

ONE NUMBER CENSUS STEERING COMMITTEE**One Number Census Local Authority Estimation**

1. This paper reports the research towards developing a methodology for estimating the populations of Local Authority Districts as part of the One Number Census. A number of different small area estimation techniques, which borrow strength across areas, are examined. These methods include direct estimates, synthetic estimates and estimates which are essentially compromises between these two extremes.
2. A simulation study was undertaken to evaluate the proposed alternatives and explore both the robustness and efficiency of the methodologies. The results of the simulations indicated that although the synthetic estimator was generally more efficient than any of the alternatives, this was not the case when large differences between LADs were present. The most robust estimator across all different conditions is that which uses a fixed LAD specific effect to model the differences between the smaller areas.
3. **Members of the Steering Committee are asked to:**
 - a) **Note the results presented in the paper;**
 - b) **Agree the recommendation that a ratio model including a LAD specific fixed effect approach be used for the Local Authority District estimation; and;**
 - c) **Provide comments at the meeting or in writing by 23rd February 2000.**

**Owen Abbott
Census Division, Room 4200W
Office for National Statistics
Segensworth Road
Titchfield
Fareham
Hampshire
PO15 5RR**

One Number Census Local Authority Estimation Strategy

Executive Summary

Introduction

The One Number Census (ONC) project aims to estimate the level of underenumeration of both households and individuals in the 2001 Census. The postcode based Census Coverage Survey (CCS) has been designed to provide the data that will facilitate the estimation of census coverage by age and sex. The design is described in ONS(ONC(SC))00/01.

The ONC has two estimation phases. Within Phase one, Dual System Estimation (DSE) methodology will be used to combine 2001 Census and CCS counts to estimate the true population in the sampled CCS postcodes. Generalisation of these DSE counts from the sampled areas to the whole population will be carried out using the strategy described in ONS(ONC(SC))00/03A. This will be used in the production of underenumeration estimates by age and sex for approximately one hundred 'design groups' in England and Wales. Each design group is an aggregation of a number of whole Local Authority Districts (LADs) and is the level at which the CCS is designed to provide direct estimates of an acceptable precision.

This paper is concerned with Phase two of the ONC estimation - the allocation of the estimated design group underenumeration to the individual LADs. These estimates must be consistent with the design group estimates produced from Phase one, and will be used as the population totals to which the ONC imputation process is constrained.

Due to the small sample sizes within each individual LAD, standard direct estimators such as those used to estimate the design group totals would yield unbiased estimates with very large standard errors. This has led to the development of techniques which borrow strength from related areas to make indirect estimates that increase the effective sample size and thus decrease the error level associated with the estimates. However, the price paid for the decreased variances are possible large biases.

A number of different approaches to the estimation of the LAD populations within the One Number Census framework were considered. The small area estimation methods examined include direct estimates, synthetic estimates and estimates which are essentially compromises between these two extremes.

Options

Four approaches are considered, all based around the Dual System Estimator used in phase one combined with a ratio model.

Approach 1 was a direct ratio estimator applied separately within each LAD. The sample is used to estimate the mean ratio between the 2001 Census postcode counts and the corresponding dual system estimates for each age-sex group in each LAD. The ratios can be interpreted as the underenumeration adjustment weights. The adjustments are applied to all postcode age-sex Census counts, the sum of which will be the population estimate of interest. This does not borrow strength across LADs, but to improve efficiency we collapse some age-sex groups (which borrows strength within the LAD) to avoid problems with zero postcode counts.

Approach 2 used a simple synthetic model to apportion the estimated design group undercount across the LADs according to the Census count within each area. For instance, if the estimated undercount at design group level was 2% then the synthetic model will use this adjustment for all LAD populations within the design group. This is a simple yet effective approach that is likely to perform well if there are no big differences between the LAD 2001 Census undercount patterns.

Approach 3 extends the synthetic model to include a fixed LAD effect to allow some variation in underenumeration across LADs. The model is fitted at design group level, estimating an overall design group coverage adjustment, with fixed deviations for each LAD.

Finally, Approach 4 contained a random LAD effect within the model which allows the underenumeration to vary between LADs. This is similar to the fixed effects model described above (Approach 3), but the key difference is that the LAD specific effects are drawn from a normal distribution – the parameters of which are estimated. This is a random slopes model that allows for unexplained differences between LADs.

All of the methodologies proposed were tested within a set of simulations under different conditions to compare their relative performance and assess robustness. The ONC small area estimation strategy must be capable of delivering accurate LAD estimates in a variety of different circumstances. Four different design groups were used containing different numbers of LADs and covering a wide range of demographic types. The undercount patterns within each was altered to assess the impact on the estimates and to allow an evaluation of robustness.

