particular problem for males aged 80 to 84 and males aged 85 and over. For these particular population groups, the cause of the problem is that for some of the generated CCS samples, the CCS fails to find any extra people over the census in the CCS sampled postcodes. Such samples lead to an estimated level of underenumeration of zero per cent with an associated estimated variance of zero. Both estimators suffer from the problem and as such situations will definitely arise in 2001, further work is needed to specify a strategy for collapsing age-sex groups to allow variances to be estimated. Overall, taking the results of Table 4.7 with Figure 4.5, the jackknife estimator performs best at the total population level and across the age-sex groups. This is with respect to both unbiased estimation of the variance and reasonable coverage of confidence intervals.

4.4.7) Implementing the Robust Estimation Strategy in 2001

The work presented in this section has taken the basic estimation strategy outlined in section 4.2, and evaluated by the simulation study in section 4.3, and looked at methods to make the strategy more 'robust' against problems that may, and will, occur with the CCS data in 2001. In particular, the strategy applies an approach similar in ethos to that of M-estimation to the prediction problem of true counts for non-sampled postcodes. The approach taken was based on option (b) in section 4.4.2. It was suggested that this approach be compared to an approach with a similar ethos to (c) such that all postcodes with census counts beyond the range of the sample data have a constant adjustment for underenumeration. In other words they all have the same

influence on the estimated underenumeration regardless of the actual value of the census count. This can be thought of as a line that is parallel to the Y=X line; the vertical distance between the two lines is determined by the slope of the ratio line at the point of the largest CCS postcode.

In general M-estimators based on influence functions like (b) have a greater negative bias than those based on (c). However, the advantage of this is usually a gain in the overall error (Mean Square Error) from reductions in variance due to the approach having a greater impact on the extreme values that increase variance. The counter argument to this is the intuitive appeal of the approach based on (c), as census users may be unhappy with an estimation strategy that essentially assumes zero underenumeration in postcodes with a large census count. Further simulations presented in ONS (2000d) indeed show that this second approach introduces less negative bias, the relative bias at the total population level is 0.06 per cent compared to -0.07 per cent in Table 4.6. Across the age-sex groups the relationship with respect to bias is also the same, one approach with a slight positive bias, one with a slight negative bias. With respect to total error measured by the relative RMSE, ONS (2000d) confirms that the approach outlined in section 4.4.1 does better for all age-sex groups although the gains are slight. Based on the simulation evidence, ONS are planning to implement the revised approach as it was felt that the intuitive appeal of the revised approach outweighed the very slight efficiency gains of the original approach.

4.5) Additional Considerations

Up to this point, the work presented here has to some extent ignored some of the more difficult issues surrounding the actual implementation of dual-system estimation. This section briefly considers two of those issues. The first is the handling of people who move between the 2001 Census day and the day the CCS interviewer attempts to contact the residents. The second is the issue of overenumeration in the 2001 Census. In addition, the work presented here has concentrated on the estimation strategy for estimation areas. In general, more than one LAD constitutes an estimation area so to

obtain estimates by age and sex for each LAD requires an additional phase that is briefly discussed here.

4.5.1) Movers and the CCS

Ideally the 2001 Census and CCS would take place on the same day or certainly during the same period as they are attempting to enumerate the resident population with respect to the same day. This is not feasible from an implementation point of view but is also not possible from a statistical point of view as the two collections must be independent. Therefore, if due to practical and statistical constraints the CCS cannot be in the field at the same time as the 2001 Census, it is desirable that the CCS is in the field as soon after the census as possible. Current plans in the UK suggest about four weeks after the 2001 Census day. This is because the CCS should be attempting to count the same population as the census. Inevitably people will move and therefore, when the CCS interviewer arrives in a sampled postcode some of the people who were resident on census night will no longer be resident (out-movers), and in some instances, the out-movers will have been replaced by new residents who were not there on census night (in-movers). Griffin (2000) considers three possible ways to deal with movers that were proposed first by Marks (1979).

a) From current residents collect proxy information to construct the resident population in the postcode as per census night. The US Census Bureau refers to this as procedure A.

b) Construct the population of the postcode as per the time of the CCS interview and for in-movers collect additional information relating to their residence on census night. At the matching stage look for a census record for the individuals concerned at the alternative location provided by the respondent. The US Census Bureau refers to this as procedure B.

c) At households with movers, collect basic proxy information on the out-movers and detailed information on the in-movers. The proxy information on out-movers is matched to the census returns within the sampled area and the achieved match-

rate is applied to the in-movers in the calculation of the post-stratum DSEs. The US Census Bureau refers to this as procedure C.

The analysis in Griffin (2000) treats the issue of movers as a potential source of heterogeneity. In particular, it is argued that the response rate for data collected on out-movers via proxy information in the follow-up survey will be lower than that achieved for non-movers. This is a disadvantage of procedure A. In 1990 the Census Bureau used procedure B, the argument in favour of procedure B being that the survey just needs to enumerate the current residents. This was not without problems in terms of matching to census records in other locations but this is the preferred approach of the US Census Bureau. The problem in the 2000 Census was that the original plan for data collection in the Census would have prevented the use of procedure B, hence the development in Griffin (2000) of procedure C. The paper argues that, assuming the match-rate estimated for out-movers is an unbiased estimate of the match-rate that would have been achieved for the in-movers if the matching could be done for in-movers, procedure C will be similar to procedure B.

The problem in the UK context is that DSE estimation takes place in individual postcodes and clusters of postcodes enumerated in the CCS sample rather than at some aggregate level with the survey data weighted-up to represent the population at that level. The strategy relies on the DSE being a good estimate of the population in the postcode on census night. This would make the application of either procedure B or procedure C problematic. Procedure A is a possibility and the CCS form has been designed to allow the collection of proxy data on out-movers. (The reality is that most data in the CCS is proxy data as the interviewer obtains the information on all the residents in the household from a single household member.) There are still two outstanding issues. The first is a concern regarding the quality of proxy data in the CCS when the entire household has moved and therefore, none of the individuals resident on census night are available to respond to the survey. The second is the legal position of collecting such data and processing it, when the survey is voluntary. The final decision with respect to the treatment of movers is still under consideration by ONS. It may well be that the treatment of movers in the 2001 CCS will rely heavily on

the short period between the Census and CCS, with out-movers being treated as missing at random non-response in the CCS.

4.5.2) Overenumeration in the 2001 Census

In the UK, the general perception of overenumeration is that compared to underenumeration it is a 'second-order' problem. The last reliable estimates, reported in Britton and Birch (1985) for the 1981 Census, estimated gross underenumeration as 0.62 per cent against gross overenumeration of 0.17 per cent. The evidence available, and general perception, suggests that while underenumeration increased quite dramatically in the 1991 Census overenumeration basically remained unchanged from the 1981 Census and therefore becoming a secondary issue. (Following the 1991 Census, imputation was carried out for completely missed households and there is evidence that shows that this imputation did introduce too many people (see Diamond, 1994). However, if a similar imputation exercise is performed in 2001, this potential source of overenumeration is easily handled by excluding the imputed data from the estimation of the true population.) The UK scenario contrasts with the situation in the US where overenumeration is a more serious issue. Hogan (1993) gives a brief discussion of gross errors and Dunstan *et al* (1999) report that gross underenumeration was estimated at 4.7 per cent balanced against an overenumeration (made-up of a range of erroneous enumerations including census imputations, fictitious census returns, and duplicates) of 3.1 per cent. Therefore, the PES in the US explicitly tackles estimation of overenumeration through the E-Sample and P-Sample design (see Hogan, 1993). The P-Sample is essentially equivalent to the CCS but the E-Sample is a sample of census returns that are check for fictitious data and erroneous or incorrect enumerations. This estimates an adjustment for overenumeration that is applied in the calculation of the DSE.

The current plan in the UK is to assume that the CCS will count people in the correct location as per their interpretation of the census definition. The CCS will also collect data on possible locations where individuals could have been counted in the census. This is similar to the Australian approach. In the UK the data processing will not allow this information to be used directly in the estimation but it is envisaged that

towards the end of data processing a secondary matching exercise will be able to give estimates for broad sub-groups and regions of the extent of overenumeration. The current working assumption is that this will be a minor issue but as a result of the exercise and the quality assurance programme, adhoc adjustments may be made to some of the age-sex estimates in a few estimation areas.

In the context of the estimation strategy, not adjusting for overenumeration in the census will inflate the DSEs for the sampled postcodes. The US approach estimates the overenumeration and corrects for it in the calculation of the DSE. In the UK context, ignoring overenumeration will not be a problem when estimating the ratio between the true count and the census, provided its impact is reasonably constant across the group for which the ratio is being estimated, as it will inflate the numerator and denominator by approximately the same factor. (If overenumeration is constant it cancels top and bottom in the estimation of the ratio.) The problem then occurs if this ratio is applied to the total census count including overenumeration as the ratio measure gross and not net underenumeration. If it is found to be a problem, adhoc adjustments could then be made to the census count. While this is a possible framework to deal with a minor overenumeration problem it still requires further investigation and the final decision on the treatment of overenumeration is still to be taken by ONS in the coming months.

4.5.3) Estimation for Individual LADs

As stated at the beginning of chapter three, the main aim of the CCS is to provided the basis for the mid-year population estimates by age and sex at the LAD level, adjusted for census underenumeration. This chapter applies to estimating the population in estimation areas, groups of LADs for which the CCS sample is considered sufficiently large to yield high quality direct estimates. Considerable additional research has been undertaken by ONS to assess different small area estimation strategies for obtaining LAD estimates. The chosen strategy, and the research that supports that choice, is presented in ONS (2000e). The basis of the approach is a synthetic estimator that assumes that the underenumeration patterns at the estimation area level apply in each of the LADs that constitute the estimation area. This is supplemented by an 'LAD

effect' that allows for the pattern to vary slightly from LAD to LAD based on the CCS data collected in each LAD.

4.6) Concluding Remarks

The work presented in this paper has shown that the proposed estimation strategy of combining dual system estimation with ratio / regression estimation techniques works well and compares favourably with the approach outlined in Wolter (1986) that was used by the US Census Bureau in 1990. The basic application of the strategy presented in section 4.2 and assessed by a simulation study in section 4.3 is not without problems. The DSE has a negative bias when applied to small areas of aggregation, such as the postcode by age-sex group. The standard ratio and regression estimators suffer from a positive bias due to model failure and outliers. Section 4.4 develops a robust strategy that addresses the problems and further simulation results suggest that the adjustments are successful in reducing the bias and variance of the estimator. Section 4.4 also addresses the important issue of variance estimation and the simulations support the use of a jackknife variance estimator.

The strategy has specifically focused on the estimation of the population for each age-sex group ignoring all the others. The justification for this is a preference for simplicity in the underlying approach to estimation. However, it seems plausible that there would be scope for further improvements in efficiency when estimating some age-sex groups by using information from other age-sex groups. For example, the number of young children is likely to be related to the number of young women. 'Borrowing strength' in this way can be achieved by the inclusion of additional auxiliary variables in either the ratio or regression model. However, the estimator no longer has a simple interpretation with the model parameters simply relating to the rate of underenumeration for each specific age-sex group.

Some of the discussion in section 4.5 points to the fact that the work in this chapter has been about creating a theoretical framework for the estimation and an overall strategy. Current work at ONS is dealing with the outstanding practical and theoretical

issues, such as movers and overenumeration, and working towards implementation of the estimation strategy for use on CCS data following the 2001 Census.

Appendix 4.1 – Creating census and CCS coverage probabilities for the simulation population.

The basic construction of the coverage probabilities for the simulation data is given in section 4.3.2. The basis of the individual probabilities are the EwC ED adjustment factors and these vary at an individual level with an underlying structure that depends on age, sex, and the HtC category of the ED. In addition, these are adjusted based on economic status measured by the census variable 'Primary Activity Last Week'. This is done by applying a power to the base probability that is greater than one to reduce coverage and less than one to increase coverage. Assume that all individuals have the same base probability p = 1 - q. If the power a_i is applied to each individual i then

$$p_i^* \cong (1 - a_i q) \text{ and } \frac{1}{N}\sum_{i=1}^{N} p_i^* = 1 - \frac{q}{N}\sum_{i=1}^{N} a_i$$

which equals p provided the average of the a_i's is one. In other words, the overall coverage remains approximately the same but across individuals it now varies by a new variable. In the simulation data the following powers are used for economic status.

Individual under 16	1
Full-time employee	0.629
Part-time employee	0.629
Employer	0.629
Self-employed	0.629
Government training scheme	2
Waiting to start a job	2
Unemployed	3
Full-time education	2
Unable to work	1
Retired	1
Homemaker	1
Other inactive	1

These average to one over all the individuals in the simulation data. The base probabilities for census coverage of households are constructed by averaging the probabilities for the individuals within each household. For households with tenure of private rented the coverage is reduced by a factor of 0.95. Further variation is introduced by using the power 2/size so households of size one have a lower coverage

and large households have a higher coverage. Note that this does not mean that within a large household the census necessarily counts the individuals correctly.

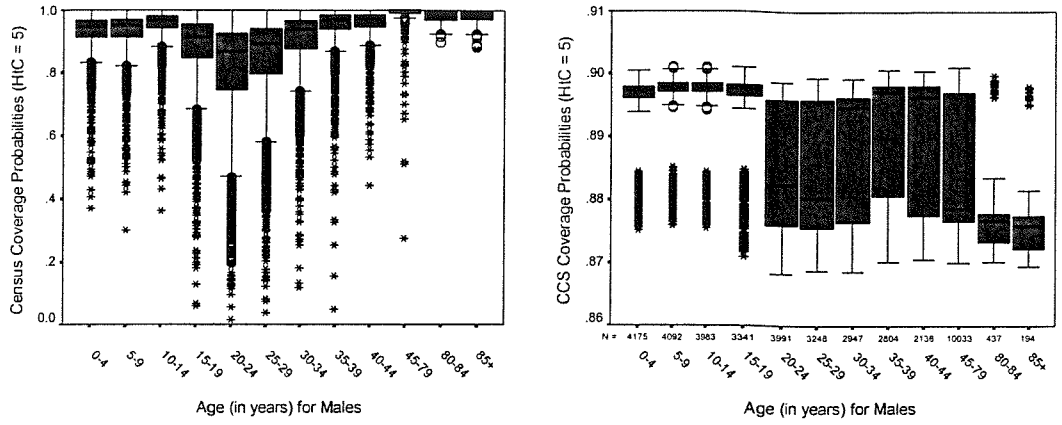
For the CCS coverage rates a constant base is set for households and then individuals in counted households. For households this base is adjusted using powers that vary by household size

One person household	1.1 +/- e
Two person household	1.05 +/- e
Three person household	1 +/- e
Four plus person household	0.83 +/- e

where e ~ N(0, 0.01) so that there is some extra heterogeneity at low levels of aggregation. As with the census, smaller households are harder to count. These powers approximately average to one over all the households so the approximate household coverage is whatever is set in the simulation. For individuals the base is adjusted using powers that vary by age and sex.

Age	Males	Females
0-4	1	1
5-9	0.98	0.98
10-14	0.98	0.98
15-19	0.98	0.98
20-24	1.1	1.05
25-29	1.075	1
30-34	1.075	1
35-39	0.98	0.98
40-44	0.98	0.98
45-79	0.98	0.98
80-84	1	1
85+	1.05	1.075

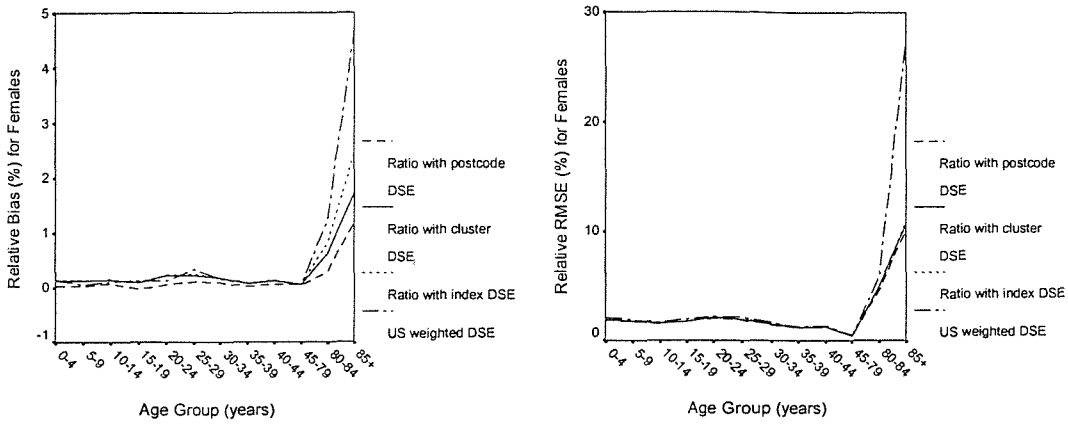
The following graphs give the variation in the individual coverage probabilities in the census and the overall coverage in the CCS for males in HtC category five after controlling for age. (The base CCS coverage is set at 0.9 for households and 0.98 for individuals and the overall coverage is the product of the two). The important point is that for the census there is considerable heterogeneity and in the CCS there is enough to ensure that the homogeneity assumption will not be completely met in the simulations at the level the DSE is applied.



Appendix 4.2 – Census coverage by age-sex group for the 100 simulated censuses

Sex	Age (years)	Mean Census Coverage (%)
Male	0-4	92.74
Male	5-9	93.79
Male	10-14	95.65
Male	15-19	91.23
Male	20-24	84.09
Male	25-29	85.47
Male	30-34	91.31
Male	35-39	95.12
Male	40-44	95.95
Male	45-79	98.38
Male	80-84	94.87
Male	85+	94.91
Female	0-4	93.37
Female	5-9	94.65
Female	10-14	96.42
Female	15-19	94.90
Female	20-24	92.23
Female	25-29	93.12
Female	30-34	96.28
Female	35-39	98.02
Female	40-44	97.72
Female	45-79	98.62
Female	80-84	91.62
Female	85+	84.26

Appendix 4.3 – Ratio estimators combined with weighted and unweighted dual system estimation for females by age



Chapter Five – CCS Design and Estimation in Northern Ireland

5.1) Introduction

The history of recent census underenumeration in Northern Ireland is different from the rest of the UK. For Great Britain, the 1981 Census was considered a successful census while the 1991 Census is judged to have had problems with underenumeration (see section 2.3). In Northern Ireland the reverse is generally considered to be the case. However, the success of the 1991 Census does not mean there was zero underenumeration and therefore it is important that Northern Ireland has a CCS in 2001 to check the coverage of the census and adjust for any estimated underenumeration, particularly as previous surveys in the UK have excluded Northern Ireland.

The make-up of Northern Ireland, both in political and geographic terms, is also rather different from the rest of the UK. It has a total population of just over 1.5 million people, and is the smallest of the countries that constitute the UK. Within Northern Ireland there are 26 local government districts (LGDs) and while they play a role in government they do not have the same relevance as LADs in the rest of Great Britain. The exception is Belfast, with a population of around 300,000 people.

The population of Northern Ireland implies that three estimation areas would be appropriate based on the fact that the total population is approximately 1.5 million persons and the design strategy is based on estimation areas with populations of approximately 0.5 million persons. The LGDs can be grouped into three based on combining a five level standard classification. This creates one estimation area for Belfast, a second for the LGDs that surround Belfast in the 'East' of Northern Ireland, and a third for the more rural LGDs in the 'West' of Northern Ireland. This chapter first considers the implementation of the CCS design discussed in chapter three to Northern Ireland. The chosen approach is assessed through a simulation study similar to the approach used in chapter four but based on the 1991 Census data for Northern Ireland. Finally, an approach is developed and tested using a simulation for the estimation of census underenumeration for other variables apart from age and sex.

This includes estimation for variables that relate to households as well as other variables that relate to individuals.

5.2) The CCS Design Applied in Northern Ireland

5.2.1) Classification Index for EDs

The design outlined in chapter three uses an index to stratify EDs into different types based on the underenumeration patterns observed in the 1991 Census. There is much less information on the patterns of underenumeration within Northern Ireland from the last census to replicate such a strategy exactly. However, an index can be constructed by considering certain unique features of Northern Ireland. Within Northern Ireland religion is an important variable. The two communities are still polarised in many areas meaning that approximately two thirds of EDs were dominated either by Protestant families or by Catholic families in the 1991 Census. This pattern is still expected to be true in 2001. The structure of each of the LGDs still tends to be a town, that is the administrative centre of the LGD, surrounded by its rural hinterland. Finally, each ED can be classified as deprived or not deprived based on the 1991 Census. These three factors can be combined to produce an eight-way classification of EDs.

There is little evidence to directly link the three factors to patterns of underenumeration in the 1991 Census. However, using data from the 1999 Census Rehearsal, the eight-way classification of the factors can be ordered based on census response rates achieved in the rehearsal. The ordering is given in Table 5.1 from the highest expected census response rate to the lowest. The distribution of the population is not even across the eight categories with the last three categories containing most of the population. Therefore, while it is desirable to spread the sample over all the eight categories to ensure the sample contains all types of EDs it will not be possible to estimate independently in all eight. This is because the small number of EDs in some categories would require very large sampling fractions to allow efficient estimation, the consequence being that when the total sample size is fixed other much larger areas would have their samples reduced, the overall effect being less efficient estimation. Instead, estimation will use a three level categorisation that combines the categories in

Table 5.1. Levels one to five will form an 'easy to count' group containing about 33 per cent of the population, levels six and seven will form a middle group containing about 50 per cent of the population, with level eight forming a 'hard to count' group. As a consequence, the three level structure also needs to be reflected in the design to ensure there is sufficient sample to allow this.

TABLE 5.1

Ranking of the ED classification index

Level of Index	Religion	Location	Deprivation Status
1	Protestant	Rural	Not Deprived
2	Protestant	Rural	Deprived
3	Catholic & Mixed	Rural	Not Deprived
4	Protestant	Urban	Not Deprived
5	Protestant	Urban	Deprived
6	Catholic & Mixed	Urban	Not Deprived
7	Catholic & Mixed	Rural	Deprived
8	Catholic & Mixed	Urban	Deprived

5.2.2) CCS Design for Northern Ireland

The design for Northern Ireland is considered as a whole rather than a series of designs for each estimation area. To be consistent with a sample of approximately 300,000 households in England and Wales, the corresponding sample size for Northern Ireland would be about 10,000 households. Assuming 15 households per postcode and five postcodes per ED this specifies an ED sample of between 130 and 135 EDs. The basic strategy outlined in section 3.5 is applied to the three estimation areas; within each estimation area the EDs are stratified by the ED classification index outlined in Table 5.1, then by size using the multivariate approach of section 3.5.2. The major difference now occurs in that the approach in section 3.5 used optimal allocation while in Northern Ireland proportional allocation is used to distribute the ED sample across the strata. A sample of postcodes is then chosen from the selected EDs. Proportional allocation is chosen as there is much less information, with respect to underenumeration in the 1991 Census, on which to base any assumptions about the

distribution of underenumeration. Therefore, it is safer to spread the sample evenly over all types of ED.

Starting with a target sample of approximately 130 EDs and a fixed set of strata for Northern Ireland the allocation proceeds in two stages. The first stage allocates the sample to the three estimation areas, and within that the three levels of the collapsed ED classification index proportional to the total population as recorded by the 1991 Census. At this stage population is used, rather than number of EDs, as the West design group has a large number of rural EDs and therefore a lower population in relation to the number of EDs when compared to the other two estimation areas. The allocation is subject to a minimum sample of eight EDs so that there is considered sufficient sample to support the use of ratio estimation within the three levels of the ED classification by estimation area. The minimum sample constraint forces the sample to eight in two levels for the Belfast estimation area and one level of the East estimation area. This specifies a fixed sample size for each of the nine groups, the three estimation areas by the three levels of the index. The second stage then allocates this fixed sample for each group to the full eight categories of the ED classification index, as per Table 5.1, and then strata defined by the multivariate approach proportional to the number of EDs. The resulting design using the 1991 Census data for Northern Ireland is given in Table 5.2.

Table 5.2 specifies a sample for Northern Ireland of 130 EDs from 3,725. In addition to this the multivariate specification of strata classified four EDs to a completely enumerated stratum based on their 1991 population counts so the final sample is 134 EDs from the 3,729 EDs. The use of proportional allocation initially based on population counts rather than number of EDs means that the sampling fraction with respect to number of EDs is slightly lower in the West estimation area due to the large number of rural EDs with small population counts. In addition the minimum sample constraint increases the sample in Belfast. However, the use of proportional allocation has led to a sample that is reasonably evenly spread across the three estimation areas and the different types of EDs.

TABLE 5.2

Specification of the CCS for the Belfast estimation area

ED Classification Index

Full	Collapsed	Number of EDs	Sample
4	1	130	5
5	1	154	6
6	2	131	8
8	3	152	8

Specification of the CCS for the East estimation area

ED Classification Index

Full	Collapsed	Number of EDs	Sample
1	1	199	6
2	1	80	3
3	1	83	3
4	1	540	17
5	1	122	4
6	2	234	9
7	2	145	5
8	3	121	8

Specification of the CCS for the West estimation area

ED Classification Index

Full	Collapsed	Number of EDs	Sample
1	1	79	2
2	1	139	4
3	1	127	4
4	1	79	2
5	1	31	1
6	2	233	7
7	2	694	17
8	3	252	11

5.3) Northern Ireland Simulation Study

The design strategy outlined in section 5.2 differs slightly from the strategy in chapter three due to the fact that Northern Ireland is somewhat different from England and Wales. Therefore, to see the effect, if any, of these slight differences on the estimation strategy outlined in section 4.2 a simulation study similar to section 4.3 is used but based on Northern Ireland data.

5.3.1) CCS Design for the Simulation Population

The basis for the simulation study is the set of anonymous individual census returns for the 1991 Census in Northern Ireland. The CCS design as outlined in section 5.2 is applied to the simulation population. However, as with the simulation study in section 4.3, the CCS design is not based directly on the ED counts in the simulation population. This is because the actual design uses 1991 data but the CCS will be conducted in 2001. Instead the design is based on ED counts derived from the simulation population but adjusted to represent the ten year gap. The adjustment is based on the proportionate changes to the LGD populations of Northern Ireland between 1981 and 1991, and is constant across EDs within the same LGD. The consequence of this is that while the simulation population corresponds approximately to the data used for the design in Table 5.2, the CCS design for the simulation is based on data that does not. Therefore, while the simulation design is indicative of Table 5.2, the total sample is still 134 EDs and the estimation area samples are similar, the actual design is not identical. There are also slight differences in the distribution of the ED classification index.

5.3.2) Running the Simulations

The simulation study developed in section 4.3 utilised a set of probabilities that defined the coverage of individuals in each simulated census. These were constructed from studies of census coverage following the 1991 Censuses. However, no such studies were undertaken for Northern Ireland. Therefore, to generate probabilities for the Northern Ireland simulation, a model was fitted to the individual probabilities defined in section 4.3. The model was then used to generate three census response

probabilities for individuals that represent three levels of census coverage in the Northern Ireland data. The probabilities vary by age, sex, and the ED classification index. In addition, LGD effects were also introduced and in particular the census coverage for Belfast was reduced relative to the other LGDs. Household probabilities were constructed from the individual probabilities, as outlined in section 4.3 and appendix 4.1, by averaging the probabilities for the individuals and adjusting them for tenure and household size.

For the simulations, CCS coverage is fixed at 90 per cent for households and 98 per cent of individuals within counted households. These then vary across households and individuals using the same approach as in appendix 4.1. The simulations then proceed in the same way by generating 100 independent censuses with 10 independent CCS samples generated for each census. The estimation strategy using the cluster level DSE with ratio estimation, defined by (4.8), is applied to each set of sample data and the whole process is repeated for the three levels of census coverage. For the high census coverage this is compared to results from applying the robust approach developed in section 4.4 and the estimator defined by (4.23).

5.3.3) Results

When analysing the simulation results, as with the analysis in section 4.3.3, the truth is known as the simulation is based on a known population. Therefore, it is possible to calculate the relative bias (4.19) and relative root mean square error (RMSE) (4.18) for the estimators over the 1,000 iterations of the simulations. The results in Table 5.3 are for estimating the total population of Northern Ireland at three different levels of census coverage using the 'standard' estimation strategy of the cluster level dual-system estimator (DSE) with ratio estimation.

TABLE 5.3

Results of the simulation study for estimates of the total population

	Performance of the Estimator	
Census Coverage (%)	Relative Bias (%)	Relative RMSE (%)
92.55	0.39	1.10
95.38	0.22	0.72
97.26	0.12	0.46

The results in Table 5.3 demonstrate that, as with the results seen in Table 4.3 of section 4.3.3, the use of ratio estimation combined with dual-system estimation at the cluster level generates a positive bias in the estimator. However, the results in Table 5.3 demonstrate that as census coverage increases, in this case from 92.55 per cent to 97.26 per cent, the bias decreases from 0.39 per cent to 0.12 per cent. In addition, the decrease in the relative RMSE from 1.10 per cent to 0.46 per cent also reflects a drop in the variance. It is reasonable to expect that as the census coverage approaches 100 per cent both the variance in the DSE and the variance in the ratio model will be reduced. In addition, as census coverage approaches 100 per cent the ratio between the truth and the census will approach one. Situations where the ratio model will not be robust will also decrease, contributing both to the drop in bias and the drop in variance.

Figure 5.1 considers the performance of the estimator by age and sex, by plotting the distribution of errors over the simulation. The plots demonstrate the phenomenon also seen in Table 5.3. As the census coverage increases the estimator becomes less variable. This is reflected in Figure 5.1 by the fact that the box plots are more tightly located around zero as census coverage increases. In addition, the impact of outliers and extreme values for the estimator decreases as census coverage increases.

The equivalent error plots for the census are in appendix 5.1. The pattern of high underenumeration in the census for young men, which can be seen in appendix 5.1, is reflected in Figure 5.1 by the more variable nature of the estimator, particularly for the ages 20 to 29. This demonstrates the trade-off between using the adjusted data rather than the census. The estimation strategy significantly reduces the error in terms

of bias but the price is an increase in variance. As the underenumeration in the census increases, the variance of the estimated population totals also increases. The estimator is also more variable for the oldest ages for both men and women. For women, the high underenumeration is partly responsible but it also reflects the fact that the group is relatively small in population terms and the ratio estimator, as discussed in chapter four, will be more variable when there is little data in the sample on which to base estimation. This problem of small populations is an even greater problem for males at the oldest ages.

The results so far have only considered estimation at the Northern Ireland level. This is of particular interest for Northern Ireland where the LGDs are of less political importance and government of the Province is more centralised. However, the estimation area results are also of importance, especially as these amalgamate to form the Northern Ireland estimates and are the basis for getting LGD estimates. Table 5.4 gives the results for the total population by estimation area from the simulations with the highest census coverage.

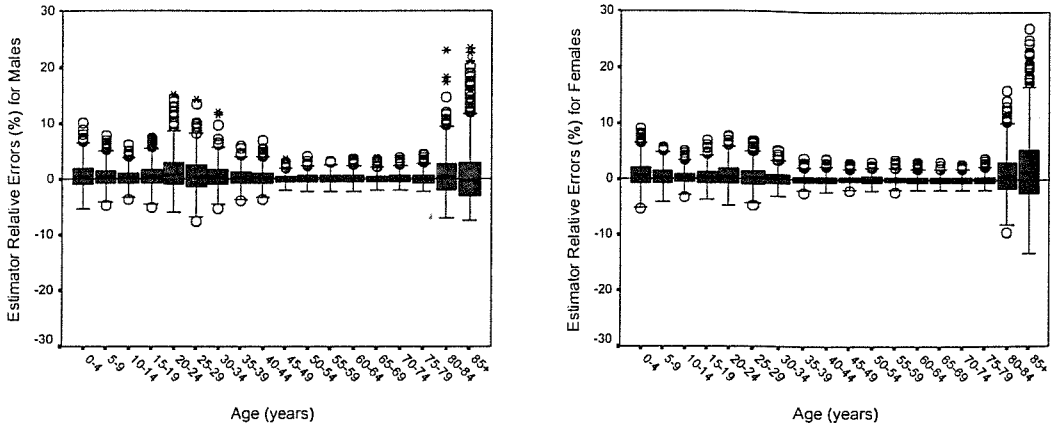
TABLE 5.4

Results of the simulation study for estimates of the total population by estimation area

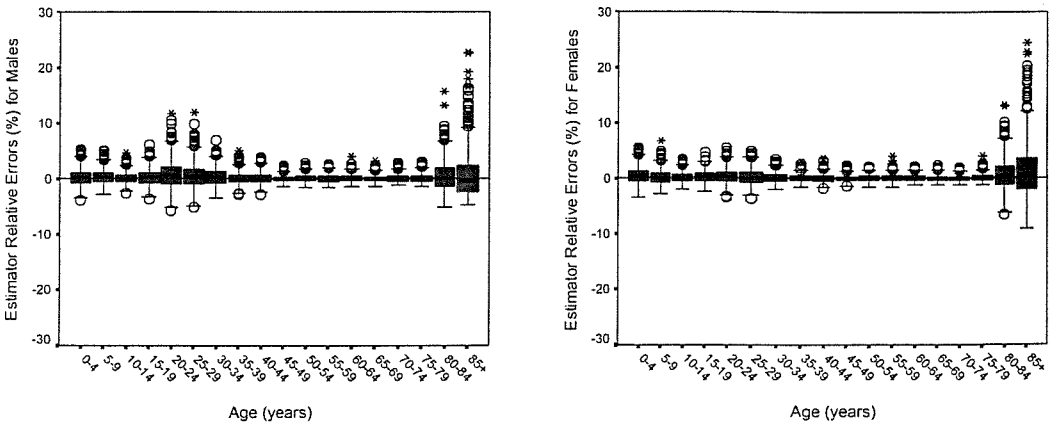
		Performance of the Estimator	
Estimation Area	Census Coverage (%)	Relative Bias (%)	Relative RMSE (%)
Belfast	92.51	0.19	1.74
East	98.68	-0.02	0.39
West	97.81	0.24	0.75
Northern Ireland	97.26	0.12	0.46

Figure 5.1: Distribution of the errors for the estimator of the total population of Northern Ireland by age and sex for three levels of census coverage

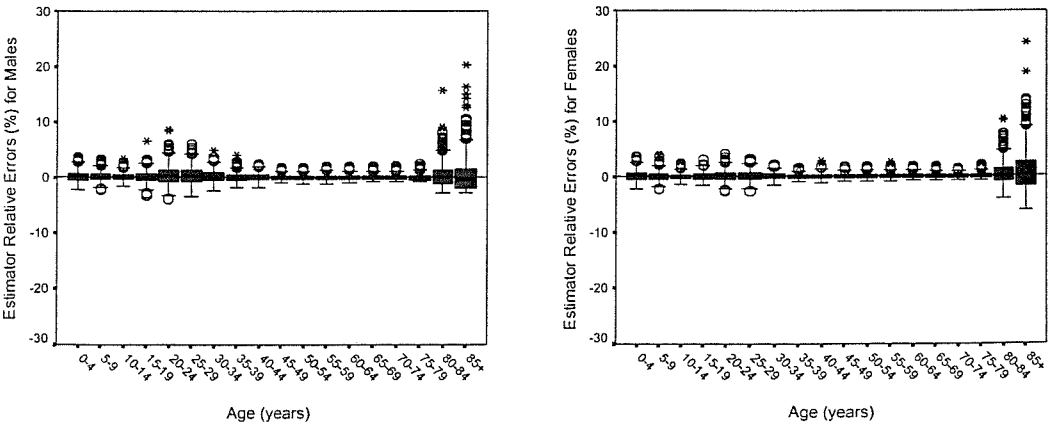
Low census coverage



Medium census coverage



High census coverage



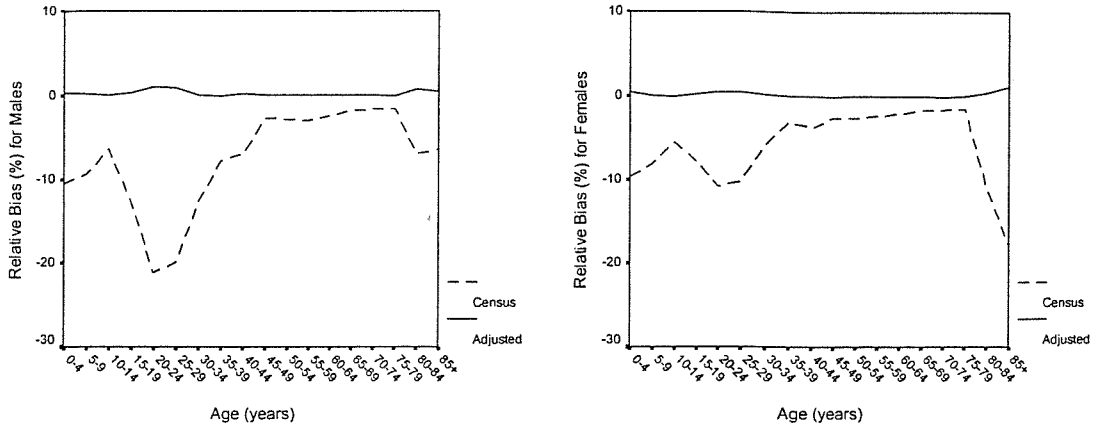
The results by estimation area in Table 5.4 are from the simulation and do not necessarily reflect real differences between the different estimation areas. Part of the difference will be due to LGD effects introduced into the simulation. However, some of the differences do reflect the variable distributions of EDs by the classification index. In particular, the West estimation area has a large number of rural Catholic EDs and Belfast has none of the groups that are considered the easiest to count by the classification given in Table 5.1. Therefore, these differences do partly represent the pattern that is expected in 2001.