Results

The results clearly show that for the majority of simulations, a simple synthetic apportionment estimator (Approach 2) provides results with the greatest precision. Under conditions where there is a relatively small overall difference in the LAD Census coverage, it is consistently better across the age-sex estimates. However, when a significant LAD effect is present, an LAD adjusted synthetic model (Approach 3) is more robust and performs better.

It is anticipated that there may be some design groups where there are large differences in the LAD Census coverage. **It is therefore recommended that an LAD adjusted synthetic approach be used for phase two of the One Number Census estimation strategy.**

SMALL AREA POPULATION ESTIMATION IN THE 2001 UK ONE NUMBER CENSUS

Owen Abbott, James Brown, Ray Chambers and Marie Cruddas.

1. INTRODUCTION

- 1.1 Considerable efforts have been made to maximise the coverage of the UK 2001 Census. However, it is only realistic to expect there will be some underenumeration, particularly of certain age-sex groups (e.g. young males, old people and very young children). The One Number Census (ONC) project aims to estimate the level of underenumeration of both households and individuals in the 2001 Census, based on data collected in a household-based post enumeration survey called the Census Coverage Survey (CCS). Data from the CCS will be matched with corresponding census data to obtain estimates of underenumeration that will then be used to adjust the census data to produce a single "best" estimate of the actual UK population distribution at the time of the census. This will define the final "one number" tabulations from the Census.
- 1.2 In order to achieve this objective in England and Wales, the CCS sample size and design will be such that age-sex population totals can be estimated to a specified level of accuracy within approximately one hundred Design Groups, each with a population of approximately 0.5 million persons. The sample design is detailed in ONS(ONC(SC))00/01, and the proposed Design Groups in ONS(ONC(SC))00/10. Population estimation within these areas will be based on a combined dual system/regression estimation methodology, described in ONS(ONC(SC))00/03A. Each Design Group is an aggregation of a number of whole Local Authority Districts (LADs), which are important administrative units for resource allocation. These LADs can vary enormously in population size, density and characteristics, and estimates for the individual LADs are needed as part of the ONC process.
- 1.3 This paper examines a number of different approaches to the estimation of the LAD populations within this framework. These methods include direct estimates, synthetic estimates and estimates which are essentially compromises between these two extremes. Synthetic and compromise "mixed model" estimates are necessary, as the sample size within an LAD is often insufficient to make reliable direct estimates for individual age-sex groups. A primary consideration is robustness of the chosen method since it will be applied in many different conditions in 2001.
- 1.4 A simulation study designed to compare each of the suggested approaches was undertaken. A description and the results are included, together with a recommended strategy for the One Number Census LAD population estimation.

2 SMALL AREA ESTIMATION

- 2.1 Design group estimates of underenumeration are relatively straightforward, given the sample design proposed in ONS(ONC(SC))00/01. The methodology combines Dual System Estimation technology with classical Ratio models and is described fully in ONS(ONC(SC))00/03A.
- 2.2 However, the application of this methodology to the majority of LADs is problematic. The reason for this is simple – there are simply not enough sampled postcodes within the individual LADs for the estimates to be sufficiently accurate or reliable. The current sampling strategy specifies that

all LADs should have at least one sampled postcode within them. There are, however, no other restrictions on selection at present.

- 2.3 Because of these small sample sizes, standard direct estimators yield estimates with very large standard errors, although they are unbiased. This has led to the development of techniques which ‘borrow strength’ from related areas to make indirect estimates that increase the effective sample size and thus decrease the variance. However, these methods usually make use of strong assumptions about the relationship between the small areas, and the relationship between the small areas and the larger aggregated area. This can mean that while the estimators can have low variances, they also can have a large bias.
- 2.4 There are a vast number of small area estimation approaches in the literature (see Ghosh & Rao (1994) for an overview of some of the methods). These vary from simple apportionment models to the more complex class of mixed models, which are essentially ‘in between’ the standard unbiased direct estimators and the (potentially very biased) synthetic apportionment estimators. These types of estimators have been used for many different applications over the last few years with the growing demand for reliable small area statistics. In particular, the U.S. Census Bureau used simple synthetic estimators within their population estimation strategy for the 1990 Census, and are planning to use these techniques again for their 2000 Census. The more complex methodologies in the literature that are not in widespread use will not be considered for this research, as there would be a danger that the methodology would not be acceptable to census users.
- 2.5 For the ONC estimation strategy, the best methodology is that which will get the correct distribution of LAD estimates. Therefore the best estimator is not necessarily the one that gives the most accurate overall predictions. A method that is not as accurate, but the accuracy is similar across all area types (such as rural, urban areas) might well be the best option. The reason for this is that the LAD estimates will all be scaled to the computed Design Group estimates (which we are assuming to be highly accurate at the Design Group level). Hence we are mostly interested in a methodology that is robust rather than the most efficient.