Table 5.4 demonstrates that for the estimation area with the best census coverage, the East estimation area, the standard procedure of combining the DSE with ratio estimation does not result in a bias as observed in Table 4.3 of section 4.3.3 and the other two estimation areas. This links with the results reported in Table 5.3 and Figure 5.1 for all of Northern Ireland. As census coverage increases the bias and variance of the estimator decrease. In addition, the East estimation area has the largest overall sample. The results suggest that the use of the robust adjustments to the ratio estimator developed in section 4.4 will have little impact in the East estimation area, but there is the possibility of improvement in the other two estimation areas.

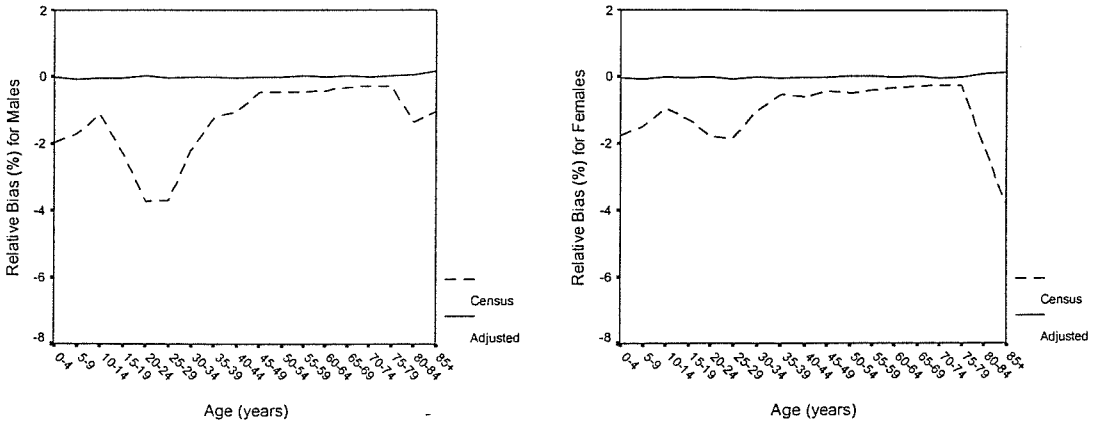
Table 5.4 gives results at the total population level but this can hide important patterns across estimates by age and sex. Figure 5.2 presents the relative bias of the counts adjusted using the standard estimation strategy by age and sex for the three estimation areas and Figure 5.3 presents the relative RMSE. The results in Figure 5.2 demonstrate the much higher levels of census underenumeration at all ages and for both sexes in Belfast. However, the estimation strategy corrects for the bias but gives a slight positive bias for the young men aged 20 to 29 and the oldest women. This reflects the very high underenumeration in the census, over 20 per cent for males aged 20 to 24, and follows the same patterns seen in Figure 4.1 of section 4.3.3. This suggests that the robust approach developed in section 4.4 should be effective at reducing the positive bias.

Figure 5.2: Relative bias of the census data and the data adjusted using the standard estimator by age and sex for the three estimation areas

Belfast estimation area



East estimation area



West estimation area

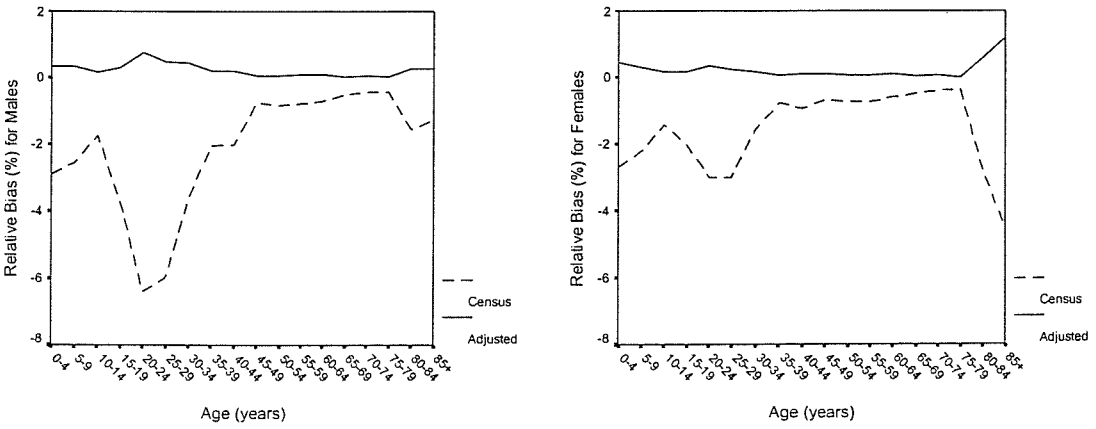
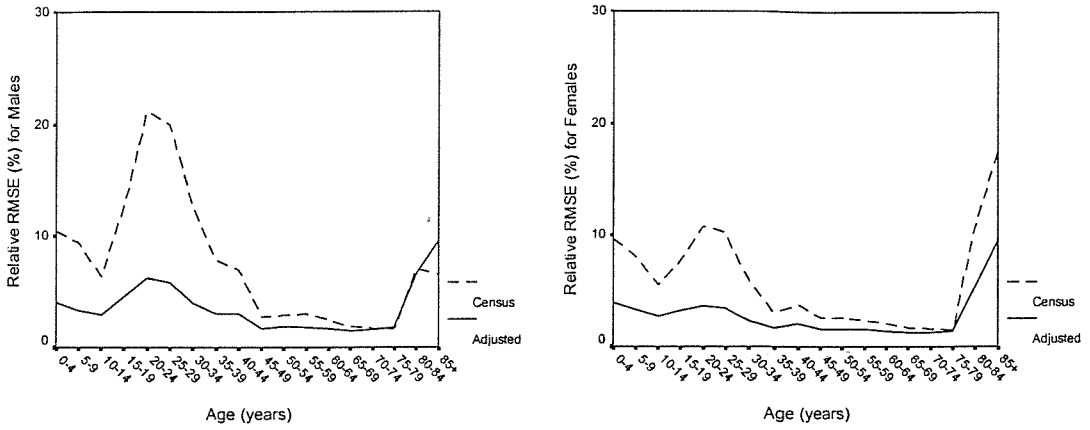
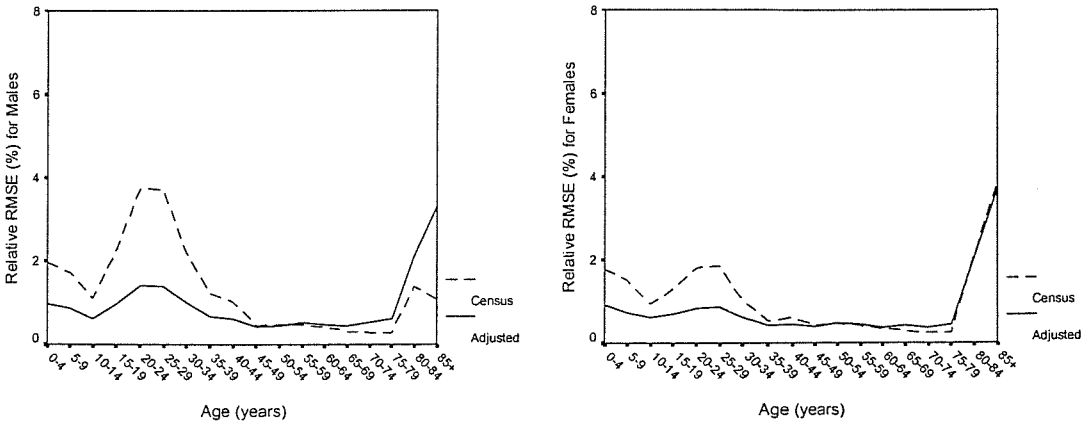


Figure 5.3: Relative RMSE of the census data and the data adjusted using the standard estimator by age and sex for the three estimation areas

Belfast estimation area



East estimation area



West estimation area

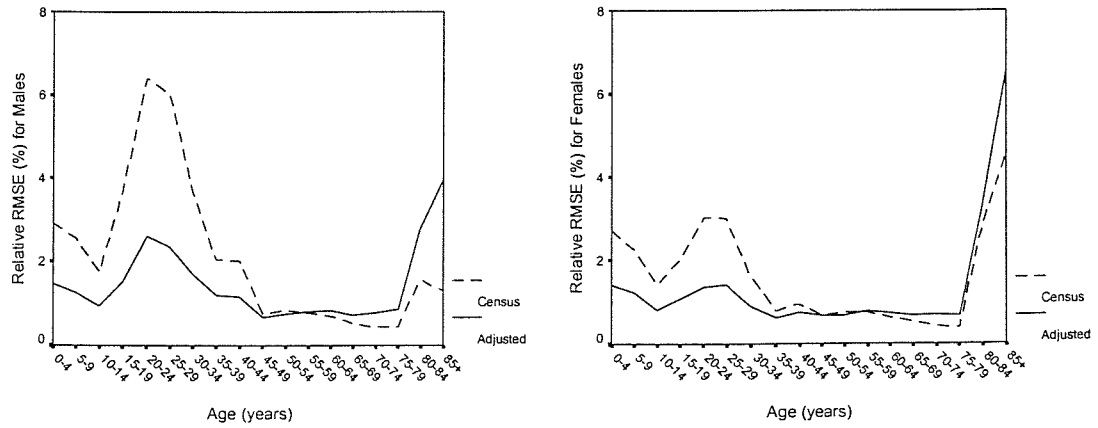


Figure 5.2 shows that the level of census underenumeration in the East estimation area is much lower but the estimation strategy still corrects for it and as Table 5.4 suggests it is effectively unbiased across all ages for both males and females. However, Figure 5.3 does demonstrate that when the level of census underenumeration is very low the total error measured by relative RMSE can be higher than for the census. This is the case in the East estimation area for people aged over 55 years. One solution to this would be to collapse across age groups when the level of census underenumeration is very low and reduce, in relative terms, the variance of the estimated population totals. This is the strategy adopted in the simulations in chapter four where a single age group is used between 45 and 79 years. However, to produce the age-sex distribution by five-year groups would then require the use of synthetic estimates. The second solution is to accept that in terms of total error, bias and variance, the estimator will not always be 'better' in statistical terms for every age-sex group of every estimation area. However, overall it will be 'better' (in Figure 5.3 the adjusted total has a lower relative RMSE for 80 out of the 108 age-sex by estimation area results) and it will always lead to a reduction in bias. There is an additional warning, Figure 5.3 demonstrates the danger of heavily over-sampling areas of high census underenumeration relative to those with low census underenumeration. In areas of low census underenumeration it will be harder for the CCS to detect the underenumeration and accurately adjust for it. Therefore, reducing the sample in such areas will make the problem worse.

Figure 5.2 gives a similar pattern for the West estimation area. The level of census underenumeration is generally low compared to Belfast but slightly higher than in the East estimation area. The patterns in the bias for the estimator are similar to Belfast again suggesting that there is the possibility of an overall gain from applying the robust strategy developed in section 4.4. However, as with the other two estimation areas, the standard estimation strategy has still been effective at adjusting for census underenumeration across all the age-sex groups. As with the east estimation area, Figure 5.3 shows that for the age-sex groups with higher levels of census underenumeration in the West estimation area, the estimation strategy does better than the census in terms of relative RMSE. However, this is reversed when the level of census underenumeration is lower (one per cent or less) with the census having a lower relative RMSE than the adjusted data.

5.3.3.1) Robust Strategy

In section 4.4 a robust strategy for the ratio estimator was developed for use with dual-system estimation. The aim is to reduce the impact of model failure on the estimator and therefore reduce the variance. As a consequence, this tends to induce a negative bias as well, which should also reduce the impact of the positive bias reported in Table 5.4 and Figure 5.2. The robust strategy has been used in the simulations for Northern Ireland and Table 5.5 compares the robust strategy with the standard ratio estimator combined with a cluster level DSE for censuses with a high coverage.

TABLE 5.5

Comparison of simulation results for the standard and robust estimators for the total population of Northern Ireland by estimation area

Estimation Area	Standard Estimator[1]		Robust Estimator[2]	
	Relative Bias (%)	Relative RMSE (%)	Relative Bias (%)	Relative RMSE (%)
Belfast	0.19	1.74	-0.19	1.53
East	-0.02	0.39	-0.03	0.37
West	0.24	0.75	0.23	0.73
Northern Ireland	0.12	0.46	0.04	0.41

1. Cluster level DSE combined with ratio estimation
2. Estimator based on the robust strategy outlined in section 4.4

Table 5.5 shows that in the East estimation area, where census underenumeration is particularly low, the application of the robust strategy makes very little impact both in terms of bias and variance. In Belfast, where census underenumeration is much higher, the robust strategy has indeed induced a negative bias overall, but this has also resulted in a drop in the relative standard error from 1.73 per cent to 1.52 per cent. In the West estimation area the application of the robust strategy has had very little impact on the overall bias and there is only a small reduction in variance. This suggests that in the West estimation area, the bias and variance in the standard approach is not due to model failure caused by zero census counts or extreme census

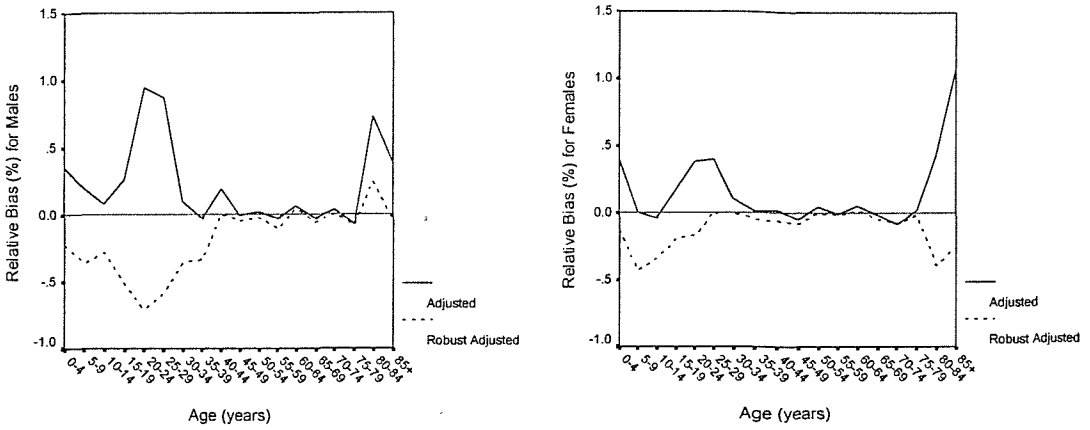
counts. Therefore, the application of the robust strategy will have very little overall impact. The cause of the bias is likely to be the fact that the ratio estimator is not unbiased over repeated sampling, and that the bias will be increasingly important for smaller sample sizes.

The results in Table 5.5 consider the total population by estimation area. Figure 5.4 gives the results for the bias across age-sex groups for each estimation area. As suggested by the results in Table 5.5, the application of the robust strategy has had little impact in both the East and the West estimation areas. In contrast, Figure 5.4 demonstrates that the robust strategy has had an impact in Belfast, particularly amongst the young men where a positive bias of approximately one per cent is replaced by a negative bias of approximately 0.75 per cent. This, on its own, is not necessarily a good result but when combined with the relative RMSE results for Belfast shown in Figure 5.5, there has been a drop in total error across all age-sex groups. This is particularly noticeable for the young men and also at the oldest ages for men and women. Figure 5.5 shows that there are also slight gains across the age-sex groups in both the East and West estimation areas. At oldest ages, population groups with small counts, this means the advantage of the census relative to an estimated population count is reduced and for females in Belfast the robust strategy has a lower total error than both the census and the standard estimator.

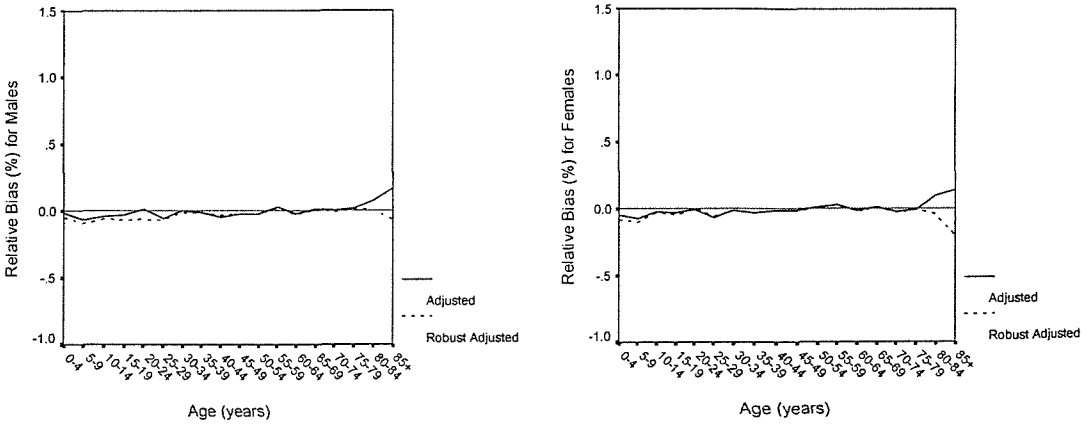
The analysis of the simulation results has so far concentrated on whether the counts produced by combining the census and CCS are 'better' than just using the census. However, as stated by Trussell (1981), it is not the count but the proportion or share of the population that matters for allocation purposes. Table 5.6 considers the mean population shares for males and females within estimation areas as a proportion of the Northern Ireland population produced by the census and the robust estimator over the 1,000 iterations of the simulation.