3 BACKGROUND

- 3.1 As time is limited, it would be impractical to examine and evaluate all alternatives to the ONC small area estimation problem. To reduce the possibilities to a manageable level, a literature review was carried out which identified a number of approaches. Preliminary exploration of the options was undertaken using a simple simulated dataset to both check that the models could be fitted and to refine them if necessary. A detailed specification of each of the four main alternative approaches is given in sections 4-7. This section provides background to the estimation environment by detailing the CCS Design, the collapsing of the categories used in the study, the notation used and the calibration to Design Group totals that is required.

Census Coverage Survey Design

- 3.2 The Census Coverage Survey (CCS) is the key component of the One Number Census (ONC) project. Data obtained from this survey will not only be used to produce population estimates, but will also, when combined with matched census counts, form the basis for models that will be used to adjust for differential underenumeration in the 2001 Census. The CCS will involve the re-enumeration of a sample of postcode areas within each Design Group, stratified by an index measuring the expected level of difficulty associated with enumerating a postcode in the census.

Since the main purpose of the CCS will be to estimate the extent of underenumeration in the census, it will be carried out as soon as possible after the census. At this stage it is anticipated that the CCS fieldwork will begin three to four weeks after census day.

- 3.3 It is expected that underenumeration in the 2001 Census will be higher in areas characterised by particular social, economic and demographic characteristics. In order to control for this effect postcodes within each Design Group will be stratified by a 'Hard to Count' (HtC) index as well as by size (population count at the previous 1991 Census). For the analysis reported below a national HtC index for 1991 Census EDs was calculated by ranking these districts with respect to the sum of a series of variables collected in 1991. These scores were divided into a 40%, 40%, 20% distribution at a national level, with each assigned an index value from 1 (easiest to count) to 3 (hardest to count). All postcodes within an ED were then allocated the HtC category for that district.
- 3.4 The CCS uses a two-stage design. The first stage of the design selects a sample of 1991 Enumeration Districts (EDs), stratified by hard to count, from Design Group areas. The second stage then samples postcodes within the selected first stage EDs. The two-stage design assumes that the aim of the CCS is to produce population estimates of specified accuracy at Design Group level for the five year age/sex groups.

Categories used in the models

- 3.5 Because of the small sample sizes involved in this analysis, the models used a different set of groupings for computing estimates than those that were used in the previous research into the Design Group estimation. By collapsing together some of the categories, we are making the assumption that the model we fit holds for the categories that have been collapsed. These collapsings are used to fit the model but predictions are calculated separately for the uncollapsed categories.
- 3.6 Design Group estimates are produced separately for all five year age/sex groups. For the estimation of the LAD totals the number of age-sex groupings were reduced to effectively increase sample sizes, as within a postcode the majority of age-sex groups are represented. This reduced the problems encountered when there are a large number of zero counts. This approach makes the assumption that the underenumeration does not vary between some age-sex groups:

0-4 year olds
5-14 year olds
15-19 year old males
15-19 year old females
20-24 year old males
25-29 year old males
20-29 year old females
30-34 year old males
35-39 year old males
30-39 year old females
40-44 year olds
45-59 year olds
60-69 year olds
70-79 year olds
80+ males
80+ females

3.7 These collapsed categories were used to fit all the models (with the exception of the simple synthetic model), and estimates were calculated for the 36 age-sex categories through use of the fitted collapsed model. More drastic age-sex collapsing were considered in an attempt to improve efficiency, but it was felt that this collapsing would be that which would be most likely to be used in 2001 given that we would have no information on which to base our choice other than the 1991 experiences.

Notation

- l = 1...L are the Local Authority Districts (LADs) in Design Group G.
- d = 1...D are the HtC strata of postcodes in the Design Group.
- a = 1...A age-sex categories (A = 36 here).
- c = 1...C collapsed age-sex categories (C = 16 here). An age-sex category a that is part of collapsed category c is denoted $a \in c$.
- k = 1...N_{dl} are the postcodes in HtC category d of LAD l, of which n_{dl} are in the CCS sample S_{dl}.
- X_{kadl} = 2001 Census count (unadjusted) for age-sex category a in postcode k in HtC stratum d in LAD l.
- Y_{kadl} = CCS-based DSE count (for those postcodes in the CCS sample) for age-sex category a in postcode k in HtC stratum d in LAD l.
- \hat{T}_{ac}^G = Design Group G estimate for age-sex category a in HtC stratum d.
- R_{ad} = Ratio parameter

Calibration to Design Group Totals

3.8 For all of the alternative approaches considered in this study, there is no guarantee that the predicted LAD totals sum to the Design Group estimate. Therefore, we have to consider a calibration adjustment to all of the LAD estimates to ensure that this condition is met. The LAD estimates were scaled to the true Design Group age-sex by HtC totals \hat{T}_{ac}^G using a simple scaling adjustment as follows:

$$\hat{T}_{adl}^{CAL} = \hat{T}_{adl} \frac{\hat{T}_{ad}^G}{\sum_{l \in G} \hat{T}_{adl}}$$

3.9 Adjusting the estimates to the true Design Group totals basically means that we are assuming that the Design Group estimation is perfect. This is reasonable for the purposes of the comparison of the different approaches. However, further work may be required to examine the effects of the Design Group estimates on the LAD estimates for the particular strategy adopted.

4 Approach 1 (Direct Ratio Estimator based on Collapsed Age-Sex Categories)

4.1 We would ideally like to replicate the ratio estimation methodology used at Design Group level within an LAD. That is, we would like to use a direct ratio estimator for each age-sex category within each HtC stratum within each LAD. However, small sample counts in an LAD mean that,

although this estimator will be unbiased, it is also likely to be extremely variable, because of the many zero counts that will occur within age-sex categories at HtC by LAD level.

- 4.2 The approach taken here therefore is to collapse across age-sex categories within an LAD to ensure that there are sufficient nonzero counts to make the ratio estimator approach viable at LAD level. This is equivalent to fitting $48 \times L$ separate models – one for each HtC stratum by collapsed age-sex category by LAD. Estimates for the original 5 year age-sex categories are computed using the coefficients for the collapsed age-sex categories. In other words it is assumed that there is no difference in the relationship between CCS Dual System Estimator based counts and Census counts in the age-sex categories making up a collapsed category at HtC by LAD level. The model implicit in this approach, and its associated estimator, are set out below.

Model

$$\begin{aligned}
 E(Y_{kadl} | X_{kadl}) &= R_{cdl} X_{kadl} \text{ for } a \in C \\
 \text{Var}(Y_{kadl} | X_{kadl}) &= \sigma_{cdl}^2 X_{kadl} \text{ for } a \in C \\
 \text{Cov}(Y_{kadl}, Y_{jbem} | X_{kadl}, X_{jbem}) &= 0 \text{ for all } j \neq k
 \end{aligned}$$

Estimator $\hat{T}_{adl} = \hat{R}_{cdl} \sum_{k=1}^{N_{cdl}} X_{kadl} \text{ for } a \in C$

where $\hat{R}_{cdl} = \frac{\sum_{s_{cdl}} \sum_{a \in C} Y_{kadl}}{\sum_{s_{cdl}} \sum_{a \in C} X_{kadl}}$.

- 4.3 This is an area specific model that does not ‘borrow strength’ between LADs within the Design Group. The number of sample points contributing to an estimate is dependant on the number of LADs in the design group. Any more than 5 or 6 LADs could lead to very small sample sizes – between 10 and 15 postcodes – and consequently zero counts for some age-sex categories (even after collapsing). Where the Census count for the sample postcodes in a HtC stratum within an LAD is zero for a particular collapsed age-sex category, the observed Census counts for an age-sex group within that collapsed category will be used as the estimate (i.e. \hat{R}_{cd} will be set to 1).

5 Approach 2 (Simple Synthetic Ratio Estimator)

- 5.1 An obvious alternative to the above approach is to "borrow strength" across all LADs for a particular age-sex category rather than to do this by collapsing age-sex categories within an LAD. That is, we assume the same regression slope values R_{adl} hold within age-sex category a across all LADs within the Design Group. This leads to a simple synthetic ratio estimator based on the following model and associated estimator.

Model

$$\begin{aligned}
 E(Y_{kadl} | X_{kadl}) &= R_{ad} X_{kadl} \\
 \text{Var}(Y_{kadl} | X_{kadl}) &= \sigma_{ad}^2 X_{kadl} \\
 \text{Cov}(Y_{kadl}, Y_{jbem} | X_{kadl}, X_{jbem}) &= 0 \text{ for all } j \neq k
 \end{aligned}$$

$$\text{Estimator} \quad \hat{T}_{adl} = \sum_{k \in S_{dl}} Y_{kadl} + \sum_{k \notin S_{dl}} \hat{R}_{ad} X_{kadl}$$

$$\text{where } \hat{R}_{ad} = \frac{\hat{T}_{ad}^G}{X_{ad}}$$

- 5.2 This estimator uses the level of underenumeration in each age-sex category by HtC stratum in the Design Group to adjust the corresponding LAD census populations. It assumes there are no differences between the underlying level of underenumeration in the LADs at this level. It must be noted that this is not likely to be true for the majority of Design Groups. This estimator will be biased where the relationship between the Census count and the CCS DSE-based count varies between LADs in the same Design Group. However, there are likely to be some Design Groups for which this assumption is not unreasonable – particularly those that contain only rural type areas.