Figure 5.4: Relative bias for counts adjusted using the standard and robust estimators by age and sex for the three estimation areas

Belfast estimation area



East estimation area



West estimation area

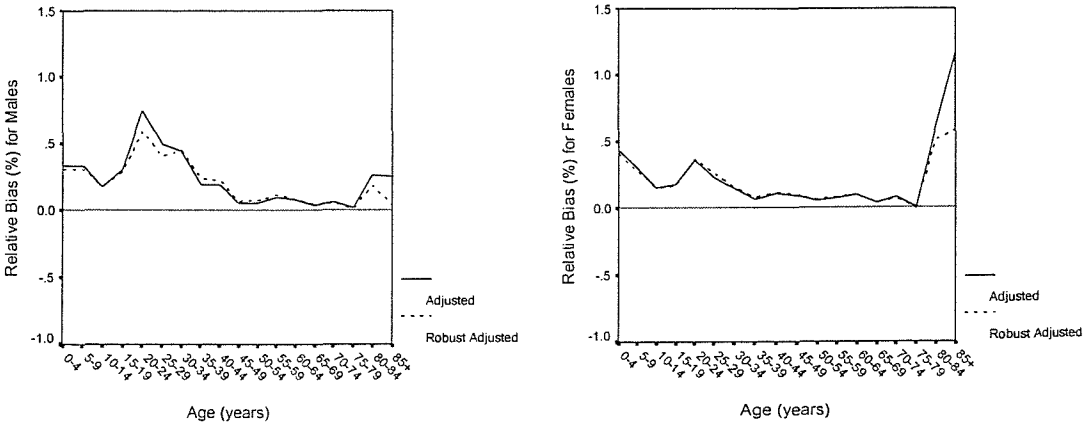


Figure 5.5: Relative RMSE of the census counts compared to counts adjusted using the standard and robust estimators by age and sex for the three estimation areas

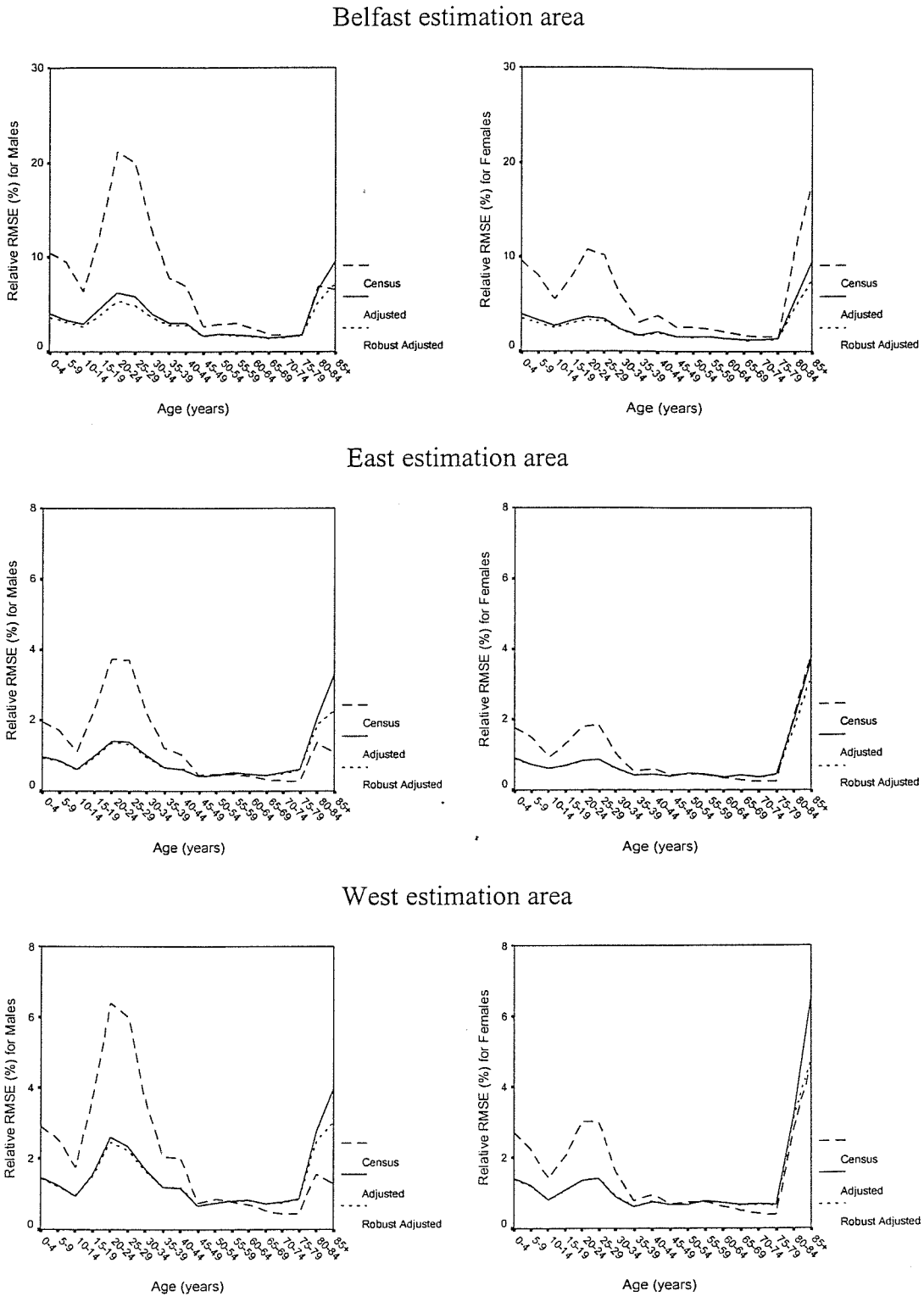


TABLE 5.6

Comparison between the census and the robust estimation strategy for population shares by sex and estimation area

Population Shares (%) of the Total Northern Ireland Population

	Census[1]		Adjusted[1]		Truth	
Estimation Area	Males	Females	Males	Females	Males	Females
Belfast	7.78	9.10	8.31	9.39	8.34	9.40
	(-0.56)	(-0.30)	(-0.03)	(-0.01)		
East	21.84	23.15	21.58	22.74	21.59	22.75
	(0.25)	(0.40)	(-0.01)	(-0.01)		
West	**18.80**	19.33	**18.84**	19.14	**18.80**	19.12
	(0.00)	(0.21)	**(0.04)**	(0.02)		

1. Difference from truth in brackets

The results in Table 5.6 demonstrate two features of census adjustment. Overall, differential levels of census underenumeration will distort population shares. In the Northern Ireland simulations the size of the male population is lower than it should be relative to the female popuation, and high levels of underenumeration in Belfast cause its size relative to the other estimation areas to be reduced. In general, the estimated counts correct for the differential underenumeration and therefore, the shares produced from the estimated counts are much closer to the truth. There is one exception in Table 5.6. If the levels of census underenumeration result in the census getting the share correct, as with males in the West estimation area, the data produced from an estimation strategy will not be able to improve on that and, as any adjusted counts will be subject to some level of sampling variation, may do slightly worse. This comes back to the discussion in the US following the 1990 Census; no adjustment procedure can be expected to improve every population distribution simultaneously. However, as Table 5.6 shows, the use of adjusted data would lead to an overall improvement with five out of six groups having a population share closer to the truth, the exception being males in the West estimation area.

5.4) Estimation for Other Variables

The work presented so far in this chapter, and in chapter four, has concentrated solely on the estimation of the population by age and sex for an estimation area. However, while this is the key characteristic for which census counts adjusted for underenumeration need to be available, the production of a One-Number Census database will require knowledge of the impact of underenumeration on other

variables. These may relate to individuals (ethnicity or in the case of Northern Ireland religion), households and individuals (tenure), or just households (household size).

The estimation strategy outlined in section 4.2 applies dual-system estimation at different levels of aggregation to the matched census CCS data for counts defined by age and sex. However, to get estimates for individuals by another variable, such as religion, dual-system estimation can be applied to counts defined by the alternative variable. In other words, instead of applying dual-system estimation to counts for each age-sex group in can be applied to counts for each religion group or each ethnicity group. Ratio estimation, within strata defined by the hard to count (HtC) index or ED classification index, can then be used to get estimates at the estimation area level by this alternative variable. The work in section 4.3.3 demonstrates that dual-system estimation applied to the cluster of postcodes works well and this approach is adopted here. However, the robust strategy developed in section 4.4 is not applied to ratio estimation. It is considered an unnecessary level of complication, as these estimates are purely control totals in the production of the One-Number Census database. In addition, estimates of the total number of individuals derived from these alternative variables will not be consistent with the estimate of the total population derived from the estimates by age and sex. Therefore, these alternative estimates will be scaled to the estimates by age and sex. The choice of calibrating to the estimates by age and sex reflects the fact that the homogeneity and independence assumptions underpinning dual-system estimation are most likely to hold when applied to counts partitioned by age and sex.

The approach outlined above gives a strategy for estimation of the total population by any individual variable collected in both the census and CCS. This can be extended to estimates for household variables by applying dual-system estimation to households rather than individuals, and combining this with ratio estimation as outlined above. Therefore it is possible to obtain estimates of the total number of occupied households by tenure within an estimation area and therefore, an estimate of the total number of occupied households.

5.4.1) Household Size

Implicit in the work on estimation that has been previously described is the fact that the characteristic being estimated does not 'change' between the census and the CCS. For example, an individual counted in the census and CCS will have the same age and sex. Any discrepancies will be due to respondent errors and in such cases the census response will be chosen. This reflects the fact that the CCS is not measuring quality and that results in the sample areas must be generalised to the non-sample areas where only census responses are available.

Not all variables fit this assumption. For example, the household size is a household variable that is derived from the number of individuals counted in the household. Although the true value does not change the response in the census and CCS will in general be different due to people being missed within counted households by both the census and the CCS. In other words, neither response can be considered a good estimate of the true value. Therefore, such variables require a slightly different approach than simply combining dual-system estimation with ratio estimation. To proceed it is necessary to state some assumptions relating to the data available from the sample areas.

a) For households identified as counted in both the census and CCS by the matching procedures it is possible to define household size as the number of individuals counted in either the census, or the CCS, or both. In other words the assumption is that for households counted by both the census and CCS no individuals within those households are missed by both and the reconciled household size is a good estimate of the true household size.

b) For households only counted by the CCS it is assumed that no individuals within the households are missed.

c) For households only counted by the census it is assumed that no individuals within the households are missed.

It is not expected that the above assumptions will hold exactly but it is certainly reasonable to assume that a) and b) will be closely approximated. This is based on the assertion that for households where the CCS makes contact the interviewer should obtain a good count of the total number of usual residents. For example, probes should reduce the phenomena seen in censuses where young babies are missed from census returns for households. The assumption c) is less tenable but for households the CCS misses and the census counts the census is the only source of information that is available.

TABLE 5.7

Proportions within the sample areas

	Reconciled Census-CCS Household Size						
Census Household Size	1	2	3	4	5	6+	
0	P_{d01}	P_{d02}	P_{d03}	P_{d04}	P_{d05}	P_{d06}	1
1	P_{d11}	P_{d12}	P_{d13}	P_{d14}	P_{d15}	P_{d16}	1
2	0	P_{d22}	P_{d23}	P_{d24}	P_{d25}	P_{d26}	1
3	0	0	P_{d33}	P_{d34}	P_{d35}	P_{d36}	1
4	0	0	0	P_{d44}	P_{d45}	P_{d46}	1
5	0	0	0	0	P_{d55}	P_{d56}	1
6+	0	0	0	0	0	1	1

Using the above assumptions, and once matching has taken place, within the sampled areas for a particular stratum d defined by the HtC index or ED classification index, it is possible to get estimates of the proportions defined in Table 5.7. They are estimated as $\hat{P}_{dij} = N_{dij} / N_{di+}$ where N_{dij} is the unweighted sample count of households with census household size i and reconciled household size j within stratum d and N_{di+} is the unweighted sample count of all households with census household size i within stratum d. The proportions represent transition probabilities between the household size in the census and the true household size as measured in the sample areas by the reconciled census-CCS household size. For example, P_{d13} represents the probability that within stratum d the census will record the household size as one when it is actually 3 and in general, P_{dij} represents the probability that the census will record the household size as i when it is actually j. There is an implicit assumption in Table 5.7;

there is no overenumeration in either the census or the CCS and that therefore the reconciled household size is always greater than or equal to the census household size. Using estimates of the probabilities in Table 5.7 it is now possible to adjust the census counts for households by size.

Before this can be done there is an additional problem. The count of the total number of occupied housing units, as estimated from the distribution of households by tenure, will include an estimate of households missed by both the census and the CCS. To estimate household size for these households an additional assumption is made. For households missed by the census, non-response in the CCS is random with respect to household size. Therefore, the first row in Table 5.7 can be used to get an estimate of the size distribution for households completely missed by the census regardless of whether they were missed by the CCS.

To demonstrate how the estimation proceeds, let X_{di} be the census count of households of size i from stratum d defined by the HtC index or ED classification index, X_d be the corresponding total census count of households, Y_{dj} be the true count of households of size j from stratum d, and Y_d the corresponding true total count of households. Therefore, assuming an estimate of Y_d is available, $\hat{X}_{d0} = \hat{Y}_d - X_d$, and the true distribution of households by size will be estimated as

$$\hat{Y}_{dj} = \sum_{i=0}^{6+} X_{di} \times \hat{P}_{ij} \qquad (5.1)$$

where (5.1) is repeated for all values of j from one to six plus. X_{d0} is replaced by its estimate and $\hat{P}_{ij}$ are the estimates of the probabilities defined in Table 5.7. This approach will produce a set of estimates of the number of households by size that will be consistent with the estimate of the total number of occupied households produced by summing estimates of numbers of occupied households by tenure.

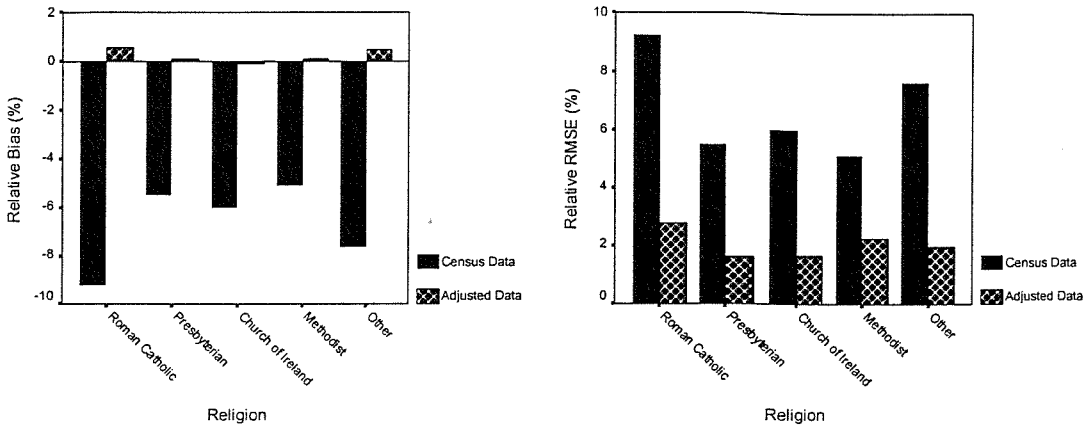
5.4.2) Simulation Results

The simulation study outlined in section 5.3 can now be applied to assess the performance of the estimation strategy with respect to other variables defined at the household and individual level. Figure 5.6 gives the results of the estimation strategy applied to the estimation of the number of individuals by religion across the three estimation areas. In terms of bias, in all three of the estimation areas the adjusted data corrects for the differential census underenumeration. (The higher level of underenumeration for Catholics is driven by the ED classification index, which assigns EDs that are predominantly Catholic the highest levels of census underenumeration.) The relative RMSE graphs show that the adjusted data in general has a lower total error than the census data and therefore gives a better overall representation of the population. In other words, the reduction in bias from using adjusted data more than compensates for any increase in variance. The one exception in Figure 5.6 is for Methodists in the West estimation area. The adjusted data have a higher relative RMSE than the census. In the West estimation area Methodists are a small proportion of the total population (about 8,000 out of half a million individuals) and as discussed earlier, the adjusted data will not always be better than the census, especially for small population groups and low census underenumeration.

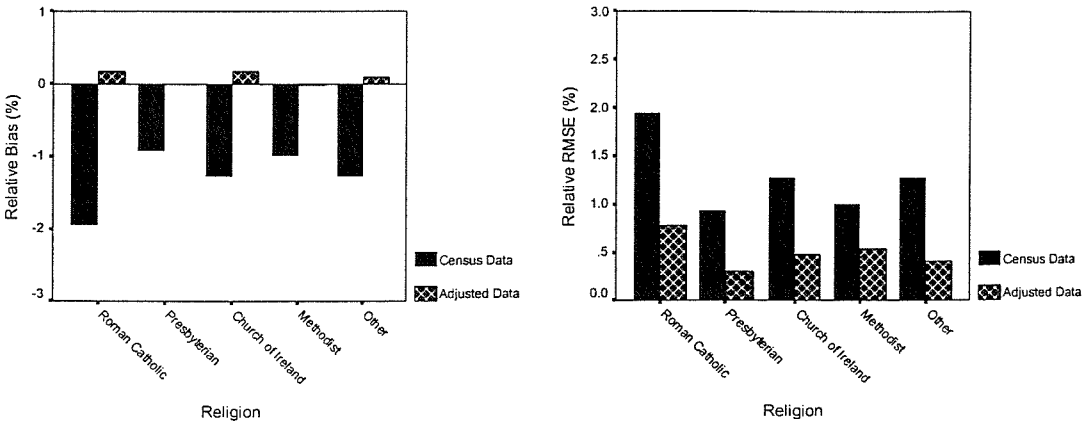
As with the population by age and sex, it is not just the counts that are important but the population shares are also of interest. Population shares by religion are of particular importance in Northern Ireland. For example, some equal opportunities legislation requires information on proportions of the population by religion. Table 5.8 gives the population shares across the estimation areas for the two main religious groups estimated from the census and the data adjusted using the estimation strategy. The results in Table 5.8 show that in the simulations the census underestimates both of the shares in Belfast. However, in the other two estimation areas the population shares for Protestants are over-estimated. Both of these results reflect the structure of the census in the simulations in which Belfast has relatively lower census coverage and the ED classification index rates Catholic EDs as, in general, harder to count. In all cases, the expected values of the shares based on the adjusted data (calculated by averaging across the simulations) are closer to the truth than the census.