6 Approach 3 (Fixed LAD Effects)

- 6.1 The small area estimation literature tends to focus on models that "explain" variation between areas by the addition of area level random effects. Such models are attractive where there are many small areas being estimated simultaneously. In the context of the One Number Census however, we do not expect any particular Design Group to contain more than 7 LADs. Consequently the argument for modelling variability between these "small areas" using a random effect formulation is less attractive. Here we explore the possibility of including fixed LAD effects in our ratio-type model.
- 6.2 The fixed effect model we use is one that includes an overall age-sex effect (defined at collapsed category level) and an LAD specific effect to distinguish between the LADs. These LAD effects are assumed to cancel out at Design Group level. The approach is implemented separately for each HtC stratum within a Design Group. The model specification underpinning this approach is:

Model

$$Y_{kadl} = (\theta_{cd} + \gamma_{dl}) X_{kadl} + \varepsilon_{kadl} \sqrt{X_{kadl}} \quad a \in C$$

$$\text{Var}(Y_{kadl} | X_{kadl}) = \sigma_a^2 X_{kadl}$$

$$\text{Cov}(Y_{kadl}, Y_{jbedl} | X_{kadl}, X_{jbedl}) = 0 \text{ for all } j \neq k$$

$$\text{Estimator} \quad \hat{T}_{adl} = \sum_{k \in S_{dl}} Y_{kadl} + \sum_{k \notin S_{dl}} (\hat{\theta}_{cd} + \hat{\gamma}_{dl}) X_{kadl} \text{ for } a \in C.$$

- 6.3 The requirement that LAD effects cancel out at the Design Group level is implemented by imposing the constraint $\sum_{l \in G} \gamma_{dl} = 0$. This means that we are fitting an overall Design Group age-sex slope parameter, and then making an adjustment to this slope to take account of the differences between the LADs. Let L_d denote the number of LADs "represented" in HtC stratum d . For $a \in C$, let \mathbf{Z}_{kadl} denote the vector of length $C + L_d$ which contains the value X_{kadl} in

position c of the first C components and in position l of the remaining L_d components. All other components of this vector are zero. Let \mathbf{Z}_d denote the matrix whose rows correspond to all the values of \mathbf{Z}_{kadl} defined by the age-sex categories and LADs represented in HtC stratum d . Furthermore let Φ_d denote the $C + L_d$ vector with θ_{cd} in position c of the first C components and γ_{dl} in position l of the remaining components. Finally, let \mathbf{V}_d denote the diagonal matrix defined by the values $\max(1, X_{kadl})$ in HtC stratum d . Then the above model can be written in matrix form as

$$\begin{aligned} E(\mathbf{Y}_d | \mathbf{X}_d) &= \mathbf{Z}_d \Phi_d \\ \text{Var}(\mathbf{Y}_d | \mathbf{X}_d) &= \sigma_d^2 \mathbf{V}_d \\ \mathbf{A}'_d \Phi_d &= 0 \end{aligned}$$

where \mathbf{Y}_d denotes the vector of CCS counts Y_{kadl} and \mathbf{A}_d is a vector of length $C + L_d$ with 0s in the first C positions and 1s in the remaining L_d positions. The generalised least squares solution to fitting this model is then

$$\hat{\Phi}_d = (\mathbf{Z}'_d \mathbf{V}_d^{-1} \mathbf{Z}_d)^{-1} \left[\mathbf{I} - \left\{ \mathbf{A}' (\mathbf{Z}'_d \mathbf{V}_d^{-1} \mathbf{Z}_d)^{-1} \mathbf{A} \right\}^{-1} \mathbf{A} \mathbf{A}' (\mathbf{Z}'_d \mathbf{V}_d^{-1} \mathbf{Z}_d)^{-1} \right] (\mathbf{Z}'_d \mathbf{V}_d^{-1} \mathbf{Y}_d)$$

- 6.4 One problem with the above solution that \mathbf{Z}_d is not of full rank. This is dealt with by using generalized inverses in the above formula. Its actual computation can be carried out using SAS PROC REG, although for comparative purposes two models will be tested – one with the restriction that the LAD effects sum to zero (which we will call C FIXED) and the other without this requirement (simply referred to as FIXED).