Figure 5.6: Performance of the census compared with the standard estimator by religion for the three estimation areas

Belfast estimation area



East estimation area



West estimation area

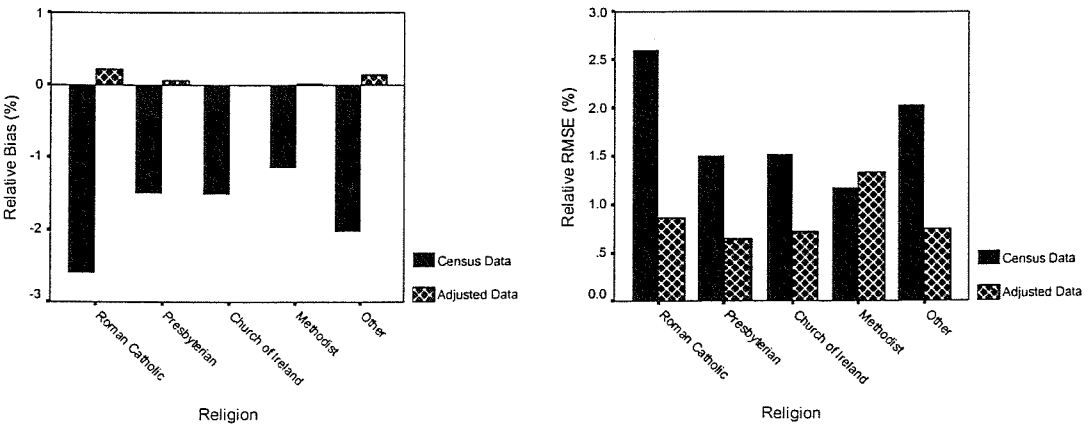


TABLE 5.8

Comparison between the census and the robust estimation strategy for population shares by religion and estimation area

	Population Shares (%) of the Total Northern Ireland Population					
	Census[1]		Adjusted[1]		Truth	
Estimation Area	Catholic	Protestant	Catholic	Protestant	Catholic	Protestant
Belfast	6.49 (-0.47)	6.90 (-0.22)	6.99 (0.03)	7.11 (-0.01)	6.96	7.12
East	10.29 (0.08)	24.09 (0.41)	10.21 (0.00)	23.66 (-0.02)	10.21	23.68
West	21.37 (0.03)	12.06 (0.16)	21.36 (0.02)	11.89 (-0.01)	21.34	11.90

1. Difference from truth in brackets

Note that the proportions do not sum to 100 per cent as about 20 per cent of the population fall into the 'other' category (which may include a few small Protestant groups).

TABLE 5.9

Results of the simulation study for estimates of the total number of occupied households by estimation area

		Performance of the Estimator	
Estimation Area	Census Coverage (%)	Relative Bias (%)	Relative RMSE (%)
Belfast	95.19	0.24	1.07
East	99.23	0.03	0.22
West	98.80	0.04	0.38

The results in Table 5.9 consider households rather than individuals and are based on summing up household counts by tenure. The levels of underenumeration are not as high for households as they are for individuals but this reflects the fact that individuals are missed in counted households and also missed in missed households. The results in Table 5.9 show that at the total population level the strategy is effective at getting the number of households correct within each estimation area. This is true even in the East estimation area where the level of underenumeration is very low at less than one per cent.

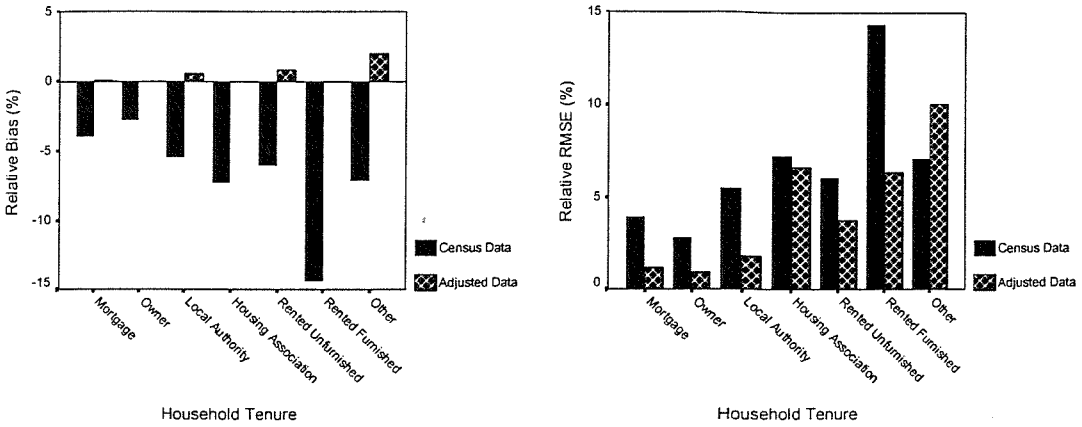
Figure 5.7 presents the household results by the tenure variable rather than just the totals for each estimation area. As with religion for individuals, Figure 5.7 demonstrates that the estimation strategy is effective at correcting the census

underenumeration. However, Figure 5.7 again demonstrates that, in terms of total error measured by relative RMSE, the adjusted data are not always as good. For example, the relative RMSE for the adjusted data for the 'other' category and the 'housing association' category tend to be higher for the adjusted data. However, those categories only represent a small proportion of households. As with the earlier results, the adjusted data are not 'better' in every case but overall they are 'better' than the census data.

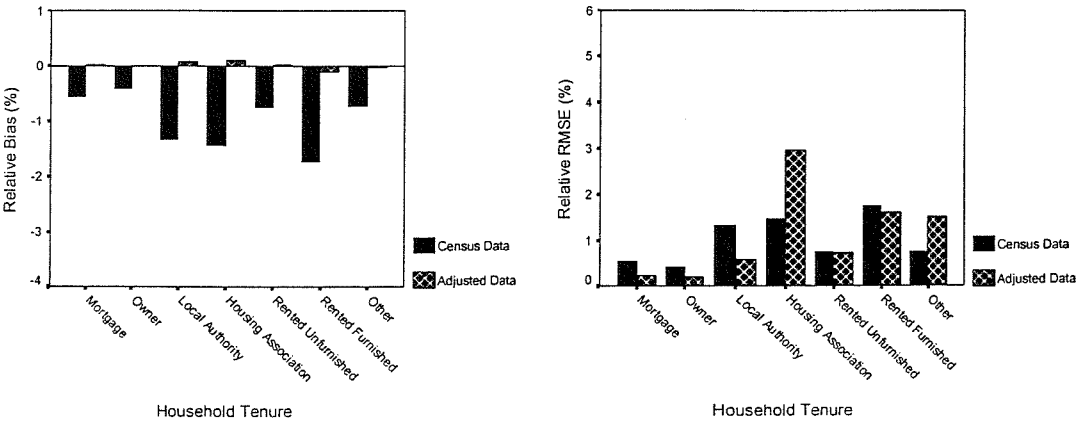
The final set of results is presented in Figure 5.8 for the household size distribution. A peculiar feature of household size is the fact that the census can overestimate the count for a particular size category due to high levels of within household underenumeration of individuals from larger households. This occurs in the Northern Ireland simulations for households of size three in the East and West estimation areas. The same phenomena results in the census getting the count correct for households of size two. As a consequence of this, the adjusted data will not have a lower relative RMSE for those particular categories but again; overall the adjusted data correct the differential underenumeration in the census. In addition, the resulting increase in variance from using the adjusted data compared to census data is usually small relative to the bias reduction and therefore overall the relative RMSE is lower.

Figure 5.7: Performance of the census compared with the standard estimator by household tenure for the three estimation areas

Belfast estimation area



East estimation area



West estimation area

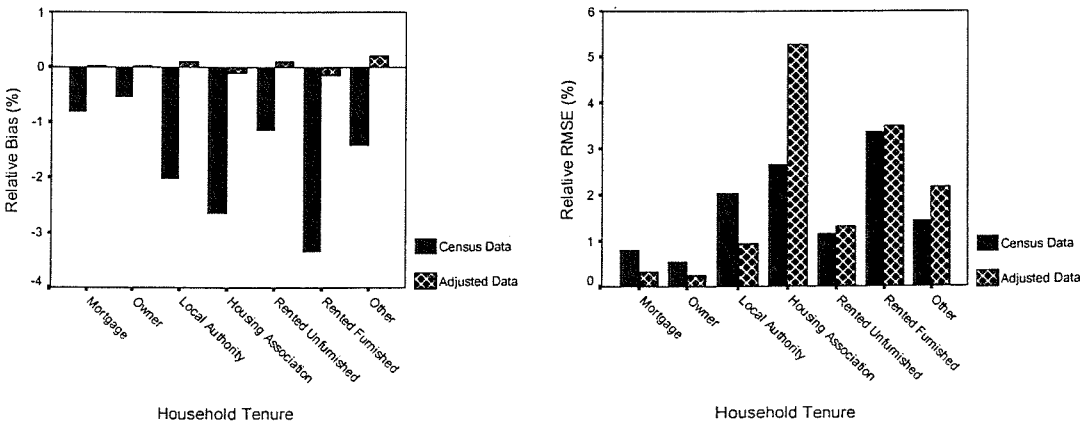
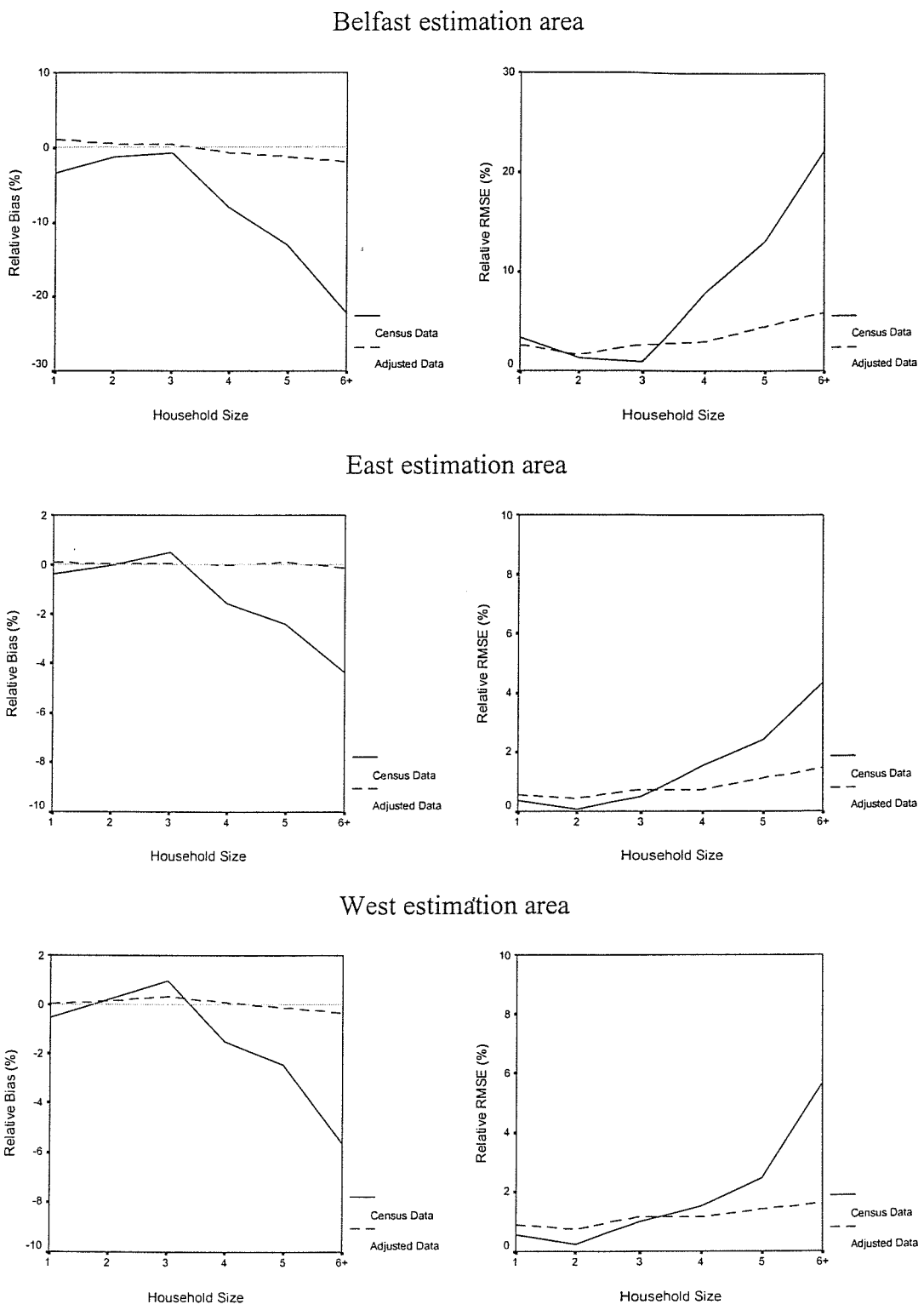


Figure 5.8: Performance of the census compared with the standard estimator by household size for the three estimation areas



5.5) Other Issues Specific to Northern Ireland

5.5.1) Estimation of Population Counts for Local Government Districts (LGDs)

As mentioned in section 4.5.3, getting counts for the population by age and sex for the estimation areas is not the final requirement. There is an additional stage needed to produce estimates for each of the LGDs that constitute the estimation area. As alluded to earlier this subsequent step is of less importance in Northern Ireland than the production of LAD counts in England and Wales. However, mid-year population estimates are still required at the LGD level and they are used in the allocation of Government funding. The strategy for LAD estimation presented in ONS (2000e) was developed assuming that most estimation areas would contain only a few (less than ten) LADs. This is the case when considering estimation areas in England and Wales. However, in Northern Ireland both the East and the West estimation areas contain over ten LGDs and estimating the necessary LGD effects would be problematic.

The proposal for LGD estimation in Northern Ireland utilises the fact that LGDs can be categorised into five groups, one for Belfast and four others, based on a standard Eurostat classification. The groups are formed so that within each group the LGDs are homogeneous in terms of their social and demographic make-up. The estimation areas are based on this, the East estimation area contains two groups and the West estimation area contains two groups. Therefore, within estimation area, the ONS strategy outlined in ONS (2000e) can be used to estimate for these smaller groups of LGDs within the estimation areas and then synthetic estimation within the ED classification index can be used to share the estimated underenumeration across the individual LGDs. Synthetic estimation at this level is not considered unreasonable as within each category of the Eurostat classification by age, sex, and the ED classification index, the constituent LGDs should be approximately homogeneous with respect to census underenumeration.

5.5.2) Selection of the Postcode Sample

The design strategy in section 3.5 suggests a fixed random sample of five postcodes within each selected ED. One motivation for this was the assumption that the cluster of postcodes would form an interviewer workload and the fixed sample would lead to approximately similar sized workloads in terms of number of households. In reality the problem is a little more complex as the distribution of households by postcode is highly skewed, large numbers of postcodes with one or two households and a few postcodes with over 50 households. The solution under consideration by ONS is to, where necessary, group together small workloads or alternatively give a large workload to a pair of interviewers. However, this is subject to change when the final data for drawing the CCS postcodes becomes available.

The original plan in Northern Ireland was to adopt the same strategy. However, the budget for the CCS only allows for an expected sample of about 10,000 households and the ED sample size was specified assuming an average of 15 households per postcode and five postcodes per ED. Across all postcodes in Northern Ireland the average number of households per postcode is indeed about 15. The problem with the final design for the CCS in Northern Ireland is it oversamples EDs in Belfast. The average number of households per postcode in Belfast is actually over 20. Conversely, in the rural West estimation area where the sample of EDs is, in relative terms, smaller the average number of households per postcode is less than 12. Therefore, applying a fixed sample of postcodes per ED results in a sample size of around 11,000 households. To overcome this problem, the approach being taken in Northern Ireland for the selection of the actual CCS sample is to select postcodes at random until the expected count of households (based on information from the Royal Mail) in the selected postcodes exceeds some value. For Belfast it will be set at around 65 households and slightly higher for the other two estimation areas. This means that on average each sample of postcodes selected from an ED will contain approximately 75 households (although this is not guaranteed) and overall the total number of households will be much closer to the target number of 10,000. However, the final postcode sample is unknown.

The overall impact of such a selection procedure on the estimation strategy should be minimal. The robust estimation strategy applies dual-system estimation at the postcode level but the postcode estimates are constrained to the cluster estimates (see section 4.4). This selection procedure should therefore improve estimation in the West estimation area as, in general, the cluster level DSEs will be more stable (based on more individuals). In addition there will, on average, be more postcodes on which to base the ratio component of the estimation strategy. Conversely, there may be a slight disadvantage in Belfast where, on average, the postcode sample will be reduced. However, it should still be around 30 postcodes per level of the ED classification index used in estimation.

5.6) Conclusions

The work in this chapter has taken the CCS design as developed in chapter three and applied it to the data available in Northern Ireland for designing the CCS. This has involved some adaptation of the design, particularly the formulation of an index to stratify EDs in to different types that reflect different expected levels of census underenumeration. The resulting design is a sample of 134 EDs with five postcodes selected per sampled ED.

A simulation for Northern Ireland has been developed using a similar strategy to the one used in section 4.3. This has been used to test the effectiveness of the estimation strategy developed in chapter four when combined with the CCS design adapted for Northern Ireland. The simulation study has demonstrated that, while applying the robust procedures developed in section 4.4 does not always lead to major reductions in variance, it is never worse than the standard approaches for counts across the age-sex distribution. For age-sex groups with very low census underenumeration, the adjusted data do not always have a lower relative RMSE than the census data. However, considering all the age-sex groups together the adjusted data have a lower relative RMSE the majority of the time. In addition, the results show that the estimated counts better represent the population in terms of population shares.

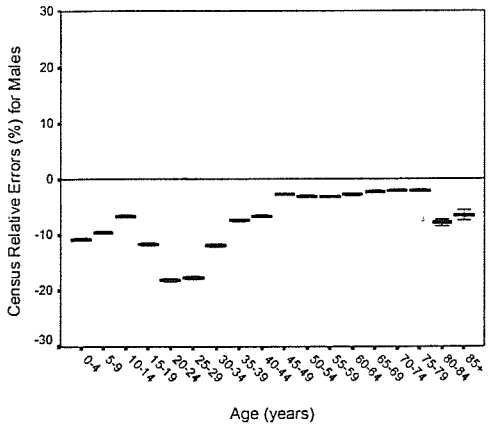
Additional work in this chapter has considered the estimation of counts for other variables at the individual level apart from age and sex and results are presented for

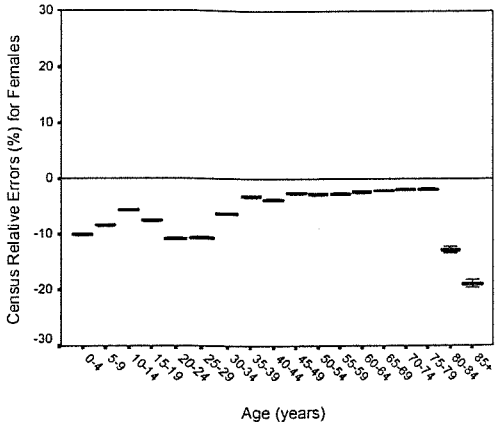
the religion variable. The results show that the general estimation strategy can be applied to the estimation of other individual characteristics such as ethnicity and economic status. The strategy has been further generalised to household variables such as tenure. A specific strategy to accommodate the estimation of household size has also been developed. Simulation results presented for religion, household tenure, and size demonstrate that the strategy works well for other variables in addition to age and sex.

At this stage it will not necessarily be clear to the reader why estimates adjusted for underenumeration for other individual variables apart from age and sex along with household variables would be needed. However, chapter six will demonstrate the importance of such estimates if the final goal is to create a 'One-Number Census' by adjusting the census database to correct for census underenumeration rather than simply informing users about the quality of the census with respect to coverage.

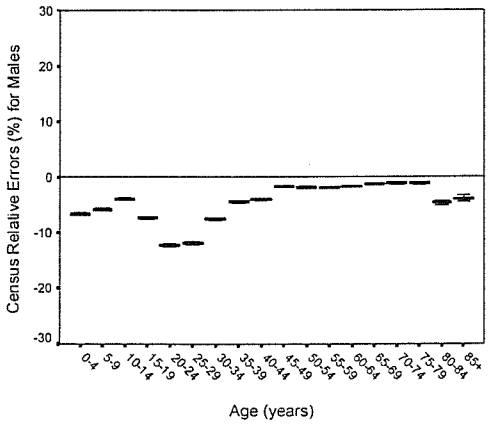
Appendix 5.1 – The distribution of census errors for Northern Ireland by age, sex, and three levels of census coverage.

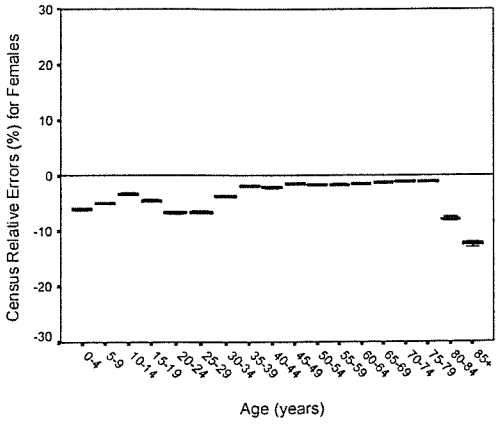
Low



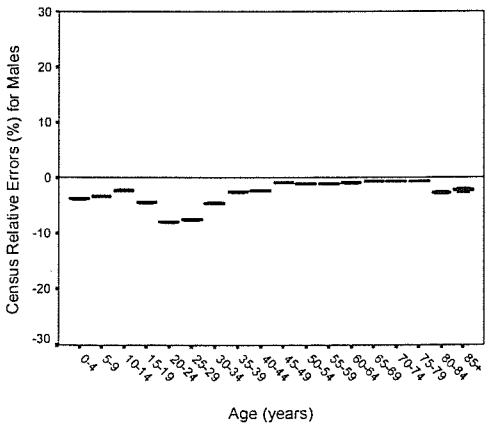


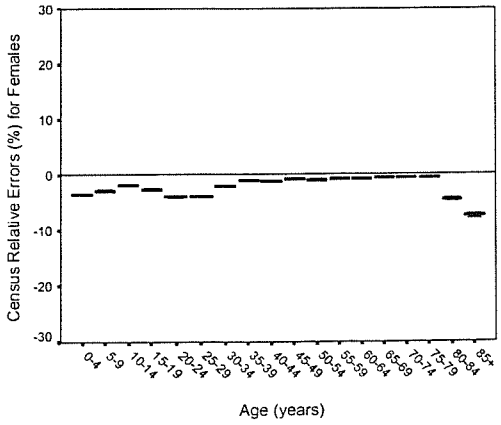
Medium





High





Chapter Six – Adjusting for Census Coverage at the Household and Individual Level

6.1) Introduction

The work in chapters three and four has concentrated on the revised design of the follow-up survey, to be called the Census Coverage Survey (CCS), and the subsequent estimation of the population by age and sex from the 2001 Census augmented by the CCS data. Achieving this must be seen as a fundamental aim of the survey and hence the concentration on this in chapter four. However, achieving accurate estimates of underenumeration by age and sex would only get UK Census users to effectively the same point they were at after the 1981 Census, although the level of disaggregation would be much lower. It would not address the concerns that census users expressed following the 1991 Census related to the higher levels of underenumeration at the national level. Users of census data would remain unsure of what to do if a particular local authority district (LAD) is estimated to have a ten per cent underenumeration for young men, but only a two per cent underenumeration for young women. This can be adjusted for in the mid-year population estimates to get the main allocation of government funds correct, but what is the impact of this on the rest of the census data at the LAD and lower levels. In statistical terms, underenumeration in the census cannot be thought of as simply missing at random; it certainly distorts the distribution of the population by age and sex, and based on the 1991 experience it will also impact on the distribution of the population by other characteristics such as employment status, housing tenure, and geographic location. Estimates of the true population by these characteristics, and household variables, are available via the estimation strategy developed in section 5.4.

The ONS response to these problems has been to research possible strategies for creating a 'One-Number Census' for the UK in 2001. A basic description of what would be involved is given in Brown *et al* (1999), the key component being the integration of estimated underenumeration into the final census database. The US Census Bureau refer to this as the creation a 'transparent file' because to the end user of the census data it appears the same as if they were working with a standard census

database but all the tabulations sum to one number, the agreed population of the UK on census night in 2001. As stated in chapter one, much of the development of this is not due to the author of this thesis. In particular, Professor Ray Chambers, with contributions from Dr. Marie Cruddas (ONS), Professor Ian Diamond, and Tim Jones (ONS), was responsible for much of the early conceptualising of the imputation and weighting framework, which the work presented in this chapter utilises. Subsequently, Dr. Fiona Steele (LSE) has made a substantial contribution in developing the actual imputation system and specifically, wrote SAS programs that enabled a simulation study to assess the proposed system. Over the last year, further development by ONS has taken the basic system developed for the simulation study and made substantial developments towards a fully implemented system for use on the 2001 Census.

The main contribution of the author of this thesis has been in the development of the modelling strategy that is a requirement for the imputation system although a description of the full simulation system, with some basic results, is presented for completeness. A fuller description can be found in Steele, Brown, and Chambers (1999).

6.2) Development of the Framework

After the census and the CCS, in the sampled areas we have a lot of information on the counted individuals and their households. In particular, we can identify individuals the census missed from households it counted and individuals it missed because it missed the whole household. An early attempt to model the process at the individual level defined the following multinomial outcome:

$Y_{ijkedlg} = 0$ when individual i of household j is counted in the census
$Y_{ijkedlg} = 1$ when individual i is missed in the census but household j is counted
$Y_{ijkedlg} = 2$ when household j is missed in the census

for individual i, a member of household j, located in postcode k, of ED e, with HtC category d, in LAD l of estimation area g. The attraction of this approach was that it attempted to capture, in a single model, the two types of underenumeration that individuals experience resulting in some missing individuals being clustered within

missing households and other missing individuals being spread throughout the counted households.

For the postcodes in the CCS sample it is possible to model this multinomial outcome. In general, the outcome will depend on the characteristics of the individual, the type of household, ED level characteristics such as HtC index and the relationships between these variables and the outcome will vary across LADs and estimation areas. The work in Brown *et al* (1998a) develops a modelling strategy for this approach and defines coverage weights for individuals based on predicted probabilities from a multinomial model. The strategy allows for the inclusion of random effects to account for the geographic clustering and Brown *et al* (1998b) investigated the possibility of estimating random effects that were spatially correlated. The simulations undertaken to assess the strategy found that the computing software used to estimate multilevel models meant it would be impractical to implement on a large-scale following the 2001 Census. In addition, the simulation results suggested that including random effects gave little or no benefit with respect to the ability to construct coverage weights.

The work in Brown *et al* (1998a) and Brown *et al* (1998b) effectively only considered estimating the coverage of individuals although the approach accounted for the clustering of missed individuals within missed households. What it did not address is the fact that if a One-Number Census database is to be created it needs to not only 'create' missed individuals (either by weighting or imputation) but also needs to 'create' missed households, which will represent some but not all of the missed individuals. Isaki *et al* (2000) comment that the work in the US following the 1990 Census did not attempt to create households for missed individuals and although census users had accepted that with respect to the 1990 Census they wanted something that was more realistic from the 2000 Census.

The work by Chambers, Cruddas, and Jones (1998), presented at the Leeds Conference for Census Users in May 1998, was the first attempt in the UK to properly consider a framework for not only estimating the missed individuals, but also missed households (and the individuals within them) and more importantly; a process either through weighting or imputation to create the individuals and households on the

database. The discussion at the conference was strongly in favour of imputation. In its favour is the fact that conceptually it is much easier to understand and much easier to get consistency over all possible census tabulations. Once an individual has been imputed onto the database, either within a counted household or as a member of an imputed household; provided this imputation has been done it such a way as to satisfy the 2001 Census edit and consistency rules the individual exists and simply contributes to any tabulation that is required.

The problem with a weighting strategy is ensuring this kind of consistency when weights are applied to both households and individuals when there are not two separate databases that represent households and individuals but a single individual level database that generates all the tabulations. Achieving such consistency would require a very careful calibration between the two sets of weights. There is an additional conceptual issue with weighting. Assume that in an ED, a young man counted by the census has been assigned a weight of two as the modelling suggests that young men have gone missing from counted households. Imputation would then 'create' a record for that missed young man and place it in a counted household. With a weighting strategy the young man will appear on the weighted ED tabulations of individuals, he will even be accounted for through the weights for households. The issue is that he does not exist as a member of any household on the database.

6.3) Controlled Imputation Methodology

As a result of the Leeds Conference research has concentrated on developing an imputation strategy based on the framework in Chambers *et al* (1998). The strategy that has been developed can be thought of as a series of steps in the creation of a database that is fully adjusted for underenumeration.

1) Modelling the census coverage of households and individuals.
2) Imputation of households completely missed by the census.
3) Imputation of individuals missed by the census in counted households.
4) Final adjustments to the database in order to satisfy the consistency requirements for a ONC.

The methodology for each step is outlined in the following sections. It should be noted that the author has made a significant contribution to the development of stage one but that the development of the subsequent stages is mainly attributable to the work of Dr. Fiona Steele.

6.3.1) Estimation of Household and Individual Coverage Weights

The difference in the modelling stage between the framework in Chambers *et al* (1998) and the modelling used in Brown *et al* (1998a) is the splitting of the two types of underenumeration. The former strategy first models the coverage of households and then conditional on the fact that the census counted a household models the coverage of individuals within that counted household. This is then reflected at the imputation stage by imputing missed households (with individuals) and missed individuals within counted households separately.

6.3.1.1) Derivation of household coverage weights

Following the census and the CCS each household within a CCS area can be placed in one of following four categories:

(1) Counted in the census, but missed by the CCS
(2) Counted in the CCS, but missed by the census
(3) Counted in both the census and the CCS
(4) Missed in both the census and the CCS

A simplifying assumption is that category four contains no households, that is no household is missed by both the census and the CCS. While an unrealistic assumption, the households missed by both are accounted for in the dual-system estimates at the estimation area level, and the final imputed database is constrained to satisfy those estimated totals. Excluding category (4), categories (1), (2), and (3) define a multinomial outcome that can be modelled for each estimation area as follows:

$$\log\left(\frac{\theta_{jke}^{(t)}}{\theta_{jke}^{(3)}}\right) = \lambda^{(t)} Z_{jke} \qquad t = 1,2 \tag{6.1}$$

where $\theta_{jke}^{(t)}$ is the probability that household j in postcode k in enumeration district (ED) e within an estimation area, with characteristics defined at the household (i.e. tenure), ED (i.e. HtC index), and LAD level by Z_{jke}, is in category t. (Model (6.1) uses category 3 as the reference category.) With matched data from the census and CCS, this model is straightforward to fit.

The estimated model based on the CCS areas is extrapolated to non-CCS areas within the estimation area to obtain predicted probabilities of being in a particular response category for each household. The probabilities for each response category estimated under model (6.1) are then used to calculate a coverage weight for each household (h/h) counted in the census that can be applied to the household database. The household coverage weight is defined as

$$w_{jke}^{h/h} = \frac{1}{\theta_{jke}^{(1)} + \theta_{jke}^{(3)}} \tag{6.2}$$

However, the resulting weighted sums of counted households will not, in general, match corresponding totals estimated at the estimation area level. Therefore the weights are calibrated to the estimation area marginal totals for key household variables estimated via the strategy in section 5.4, such as tenure, using iterative proportional scaling.

6.3.1.2) Derivation of Individual Coverage Weights

To calculate coverage weights for those individuals counted in counted households, two assumptions are necessary regarding coverage of individuals in CCS areas. If a household is only counted by the census, then no individuals from that household are missed by the census. Similarly, if only the CCS counts the household then no individual from that household are missed by the CCS. These assumptions are necessary because a household counted by only one source has no second list against

which counted individuals can be compared. Although this assumption does not hold in general, people missed as a consequence are accounted for through constraining to population totals at the estimation area level. In this case the possible categories of counted individuals are:

(a) Counted in the census, but missed by the CCS
(b) Counted in the CCS, but missed by the census
(c) Counted in both the census and the CCS

These categories are then used to define the outcome in a second multinomial model:

$$\log\left(\frac{\pi^{(r)}_{ijke}}{\pi^{(c)}_{ijke}}\right) = \beta^{(r)}X_{ijke} + \gamma^{(r)}Z_{jke} \qquad r = a, b \tag{6.3}$$

where $\pi^{(r)}_{ijke}$ is the estimated probability that individual i in household j in postcode k in ED e within an estimation area with individual characteristics defined by X_{ijke} and household/ED/LAD characteristics defined by Z_{jke} is in category r. (Model (6.3) uses category c as the reference category.) As the work in Brown *et al* (1998b) suggests, (6.3), and if desired (6.1), can be extended to include random effects but as noted earlier in section 6.2, their inclusion requires considerable additional computations and the early simulations implied little or no gain for the extra complexity.

As with the household model, the fitted model for individuals in counted households is then extrapolated to non-CCS areas to obtain predicted probabilities of being in a particular response category for each individual. The probabilities estimated under the model are used to calculate a coverage weight for each individual (ind) that can be applied to the individual database. The individual coverage weights are calculated as

$$w^{ind}_{ijke} = \frac{1}{\pi^{(a)}_{ijke} + \pi^{(c)}_{ijke}} \tag{6.4}$$

As before, the resultant weighted sums of census counted individuals will not be equal to the corresponding estimation area totals. At the final stage of the imputation

procedure, further adjustments are necessary to meet agreed estimation area totals by age, sex and household size. To minimise the amount of adjustment required at this stage, individual coverage weights are calibrated to the agreed age-sex totals following the household imputation but before the imputation of individuals.

6.3.2) Imputation of Households

The household-based file of counted households in an estimation area is matched to the file of calibrated household coverage weights (as described in Section 6.3.1.1). This file is sorted by coverage weight, and by geographical location. For more efficient processing, households are then grouped into impute classes defined by the characteristics on which the household coverage weights are based. Weights are grouped into bands to give impute classes. The processing block is an impute class within an estimation area.

Within each processing block, households are processed sequentially and running totals are retained of the unweighted household count and the weighted household count (calculated using calibrated coverage weights). Whenever the weighted count exceeds the unweighted count by more than 0.5, households are imputed into the ED currently being processed until the difference between the weighted and unweighted running totals is less than or equal to 0.5. An imputed household is assigned a household coverage weight of zero. In order to assign characteristics to the imputed households, a donor imputation method is used. For each imputed household, a donor is selected at random from among the counted households with the same weight and in the same ED as the counted household that was processed immediately before the imputation. Once a donor has been selected, the characteristics of the household and its occupants are copied to the imputed household. The imputed household is then assigned at random to a postcode within the ED.

6.3.3) Imputation of Individuals into Counted Households

The individual weights estimated in section 6.3.1.2 are not calibrated to population totals when calculated. However, it is necessary to do this to ensure that enough extra individuals with the correct characteristics are added. This is achieved by using

iterative scaling to calibrate the weights to population totals that reflect the individuals already imputed by the household imputation described in section 6.3.1.

The individual-based file of counted individuals is then sorted by weight, and by geographical location. Impute classes are defined by the characteristics on which the individual coverage weights are based. Individual coverage weights are grouped into bands to give impute classes. Within a processing block (impute class within an estimation area), counted individuals are processed sequentially. When the weighted count of individuals exceeds the unweighted count by more than 0.5, individuals are imputed in the current ED until the difference is less than or equal to 0.5.

Individual and household characteristics are assigned to the imputed individuals in two separate stages. Some of an imputed individual's characteristics are determined by the weight of the last counted individual that was processed before the imputation. The remaining individual characteristics are copied from a suitable donor. The search for a donor is carried out in the same way as described above for the household imputation. The donor is selected at random from among the counted individuals with the same coverage weight and in the same ED as the counted individual that was processed immediately before the imputation. When a donor is found, the LAD is searched for a suitable recipient household in which to place the imputed individual. The household characteristics for an imputed individual come from the selected recipient.