7 Approach 4 (Random LAD Effects)

- 7.1 This is the standard way of allowing variability across LADs within a Design Group while at the same time "borrowing strength" across these LADs for estimation. Its rationale is the same as that underpinning the fixed LAD effect model used in Approach 3 above. However here the LAD effects are assumed to be normally distributed with expectation zero rather than constrained to sum to zero. Again, collapsed age-sex categories are used for fixed effects and separate models are fitted within each HtC stratum using SAS PROC MIXED. The underlying model assumed is:

Model

$$\begin{aligned} Y_{kadl} &= (\beta_{cd} + u_{adl}) X_{kadl} + \varepsilon_{kadl} \sqrt{X_{kadl}}; \quad a \in c \\ \begin{pmatrix} u_{adl} \\ \varepsilon_{kadl} \end{pmatrix} &\sim \text{independent } N \left[\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{bmatrix} \sigma_{ud}^2 & 0 \\ 0 & \sigma_{\varepsilon d}^2 \end{bmatrix} \right] \text{ across LADs} \end{aligned}$$

It follows

$$\text{Cov}(Y_{kadl}, Y_{jbem} | X_{kadl}, X_{jbem}) = \begin{cases} \sigma_{ed}^2 X_{kadl} + \sigma_{ud}^2 X_{kadl}^2 & \text{if } j=k, a=b, d=e \text{ and } l=m \\ X_{kadl} X_{jbem} \sigma_{ud}^2 & \text{if } a=b, d=e \text{ and } l=m \\ 0 & \text{otherwise} \end{cases}$$

Estimator $\hat{\tau}_{adl} = \sum_{k \in S_{dl}} Y_{kadl} + \sum_{k \notin S_{dl}} (\hat{\beta}_{cd} + \hat{u}_{adl}) X_{kadl}$ for $a \in c$.

7.2 One problem with using SAS PROC MIXED for fitting this model is dealing with the heteroskedasticity implicit in its specification. The standard assumption in random effects (or multilevel) models is that the random effects have constant variances. We consider four different ways of handling this problem:

7.2.1 Option 1 - Ignore the heteroskedasticity in the observed data and fit the constant variance model $Y_{kadl} = (\beta_{cd} + u_{adl})X_{kadl} + \varepsilon_{kadl}$. This will be referred to as the Homogenous Mixed (H MIXED) model throughout this paper.

7.2.2 Option 2 - Scale the response variable (Y) and dependant variable (X) by the reciprocal of \sqrt{X} and then fit the homoskedastic model in option (1) above to these scaled variables. This will be known as the Scaled Mixed (S MIXED) model.

7.2.3 Option 3 - Model the heteroskedastic level one error term ε_{kadl} as a level two error – i.e. fit the model $Y_{kadl} = (\beta_{cd} + u_{adl})X_{kadl} + \xi_{cd}\sqrt{X_{kadl}} + \varepsilon_{kadl}$. The ξ_{cd} term is then an additional random effect $\sim N(0, \sigma_{\xi_{cd}}^2)$ used to model the heteroskedasticity in the data. This will be called the Combined Mixed (C MIXED) model.

7.2.4 Option 4 - Use the WEIGHT option in PROC MIXED to produce weighted least squares estimates – the SAS documentation does not give details sufficient to determine whether this will produce the heteroskedastic model solution we require. This last alternative will be referred to as the Weighted Mixed (W MIXED) model.

7.3 Option 2 ran into problems where the census count X is zero (we cannot divide by zero). Therefore in options 2 and 3 above \sqrt{X} was replaced by $\max(1, \sqrt{X})$.

8 SIMULATION METHODOLOGY

8.1 In order to evaluate the models proposed above, each was implemented within a set of simulations in order to compare their relative performance. The simulations were carried out on the same basis as those used for previous One Number Census research (see Brown *et al* (1999)), albeit on a slightly more extended basis as described below.

Methodology

8.2 Census underenumeration was simulated by each individual in the population being given a probability of being counted in a census. These probabilities depend on individual characteristics such as age and sex and are based on research by the ‘Estimating With Confidence Project’

(Simpson *et al*, 1997) following the 1991 Census. They also vary by the HtC index score and the census variable 'Primary Activity Last Week'. Whole households are also assigned probabilities of being counted in the census. These are based on averaging the individual probabilities associated with the adults within the households. Household probabilities also vary according to the tenure of the household and the household size. Each individual and household is also assigned a factor that defines the differential nature of response in the CCS. These mirror the same pattern as the census probabilities.

- 8.3 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 missed. In these simulations the census and CCS outcome for households and individuals are independent.
- 8.4 Coverage in the CCS is set at approximately 95 per cent for households with 98 per cent of individuals within those households being counted. The CCS design for the selection of 1991 EDs was fixed throughout each simulation, and 5 postcodes per ED were randomly taken at the second stage. The census counts and CCS counts are then used within a 1991 ED cluster level Dual System Estimator to provide the best estimate of the population for the sampled postcodes. This assumed that the matching process required to compute the DSE was perfect, and that the census and CCS counts were independent. For each Census and corresponding CCS, the proposed LAD estimation approaches were implemented.
- 8.5 For each census ten CCS postcode samples are selected based on the design. The whole process is repeated for one hundred independent censuses.