In order to maintain sensible household structures for households into which individuals have been imputed, the type of recipient household sought depends on certain characteristics of the donor. In the simulation study that follows the choice of recipient depends on the age, marital status and household structure of the donor. Household structure is defined using both census and CCS information. Therefore, if an individual who was missed by the census is found in the CCS, the structure of their household will be edited accordingly. To illustrate the recipient search, consider an individual that the coverage weights suggest needs to be imputed. Suppose that a married person went missing from a 'couple without children' household. The household structure(s) that would result after exclusion of the imputed person defines the structure required for the recipient household. Thus the recipient for this

individual must be a single person household. In this case, the marital status of the single person would be edited to married after the imputed person is added to the household. In a further attempt to maintain sensible households, the age-sex composition of the donor's household is also taken into account in the search for a recipient. After selection of a suitable recipient, the imputed individual is placed in the chosen household and is assigned the recipient's household characteristics.

6.3.4) Final Calibration ('pruning and grafting')

Due to the calibration of household coverage weights carried out before the household imputation, the number of households in each impute class will be within one household of the weighted total for that class. Further, the distribution of the household variables to which household weights are calibrated will be almost exactly the same as the target distributions. However, the household size distribution will be incorrect. This is due to individuals being imputed in both Step 2 and Step 3 that leads, in general, to too many larger households. In the final calibration stage, the post-imputation database is adjusted to ensure that the household size distributions and age-sex distributions derived from the ONC database agree with the ONC estimates of their distributions at the LAD level. To achieve this aim some addition and/or deletion of imputed individuals from imputed and counted households will be necessary.

The basic idea of the 'pruning and grafting' procedure is to start at the largest households and work down to households of size one, adding ('grafting') and deleting ('pruning') people to move households up or down in size. The addition of individuals follows the same process as individual imputation while the deletion is at random from a set of possible imputed individuals. This is controlled so that the age-sex distribution after pruning and grafting is exactly calibrated to the control distribution.

6.4) Simulation Study

A comprehensive simulation study has been developed and a full description of its implementation is given Steele *et al* (1999). However, to illustrate that the imputation is feasible, results indicative of the overall performance of the imputation system are

presented. The basis for the simulation is the same data previously analysed in chapter four. In this case ten censuses have been simulated using the methodology outlined in section 4.3, with each census a CCS sample is also selected as per Table 4.1 with a household coverage of 90 per cent and a 98 per cent coverage of individuals within counted households. An important point to note is that the development of the imputation methodology has so far only considered estimation areas that contain a single LAD, as is the case with the data used in the simulation. As part of developing a fully implementable imputation system for the 2001 Census, ONS are undertaking research to assess the most appropriate way to incorporate estimation areas containing multiple LADs.

6.4.1) The Household Coverage Model

For the purposes of modelling household coverage some additional variables have been calculated based on the households and individuals counted in each simulated census. These are household structure and household ethnicity. For modelling purposes, the household structure variable needs to be calculated based on the household structure including the CCS data as, in general, individuals found by the CCS will change the structure of a household. Household structure is categorised as follows: 1) single person, 2) single parent with all children aged under 16, 3) married couple, 4) married couple with all children under 16, 5) unrelated adults, and 6) mixed (including families with children aged 16 or over). The explanatory variables used in the household model (6.1) are tenure, household ethnicity, household structure, and the enumeration district's HtC index.

In general, unless necessary as with the household structure variable, census values for variables are preserved over the CCS values as the model is applied to census (and not CCS) data in the non-sampled areas. For example, if a household changes from owner-occupied in the census to private rented in the CCS the census answer is used in the model. In addition, it is important to remember that the CCS does not measure quality, it is just an independent re-enumeration and having two different values for the tenure variable reflects that individuals within the household give different answers to the same question. We do not know which one is correct.

The household coverage weights defined by (6.2) are calibrated to satisfy marginal distributions estimated at the estimation area level. For this simulation the 'true' marginal distributions have been used, as the aim here is to test the imputation methodology rather than the ability to estimate totals at a higher level. The weights have been calibrated to the true distributions by tenure, household ethnicity, and HtC index. Using the HtC index ensures that, in general, the hardest to count enumeration districts will get more imputed households. The calibration was carried out using an iterative scaling algorithm that converged very rapidly.

6.4.2) The Individual Coverage Model

In the model for individual coverage within counted households children have been considered separately from adults, as they do not have an economic status (as measured by the census). Therefore, two versions of (6.3) are fitted, one for children (those aged under 16) and one for adults. The explanatory variables in the model for children are sex and age group at the individual level, tenure and the number of counted adults based on the household structure variable at the household level, along with the enumeration district's HtC index. The model for adults extends the explanatory variables to include economic status and marital status at the individual level with the full household structure variable at the household level. It is important to remember that census data are always used when these are available. CCS data are only used for the individual characteristics of individuals missed by the census in counted households.

The individual coverage weights defined by (6.4) are approximately calibrated to marginal distributions after accounting for the individuals added by the household imputation. As with the household calibration the 'true' marginal distributions have been used and these are for the 24 category age-sex variable, the HtC variable, the tenure variable at the individual level, and the economic status variable. The calibration is only approximate as it is possible that the household imputation will have added too many women aged 85+. If, for example, this is the case all the individual coverage weights associated with women aged 85+ are constrained to one and means that too many individuals will exist on the database after the two

imputations. It is such situations that mean the final stage (pruning and grafting) is necessary to get the age-sex distribution exactly correct at the estimation area level.

6.4.3) Simulation Results

A more rigorous evaluation of the simulation results is given in Steele *et al* (1999) but the results included here demonstrate that the imputation process to produce a One-Number Census (ONC) is viable. In the evaluation, census underenumeration is thought off as being a negative bias in the estimation procedure of the census. The imputation procedure aims to reduce the bias by adding households and people into the database. However, to achieve this the imputation procedure introduces variability into the database. Therefore, to evaluate the accuracy of estimates at the ED level this trade-off is considered by comparing census and adjusted estimates. The relative average bias and relative root average mean square error (RRAMSE), a combination of variance and bias, are calculated across EDs for selected household and individual variables. Relative average bias is calculated as

$$\text{Relative Average Bias} = \frac{100}{\overline{T}} \times \frac{\sum_{e=1}^{N_e} \sum_{i=1}^{10} (T_{ei}^{(adj)} - T_e)}{10\, N_e} \qquad (6.5)$$

where T_e is the true number of households (individuals) in ED e, $\overline{T}$ is the true mean number of households (individuals) per ED, and $T_{ei}^{(adj)}$ is the number of households (individuals) in ED e in the adjusted census database for simulation i. In other words, the bias for a particular ED is estimated across the ten simulations and then this is averaged over all the EDs. The RRAMSE is calculated as

$$\text{RRAMSE} = \frac{100}{\overline{T}} \times \sqrt{\frac{\sum_{e=1}^{N_e} \sum_{i=1}^{10} (T_{ei}^{(adj)} - T_e)^2}{10\, N_e}} \qquad (6.6)$$

where as above T_e is the true number of households (individuals) in ED e, $\bar{T}$ is the true mean number of households (individuals) per ED, and $T_{ei}^{(adj)}$ is the number of households (individuals) in ED e in the adjusted census database for simulation i. In other words, this is the mean squared error for an ED over the ten simulations averaged over all EDs. The measures (6.5) and (6.6) are calculated within each category of selected household and individual level variables. To assess the performance of the imputation procedure relative to the census, these are compared to the relative average bias and RRAMSE that contrast ED totals in the unadjusted census file with the true ED totals.

TABLE 6.1

Relative average bias and relative root average mean square error across EDs for ten simulations: number of households by HtC index

	Adjusted Data		Census Data	
HtC Index	Relative Average Bias	RRAMSE	Relative Average Bias	RRAMSE
Very easy	-0.05	1.29	-1.75	2.04
Easy	0.02	1.42	-2.27	2.61
Medium	0.00	1.72	-2.99	3.42
Hard	-0.02	2.18	-4.19	4.69
Very hard	0.04	3.14	-7.14	7.81
Overall	0.00	1.95	-3.56	4.23

The results in Table 6.1 first consider the placement of households within EDs for each category of the HtC index and then for the estimation area as a whole. The results across the HtC index demonstrate that in terms of bias the adjusted data is an improvement over the census data and there is also an improvement in terms of overall error. In other words, not only is the imputation putting in the right number of households by HtC index, but they are also being placed in sensible EDs. The overall result of zero bias in the adjusted data in Table 6.1 just reflects the use of true distributions at the calibration stage and the fact that the imputation process preserves this calibration. In reality this implies that provided the estimated control totals are

unbiased, the imputation process will produce a database that, with respect to the calibrated variables, is also unbiased.

TABLE 6.2

Relative average bias and relative root average mean square error across EDs for ten simulations: number of individuals by HtC index

	Adjusted Data		Census Data	
HtC Index	Relative Average Bias	RRAMSE	Relative Average Bias	RRAMSE
Very easy	0.16	2.02	-2.80	3.06
Easy	0.14	1.46	-3.57	3.89
Medium	0.11	1.58	-4.26	4.60
Hard	0.04	1.81	-5.76	6.21
Very hard	-0.49	2.17	-9.34	10.03
Overall	0.00	2.98	-5.10	5.84

The results in 6.2 are for the same variable but at the individual level. As with households, the imputation system is placing individuals in sensible EDs although the bias results for the adjusted data by HtC index are not quite so good as the household results in Table 6.1. However, compared to households individuals need considerably more correction, an overall bias of over five per cent in the census compared to three and a half per cent. In addition, the pruning and grafting stage of the imputation process has a much greater impact on individuals and their distributions. As with Table 6.1, the overall zero bias in the adjusted data in Table 6.2 just reflects the use of true distributions at the calibration stage and the fact that the imputation process combined with pruning and grafting preserves this calibration.

6.5) Discussion

Since the completion of the simulation study, of which some basic results are presented in section 6.4.3, ONS has been working on developing a full imputation system for use on the 2001 Census data. Linked with this has been the need to

develop the interfaces between the actual census database and the imputation system to facilitate the creation of the final adjusted database.

The results from the simulation study demonstrate that imputation can be done, but further work by ONS is considering the use of 'dummy forms' to further improve the system. A dummy form is returned by an enumerator when they consider there to be a non-vacant household in their ED for which they are unable to get a completed census form. It is expected that if the dummy forms are of reasonable quality, they can be used in the imputation system to place 'missed' households and this should further reduce the RRAMSE for the adjusted data by placing households in EDs where there is evidence from the field to support the fact that a household was indeed missed by the 2001 Census. Development work at ONS has also included a rethink of the pruning and grafting strategy. This has resulted in a system that is more stable, this part of the imputation caused considerable problems in the simulations reported by Steele *et al* (1999), and one that achieves the adjusted database in considerably less time.

All the work so far has concentrated on the use of an estimation area with a single LAD. The final area of research and development facing ONS is to include multiple LAD estimation areas into the system. How this is achieved is particularly important at the final pruning and grafting stage where the constraints on the age-sex distribution must be at the LAD and not estimation area level so that the database is consistent with the LAD age-sex estimates that will have already been produced from the CCS via the strategy in chapter four and ONS (2000e).

Chapter Seven – Conclusions

7.1) Introduction

As stated in chapter one of this thesis, the aim of the work presented here has not been to cover all the aspects involved in undertaking a 'One-Number Census' in the UK in 2001. Instead, the thesis describes the development of much of the methodology that will be needed. It should be noted that the practical issues are very important: issues such as efficiently conducting the 2001 Censuses, carrying-out independent follow-up surveys, accurately processing the data from both data collections and, perhaps most challenging, matching the two datasets. Much research has been done by ONS to ensure that the practical problems are effectively overcome. This chapter reviews the work that has been presented in this thesis. The following section considers the design of the census coverage survey, followed by a section on estimation using the survey, and finishing with a section on the imputation methodology for creating the final 'One-Number Census' database.

7.2) The Census Coverage Survey Design

The first goal of this thesis has been to re-consider the design of census follow-up survey in the UK. The 1991 follow-up survey, called the 1991 Census Validation Survey (CVS), was unable to estimate the increased level of underenumeration in the 1991 Censuses. As the two per cent underenumeration was not particularly different from that observed in other countries such as the US, Canada, Australia, and New Zealand, it is reasonable to assume that a similar level of underenumeration will exist in 2001. Therefore, as there were methodological issues with the CVS it is unlikely that adopting the same strategy for the estimation of census underenumeration in 2001 will be any more successful than it was in 1991.

Chapter three specified the basic framework for the design of a follow-up survey for the 2001 Censuses to be called the Census Coverage Survey (CCS) and evaluated several different options within that framework for the allocation of the sample. As the name suggests the biggest change in the proposed approach from the CVS in 1991

has been to separate estimation of coverage from assessment of census quality. This change alone makes undertaking an independent follow-up survey more practical, can simplify the data collection procedures for the interviewer, and reduces the amount of information the follow-up survey needs to collect. The first point opens the door to a range of estimation techniques that could not be utilised by the 1991survey, and these are addressed in chapter four. The second point relates to the fact that the 1991 CVS required the interviewers to re-list a large area, compare their results with the 1991 Census listing, and then sample the different types of households (vacant, multi-occupied, co-operated with census, missed by census) at different rates. This desire to check each part of the census data collection process is important for quality but, from a coverage only point of view, the survey just needs to find the missed people. Therefore, measurement of coverage can be achieved by getting interviewers to re-enumerate small areas without any reference to what the census did. The CCS intends to give interviewers maps of postcodes and ask them to re-enumerate all households identified by the map as being in the postcode, as well as to check the boundaries of the postcode.

The third point has particular relevance for the CCS design strategy adopted in chapter three as it allows for a much larger sample size. The problem with census underenumeration is that many users are not particularly interested in the national level. It is the underenumeration at much smaller geographic levels such as local authority districts (LADs) and by characteristics such as age and sex that is of more interest. The 1991 CVS could only achieve this by grouping the LADs into very broad groups. The strategy in chapter three utilised the increased sample size to form estimation areas at a much lower level of geographic aggregation to avoid the need to make homogeneity assumptions that, for example, assumed Birmingham, Liverpool, Manchester, Leeds, and Bradford to be the same with respect to census underenumeration.

Within the estimation area, the design strategy looks to spread the sample across all types of EDs. This is achieved by using a national hard to count (HtC) index to stratify the EDs within estimation areas. The aim is to ensure that the sample contains all types of EDs including those that are expected to be easy to count as well as those that will be hard to count. Within the HtC index the EDs are further stratified based

on 1991 Census age-sex counts. The sample of EDs is then allocated using optimal allocation based on a design variable constructed by combining several age-sex counts for the EDs based on the 1991 Census. The final stage selects a sample of five postcodes per selected ED. The work in chapter three considered several other options for selecting the postcode sample but this approach has the advantage of clustering the postcode sample for cost efficiency as well as giving good statistical efficiency, based on 1991 Census data, for the estimation of all the age-sex groups.

7.2.1) The Census Coverage Survey Design for Northern Ireland

Chapter five of the thesis considers the implementation of the design strategy developed in chapter three within Northern Ireland. Northern Ireland is unique in the fact that religion defines two communities that have quite different demographic characteristics and in many areas are highly clustered geographically. This is reflected by the fact that the equivalent of the HtC index, the ED classification index, includes dominant religion of the ED as one of its constituent variables. The other variables are an urban / rural identifier for each ED and whether, based on the 1991 Census, the ED was classified as deprived. The urban / rural variable reflects the fact that the local government areas within Northern Ireland are still often a single town or well defined urban area surrounded by rural areas. The final variable reflects the fact that census underenumeration is often associated with variables such as high unemployment which are also associated with measures of deprivation.

The formation of estimation areas in Northern Ireland has also required a slightly different approach. In chapter three a bottom-up approach is taken, with local authority districts grouped to form estimation areas. However, in Northern Ireland the local government districts are much smaller and the approach has been more top down, utilising a standard Eurostat[1] classification of Northern Ireland into three areas based on an analysis of the 1991 Census. In general, each estimation area contains many more local government districts than in England and Wales; except for the estimation area that just contains Belfast. The reason that Belfast is on its own is that,

[1] Eurostat are the agency with responsibility for the collection and quality of comparable statistics from across all member states of the European Union.

in terms of population size, and its social and demographic characteristics, it is significantly different from all other local government districts in Northern Ireland.

The sampling strategy is also slightly different in the Northern Ireland design. Rather than using optimal allocation based on the distribution of a variable constructed from 1991 Census data, proportional allocation is used with respect to population size and the number of EDs. This use of proportional allocation with respect to the number of EDs helps spread the sample evenly across all areas within each estimation area, while the use of proportional allocation with respect to population size ensures that the sample in the West estimation area, a large mainly rural area with a low population, is not over-inflated due to large numbers of small EDs. An approach that spreads the sample as evenly as possible has two advantages. Firstly, as there is very little information from the 1991 Census on which to base assumptions about the likely patterns of underenumeration in 2001, such an approach is sensible and should lead to a sample that is 'representative' of all areas. Secondly, to a politician it looks 'fair' as all groups are evenly represented, particularly important if a full One-Number Census is to undertaken. The potential disadvantage is a slightly less efficient design.

7.3) The Census Coverage Survey Estimation Strategy

As stated in section 7.2, the move to an independent follow-up survey allows for the use of different estimation methodologies, specifically capture-recapture estimation methods. When there are only two 'captures', in this case the census and the CCS, this is referred to as dual-system estimation. A multinomial model that assumes independence between the census count and the CCS count underpins the classical dual-system estimator. This is impossible to guarantee but careful implementation of the two data collections should ensure it is well approximated. It also assumes that the same 'capture probability' applies independently to each individual in the population. On the practical side, it requires very accurate matching of individuals in the census to individuals in the CCS. This allows the identification of individuals counted twice, those counted only once, and then the dual-system estimator accounts for those missed by both.

The approach taken to the use of dual-system estimation in chapter four is different from the general methods reviewed in chapter two. The main difference is that the strategy in chapter four treats dual-system estimation as a method of adjusting the CCS count for non-response so that it gives a 'good' estimate of the true population in each sample postcode. The advantage of such an approach is that the homogeneity assumption required for dual-system estimation is most plausible for small populations. These estimated true counts are then combined with 2001 Census data to get an estimate at the total population level. Initial work in chapter four considered both ratio and regression models for combining census counts as an auxiliary variable with these estimated 'true' counts. Simulations showed the ratio model to be more appropriate and the variance assumption in a ratio model is certainly more appropriate when using count data. However, the approach was not without problems. One advantage of the approach adopted in chapter four is that as it does not calculate the DSE for 'large' sub-populations the homogeneity and independence assumptions only need to be satisfied at a much lower level of aggregation. However, the results in chapter four demonstrated problems with using dual-system estimation at very low levels of aggregation that meant unconstrained estimation at the postcode level was not possible. The simulation results demonstrated that using the DSE for each cluster of five postcodes lead to more stable estimates.

There were two further problems. First, while the ratio model is preferable to the regression model when using count data, it is sensitive to situations where the census count is zero but the CCS count is greater than zero. Such situations did occur in the simulations and will occur in practice. In addition, the estimated counts for a non-sample postcode that were much larger than any sampled postcodes were rather unstable.

The result is the development of an estimation strategy in chapter four that attempts to be robust to these problems. In particular, the strategy separates estimation for postcodes with a zero census count from postcodes with a non-zero census count and applies adjustments to the ratio model when predicting for postcodes with census counts larger than those in the sample. Finally, although the approach uses postcode counts, the DSEs for the individual postcodes within a cluster of five postcodes are constrained to sum to the single DSE calculated by combining the five postcodes

within a selected ED. Simulation results in chapter four demonstrate that the approach introduces a small negative bias into the estimation. However, it also leads to a drop in variance and a drop in the mean square error of the estimator. In particular, the robust strategy is less subject to extreme overestimates of the population total.

7.3.1) Estimation of the Population by Other Characteristics

The simulations in chapter four considered the estimation of the population by age and sex for each estimation area. However, if the intention is to adjust the census database to reflect underenumeration it is necessary to understand the impact of underenumeration on the distribution of other variables such as ethnicity, economic status, and tenure. Simply to adjust the database to reflect underenumeration by age and sex would miss for example the fact that in 1991 it is thought that those in privately rented accommodation had higher underenumeration than those in accommodation with other tenures (e.g. homeowners).

An approach to overcoming this estimation problem is outlined in chapter five and, using simulations for Northern Ireland, is applied to the estimation of the population by religion. Religion was chosen because of its sensitivity within the Northern Ireland context. The simulation results demonstrate that the proposed strategy is effective at not only producing estimates of counts that are closer to the true counts but it also corrects for the impact of the differential underenumeration on the underlying distribution.

7.3.2) Estimation of Occupied Housing Units

A census is not just a count of individuals; it is a count of both households and then the individuals within those households. Therefore, underenumeration of individuals suggests the possibility that some households will have been completely missed by the census. If a full 'One-Number Census' is to be created it is necessary to know about the number and type of households that have been missed. Chapter five develops a strategy for estimating the number of households by tenure and tests the strategy using simulations for Northern Ireland. A special approach is also adopted for the estimation of households by size. The results demonstrate that overall the

estimation strategy works well. However, they also highlight the important point that the estimation strategy will not always be 'better' than the census for every category of every variable within all estimation areas. In the simulations the proposed strategy always reduces bias but sometimes the increase in variance from using sample data leads to an increase in the mean square error.

7.4) Production of a One-Number Census

Much of the work presented in chapter six is not the sole work of the author of this thesis whose main contribution was in the development of the models to estimate the coverage weights needed by the imputation system. However, the imputation system in its entirety is included to allow the reader to have a complete picture of the 'One-Number Census' methodology. The imputation system recognises the fact that individuals are missed by the census through one of two processes. The census misses the entire household in which the individual resides or alternatively, the census misses an individual or individuals within a household but counts the households and some of its residents. Therefore, the imputation system approaches the problem through a series of stages. It first models the types of households that are missed and then uses a donor imputation system to add households (and as a consequence the individuals within those households) to the database. It then models the individuals missed from counted households and again uses a donor imputation system to add the individuals into counted households on the census database. At all stages the imputation system utilises a whole range of estimated totals to control the counts of individuals and households by certain variables within the final database. There is also a final stage that ensures an exact match with the agreed population estimates for the 2001 Census. Simulation results are presented at the end of chapter six to demonstrate that the approach is feasible.

7.5) Concluding Remarks

The work in this thesis shows the development of some of the key aspects of the methodology needed to create a 'One-Number Census' for the UK in 2001. It does not cover all areas of this methodology. In particular, work at ONS has developed the methodology to be used to estimate populations at the local authority district level

(ONS, 2000e) and there is still work to be done on the exact implementation of the imputation system within an estimation area that contains more than one local authority district.

The work presented in this thesis has also included a brief discussion of some of the practical problems associated with the use of dual-system estimation. These include the treatment of movers between the census and the CCS as well as the adjustment for overenumeration, although the exact approach that ONS will use in 2001 is still to be finalised. Unlike the situation in the US, there is no history of using dual-system estimation for estimating census underenumeration in the UK. Consequently, the main task has been to develop an appropriate methodology for its use in the UK context. The work presented in this thesis goes some considerable distance towards that aim and demonstrates a viable and efficient methodology to create a 'One-Number Census' for the UK in 2001.

References

Akers, D. S. (1962). Estimating Net Census Undercount in 1960 Using Analytical Techniques. Paper presented at the annual meeting of the *Population Association of America*, Madison, Wisconsin, May 1962.

Alho, J. M. (1990). Logistic Regression in Capture-Recapture Models. *Biometrics*, **46**, 623-635.

Alho, J. M., Mulry, M. H., Wurdeman, K., and Kim, J. (1993). Estimating Heterogeneity in the Probabilities of Enumeration for Dual-System Estimation. *Journal of the American Statistical Association*, **88**, 1130-1136.

Alho, J. M. (1994) Analysis of sample-based capture-recapture experiments. *Journal of Official Statistics*, **10**, 245 - 256.

Andrews, D. F., Bickel, P. J., Hampel, F. R., Huber, P. J., Rogers, W. H., and Tukey, J. W. (1972). Robust Estimates of Location: Survey and Advances. Published by *Princeton University Press*: Princeton, New Jersey.

Bailar, B. A. and Jones, C. D. (1980). The Evaluation of the 1980 Decennial Census. *Statistician*, **29**, 223-235. (Proceedings of the 1980 I.O.S. Annual Conference on Censuses and Sample Surveys.)

Belin, T. R., Diffendal, G. J., Mack, S., Rubin, D. B., Schafer, J. L., and Zaslavsky, A. M. (1993). Hierarchical Logistic Regression Models for Imputation of Unresolved Enumeration Status in Undercount Estimation. *Journal of the American Statistical Association*, **88**, 1149-1159.

Belin, T. R. and Rolph, J. E. (1994). Can We Reach Concensus on Census Adjustment? *Statistical Science*, **9**, 458-475.

Belley, C., Clark, C., Ha, B., Switzer, K., and Tourigny, J. (1999). Coverage: 1996 Census Technical Reports. *Statistics Canada*, Catalogue No. 92-370-XIE.

Breiman, L. (1994). The 1991 Census Adjustment: Undercount or Bad Data? *Statistical Science*, **9**, 486-508.

Britton, M. and Birch, F. (1985). 1981 Census Post-Enumeration Survey: An enquiry into the coverage and quality of the 1981 Census in England and Wales. Published by *HMSO* (London) for OPCS.

Brown, J. J., Chambers, R. L., Diamond, I. D., and Buckner, L. J. (1998a). Modelling down to small areas. In *A One Number Census: Proceedings of a Research Workshop* edited by Stephen Ludi Simpson. CCSR Occasional Paper 15: University of Manchester.

Brown, J. J., Diamond, I. D., Chambers, R. L., and Buckner, L. J. (1998b). Statistical Models to Estimate Underenumeration in the 2001 Census of England and Wales. Unpublished paper presented at the *Population Association of America Annual Conference*, Chicago, 2nd to 4th April 1998.

Brown, J. J., Diamond, I. D., Chambers, R. L., Buckner, L. J., and Teague, A. D. (1999). A methodological strategy for a one-number census in the UK. *Journal of the Royal Statistical Society A*, **162**, 247-267.

Cantwell, P. (2000). Accuracy and Coverage Evaluation: Missing Data Procedures. *DSSD Census 2000 Procedures and Operations Memorandum Series*, **Q-19**, US Census Bureau.

Chambers, R., Cruddas, M., and Jones, T. (1998). Small Area Estimation for a One Number Census: Weighting versus Imputation. In *A One Number Census: Proceedings of a Research Workshop* edited by Stephen Ludi Simpson. CCSR Occasional Paper 15: University of Manchester.

Chapman, D. G. (1951). Some properties of the hypergeometric distribution with applications to zoological censuses. *Univ. Calif. Public. Stat.* **1**, 131-160.

Citro, C. F. and Cohen, M. L. (eds) (1985). The Bicentennial Census: New Directions for Methodology in 1990. *National Academy Press*, Washington DC.

Coale, A. J. (1955). The Population of the United States in 1950 Classified by Age, Sex, and Color – A Revision of Census Figures. *Journal of the American Statistical Association*, **50**, 16-54.

Cochran, W. G. (1977). Sampling Techniques, third edition. Published by *John Wiley & Sons*: New York.

Cowan, C. D. and Malec, D. (1986). Capture-Recapture Models When Both Sources Have Clustered Observations. *Journal of the American Statistical Association*, **81**, 347-353.

Cressie, N. (1989). Empirical Bayes Estimation of Undercount in the Decennial Census. *Journal of the American Statistical Association*, **84**, 1033-1044.

Darga, K. (1999). Sampling and the Census: A Case Against the Proposed Adjustments for Undercount. Published by *AEI Press*.

Darroch, J. N., Fienberg, S. E., Glonek, G. F. V., and Junker, B. W. (1993). A Three-Sample Multiple-Recapture Approach to Census Population Estimation With Heterogeneous Catchability. *Journal of the American Statistical Association*, **88**, 1137-1148.

Diamond, I. (1994). Where and who are the missing million? Measuring Census of Population Undercount. In *Regional and Local Statistics*, Statistics Users Council. Esher: IMAC Research Ltd.

Diamond, I. and Skinner, C. (1994). Comment on Census Adjustment. *Statistical Science*, **9**, 508-510.

Dunstan, K., Heyen, G., and Paice, J. (1999). Measuring Census Undercount in Australia and New Zealand. *Australian Bureau of Statistics*, Demography Working Paper 99/4.

Ericksen, E. P. (1974). A Regression Method for Estimating Population Changes of Local Areas. *Journal of the American Statistical Association*, **69**, 867-875.

Ericksen, E. P. and Kadane, J. B. (1985). Estimating the Population in a Census Year: 1980 and Beyond. *Journal of the American Statistical Association*, **80**, 98-109.

Ericksen, E. P., Kadane, J. B., and Tukey, J. W. (1989). Adjusting the 1980 Census of Population and Housing. *Journal of the American Statistical Association*, **84**, 927-944.

Fay, R. E. and Herriot, R. A. (1979). Estimates of Income for Small Places – An Application of James-Stein Procedures to Census Data. . *Journal of the American Statistical Association*, **79**, 269-277.

Fienberg, S. E. (1972). The Multiple Recapture Census for Closed Populations and Incomplete 2^k Contingency Tables. *Biometrika*, **59**, 591-603.

Freedman, D. A. and Navidi, W. C. (1986). Regression Models for Adjusting the 1980 Census. *Statistical Science*, **1**, 3-11.

Freedman, D. and Wachter, K. (1994). Heterogeneity and Census Adjustment for the Intercensal Base. *Statistical Science*, **9**, 476-485.

Ghosh, M. and Rao, J. N. K. (1994). Small Area Estimation: An Appraisal (with discussion). *Statistical Science*, **9**, 55-93.

Griffin, R. (2000). Accuracy and Coverage Evaluation: Dual System Estimation. *DSSD Census 2000 Procedures and Operations Memorandum Series*, **Q-20**, US Census Bureau.

Griffin, R. and Haines, D. (2000). Accuracy and Coverage Evaluation: Post-Stratification for Dual System Estimation. *DSSD Census 2000 Procedures and Operations Memorandum Series*, **Q-21**, US Census Bureau.

Hansen, M. H., Hurwitz, W. N., and Pritzker, L. (1953). The Accuracy of Census Results. *American Sociological Review*, **18**, 416-423.

Heady, P., Smith, S., and Avery, V. (1994). 1991 Census Validation Survey: coverage report. Published by HMSO (London) for OPCS.

Hogan, H. (1992). The 1990 Post-Enumeration Survey: An Overview. *The American Statistician*, **46**, 261-269.

Hogan, H. (1993). The 1990 Post-Enumeration Survey: Operations and Results. *Journal of the American Statistical Association*, **88**, 1047-1060.

Hogan, H. (2000). Accuracy and Coverage Evaluation: Theory and Application. Prepared for the 2000 DSE Workshop, 2nd – 3rd February 2000, National Academy of Science Panel to Review the 2000 Census.

Horvitz, D. G. and Thompson, D. J. (1952). A generalization of sampling without replacement from a finite universe. *Journal of the American Statistical Association*, **47**, 663-685.

Isaki, C. T., Ikeda, M. M., Tsay, J. H., and Fuller, W. A. (2000). An Estimation File that Incorporates Auxiliary Information. *Journal of Official Statistics*, **16**, 155-172.

Iversen, R. R., Furstenberg Jr., F. F., and Belzer, A. A. (1999). How Much Do We Count? Interpretation and Error-Making in the Decennial Census. *Demography*, **36**, 121-134.

Jones, G. C. (1997). Planning for the 2001 Census: only four years to go. *Population Trends*, **88**, 31-35.

Kaplan, D. (1970). Plans for the 1970 Census of Population and Housing. *Demography*, **7**, 1-18.

Lyberg, L. and Lundström (1994). Comment on Census Adjustment. *Statistical Science*, **9**, 515-517.

Marks, E. S., Parker Mauldin, W., Nisselson, H. (1953). The Post-Enumeration Survey of the 1950 Census: A Case History in Survey Design. *Journal of the American Statistical Association*, **48**, 220-243.

Marks, E. S. and Waksberg J. (1966). Evaluation of Coverage in the 1960 Census of Population Through Case-by-Case Checking. *American Statistical Association: Proceedings of the Social Statistics Section*, **1966**, 62-90.

Marks, E. S. (1979). The Role of Dual System Estimation in Census Evaluation. In *K. Kroti, Recent Developments in PGE*, 156-188. University of Alberta press.

Mulry, M. H. and Spencer, B. D. (1993). Accuracy of the 1990 Census and Undercount Adjustments. *Journal of the American Statistical Association*, **88**, 1080-1091.

ONS (1998a). Census Coverage Survey: Precision of population estimates for different sample sizes and design areas. *One Number Census Steering Committee Paper* **98/12**.

ONS (1998b). Matching Strategy for a One Number Census. *One Number Census Steering Committee Paper* **98/??**.

ONS (2000a). 2001 Hard to Count Index. *One Number Census Steering Committee Paper* **00/15**.

ONS (2000b). Design Groups for Use in the 2001 One-Number Census. *One Number Census Steering Committee Paper* **00/10**.

ONS (2000c). Methodology for a One Number Census. *One Number Census Steering Committee Paper* **00/15**.

ONS (2000d). One Number Census Estimation Update. *One Number Census Steering Committee Paper* **00/16**.

ONS (2000e). One Number Census Local Authority Estimation. *One Number Census Steering Committee Paper* **00/03B**.

OPCS (Spring 1993). How complete was the 1991 Census? *Population Trends*, **71**, 22-25.

OPCS (Autumn 1993). Rebasing the annual population estimates. *Population Trends*, **73**, 27-31.

Royall, R. M. (1970). On finite population sampling under certain linear regression models. *Biometrika*, **57**, 377-387.

SAS Institute (1990). The CLUSTER procedure: clustering methods. In *SAS/STAT Users Guide Version 6*, 4th edition, volume 1, 529-536. Cary: SAS Institute.

Scott, A. J. and Holt, D. (1982) The effect of two-stage sampling on ordinary least squares methods. *Journal of the American Statistical Association* **77**, 848-854.

Schirm, A. L., and Preston, J. (1987). Census undercount adjustments and the quality of geographic population distributions (with discussion). *Journal of the American Statistical Association*, **82**, 965-990.

Seber, G. A. F. (1982). The estimation of animal abundance and related parameters. Second edition published by *Charles Griffin & Company Ltd*, London.

Sekar, C. C. and Deming W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association*, **44**, 101-115.

Siegel, J. S. (1974). Estimates of Coverage of the Population by Sex, Race, and Age in the 1970 Census. *Demography*, **11**, 1-23.

Simpson, S. (1994). Coverage of the Great Britain census of population and housing. *Estimating With Confidence Project*, **Working Paper 8**. Available from Social Statistics, University of Southampton SO17 1BJ.

Simpson, S., Cossey, R. and Diamond, I. (1997) 1991 population estimates for areas smaller than districts. *Population Trends* **90**, 31-39.

Steel, D. (1994). Comment on Census Adjustment. *Statistical Science*, **9**, 517-519.

Steele, F., Brown, J., and Chambers, R. (1999). A Controlled Donor Imputation System to Create a Census Database Fully Adjusted for Underenumeration. Submitted to *Journal of the Royal Statistical Society A*, December 1999. A working paper version is available from the authors upon request.

Steinberg, J., Gurney, M., and Perkins, W. (1962). The Accuracy of the 1960 Census Count. *American Statistical Association: Proceedings of the Social Statistics Section*, **1962**, 76-79.

Taeuber, C. and Hansen, M. (1964). A Preliminary Evaluation of the 1960 Censuses of Population and Housing. *Demography*, **1**, 1-14.

Trussell, J. (1981). Should State and Local Area Census Counts be Adjusted? *Population Index*, **47**, 4-12.

Waksberg, J. and Perkins, W. M. (1971). The Role of Evaluation in US Censuses of Population and Housing. *Statistician*, **20**, 33-46. (Population Censuses.)

Webber, R. (1977). The national classification of residential neighbourhoods: an introduction to the classification of wards and parishes. *PRAG Technical Papers*, **TP 23**. Centre for Environmental Studies.

Werker, Henry F. (1981). Results of the 1980 US Census Challenged. *Population and Development Review*, **7**, 155-167.

Wolter, K. M. (1986). Some Coverage Error Models for Census Data. *Journal of the American Statistical Association*, **81**, 338-346.

Wright, T. and Hogan, H. (1999). Census 2000: Evolution of the Revised Plan. *Chance*, **12**, 11-19.

Zaslavsky, A. M. and Wolfgang, G. S. (1990). Triple-System Modelig of Census, Post-Enumeration Survey, and Administrative List Data. *American Statistical Association: Proceedings of the Survey Research Section*, **1990**, 668-673.

One Number Census Steering Committee Papers are public documents and copies are available upon request to:

One Number Census Team
Room 4200W, Census Division
Office for National Statistics
Segensworth Road, Titchfield
Fareham, Hants PO15 5RR, UK.