Simulation Data

- 8.6 It is imperative that any ONC small area estimation strategy is capable of delivering accurate estimates in all different circumstances. Therefore, all of the methodologies proposed were tested under different conditions to assess how robust they were. Firstly, four different design groups were used containing two, three, four and five LADs respectively. It is important to evaluate how the alternative methodologies coped with differing numbers of LADs, as some methods may work better with fewer LADs. Furthermore, the Design Groups cover a wide range of demographic types, some including mostly urban areas and some mostly rural. The makeup and characteristics of the four design groups are shown in **Table 1**.

Table 1 – Design Groups used in the simulation study.

| Design Group | LADs | Simulation LAD resident population | 1991 Area Type* | Estimated 1991 underenumeration (percent)* |
|---------------------|--------------|---|------------------------|---|
| West Yorkshire | Calderdale | 191,654 | Other Metropolitan | 3% |
| | Kirklees | 376,011 | Other Metropolitan | 3% |
| Hampshire A | Eastleigh | 106,886 | Mixed Urban/Rural | 2% |
| | Southampton | 199,319 | Non Metropolitan City | 5% |
| | Test Valley | 101,359 | Mixed Urban/Rural | 2% |
| Hampshire B | Fareham | 99,367 | Mixed Urban/Rural | 2% |
| | Gosport | 74,979 | Urban | 2% |
| | Havant | 120,263 | Mixed Urban/Rural | 2% |
| | Portsmouth | 175,472 | Non Metropolitan City | 5% |
| Hampshire C | Basingstoke | 145,676 | Urban | 2% |
| | E. Hampshire | 103,450 | Mixed Urban/Rural | 2% |
| | Hart | 80,732 | Mixed Urban/Rural | 2% |
| | Rushmoor | 80,491 | Mainly Rural District | 2% |
| | Winchester | 96,300 | Mixed Urban/Rural | 2% |

* area types and estimated underenumeration taken from the 1991 Census Validation Survey: Coverage Report (HMSO, 1994).

- 8.7 For realism, each of the Design Groups are one of those proposed for 2001, as detailed in paper ONS(ONC(SC))00/10.
- 8.8 To further evaluate the robustness of the methodologies, the individual level probabilities used for the simulations were modified within each Design Group separately to produce different undercount patterns. These patterns were chosen to represent some of the possible situations that might occur in 2001.
- a) **Version 1:** 1991 patterns of census coverage, 95% CCS coverage of HHs, 98% CCS coverage of persons within HHs. This was the basic set of individual level probabilities as used in previous simulations;

- b) **Version 2:** As (a) but reduce the probabilities of being counted in the Census by raising to the power 1.9. This reduced the overall coverage and increased the differential patterns; and;
- c) **Version 3:** As (a) but reduce the probabilities of being counted in the Census in Basingstoke, Kirklees, Portsmouth and Southampton LADs by raising to the power 1.9. This attempted to introduce an overall LAD effect in an urban area where the undercount might have been quite high already.

8.9 Each Design Group therefore had three variants – each having different individual level probabilities of being counted in a census. Each proposed method was implemented for every variant of every Design Group. This resulted in the production of 8 different estimates for 12 separate simulation runs.

8.10 The mean levels of undercount by LAD is summarised for each Design Group in **Table 2**. The shaded figures highlighted in the table are those particular design group versions where the coverage in the LADs is most different – these will be the simulations that may tell us the most about how well the estimators coped with a different level of underenumeration across the areas. Note that we have highlighted version three of the West Yorkshire group – although the coverage is not that different in percentage terms, Kirklees has just under twice the population of Calderdale and therefore the actual undercount will be quite different (approximately 9100 compared to 16,000).

Table 2: Mean Simulation Coverage by LAD for each Design Group variant

| Local Authority District | Design Group | Mean Census LAD coverage (percentage to 2 d.p.) by simulation | | |
|--------------------------|----------------|---|-----------|-----------|
| | | Version 1 | Version 2 | Version 3 |
| Calderdale | West Yorkshire | 95.72 | 95.15 | 95.72 |
| Kirklees | West Yorkshire | 95.71 | 95.08 | 95.08 |
| Eastleigh | Hampshire A | 95.87 | 95.49 | 95.87 |
| Southampton | Hampshire A | 93.99 | 92.19 | 92.19 |
| Test Valley | Hampshire A | 95.62 | 94.68 | 95.62 |
| Fareham | Hampshire B | 96.06 | 95.66 | 96.06 |
| Gosport | Hampshire B | 95.01 | 93.61 | 95.01 |
| Havant | Hampshire B | 95.97 | 95.60 | 95.97 |
| Portsmouth | Hampshire B | 93.08 | 90.52 | 90.52 |
| Basingstoke | Hampshire C | 95.99 | 95.53 | 95.53 |
| E. Hampshire | Hampshire C | 95.87 | 95.27 | 95.87 |
| Hart | Hampshire C | 95.74 | 94.81 | 95.74 |
| Rushmoor | Hampshire C | 93.46 | 91.01 | 93.46 |
| Winchester | Hampshire C | 95.53 | 94.67 | 95.53 |

9 RESULTS

- 9.1 The key measure of performance of the different approaches is the Relative Root Mean Squared Error (RRMSE) of the total population estimates for the five year age-sex group within collapsed hard to count categories with each LAD. These actual error values can be calculated since the true value of these populations are known. The great benefit of using this particular measure of error is that it contains both a variance and bias component – and therefore our evaluation can take both into account. The mean of these errors across all 1000 simulations for each subgroup are used in the analysis. For presentational purposes, the natural logarithm (i.e. Log_e) of the Mean RRMSEs are utilised as these values are extremely skewed. A description of the RRMSE calculations can be found in Annex A.
- 9.2 This measure of error will give us information about the precision achieved over all the 1000 set of estimates from each simulation. However, we must also be careful to ensure we examine how variable each estimate is across all the simulations – the simulation error. The overall mean precision for a particular estimate may be low, but the distribution around the mean may be extremely skewed or spread out across a wide range.
- 9.3 Results for the initial simulation runs were examined, and it became clear that some of the estimation variants were not performing as well as similar estimators. For instance, the weighted random effects model performed better than all of the other models that included random effects. Therefore for presentational purposes, the decision was made to only present results for the following models:
- direct ratio estimator specified in Approach 1 – this is labelled as **Direct**;
 - simple synthetic estimator from Approach 2 – this is labelled as **Synthetic**;
 - constrained fixed effects model in Approach 3 – this is labelled as **Cons. Fixed**;
 - weighted mixed model from Approach 4 – this is labelled as **WMixed**.
- 9.4 To provide an initial overall evaluation of these methodologies, **Tables 3.1, 3.2, 3.3 and 3.4** show the overall mean Ln Relative Root Mean Square Error for each methodology in the Design Group variants. This is the mean of the Ln RRMSEs across the HtC index, age-sex groups and LADs. The tables indicate that overall, the Synthetic estimator has the lowest mean error for every design group variant. The second best is the random effects model, followed closely by the constrained fixed approach. As we would expect, the Direct ratio estimator performs the worst and becomes less accurate as the number of LADs within the design group increases.

Table 3.1 – Overall Mean Ln RRMSE by methodology for West Yorkshire variants.

| | West Yorkshire (2 LADs) | | |
|--------------------------|-------------------------|-----------|-----------|
| | Version 1 | Version 2 | Version 3 |
| Direct | 0.77 | 0.86 | 0.79 |
| Constrained Fixed | -0.22 | -0.07 | -0.12 |
| Synthetic | -1.03 | -0.94 | -0.74 |
| Weighted Mixed | -0.22 | -0.09 | -0.15 |

Table 3.2 – Overall Mean Ln RRMSE by methodology for Hampshire group A variants.

| | Hampshire Group A (3 LADs) | | |
|--------------------------|----------------------------|-----------|-----------|
| | Version 1 | Version 2 | Version 3 |
| Direct | 1.29 | 1.40 | 1.34 |
| Constrained Fixed | 0.35 | 0.57 | 0.52 |
| Synthetic | -0.25 | -0.05 | 0.04 |
| Weighted Mixed | 0.21 | 0.44 | 0.41 |

Table 3.3 – Overall Mean Ln RRMSE by methodology for Hampshire group B variants.

| | Hampshire Group B (4 LADs) | | |
|--------------------------|----------------------------|-----------|-----------|
| | Version 1 | Version 2 | Version 3 |
| Direct | 1.41 | 1.55 | 1.46 |
| Constrained Fixed | 0.44 | 0.70 | 0.63 |
| Synthetic | -0.01 | 0.25 | 0.33 |
| Weighted Mixed | 0.33 | 0.63 | 0.61 |

Table 3.4 – Overall Mean Ln RRMSE by methodology for Hampshire group C variants.

| | Hampshire Group C (5 LADs) | | |
|--------------------------|----------------------------|-----------|-----------|
| | Version 1 | Version 2 | Version 3 |
| Direct | 1.65 | 1.77 | 1.66 |
| Constrained Fixed | 0.56 | 0.80 | 0.58 |
| Synthetic | 0.01 | 0.24 | 0.04 |
| Weighted Mixed | 0.47 | 0.74 | 0.48 |

- 9.5 As discussed earlier, we are not necessarily looking for the method that provides the lowest overall precision since this is likely to be driven primarily by the accuracy of the design group estimates. Therefore, although the above tables demonstrate that the synthetic model has the lowest overall mean precision it is important to examine the distribution around this mean.
- 9.6 **Diagrams 1.1, 1.2, 1.3, 1.4 and 1.5** show the distribution of the age-sex group by hard to count index estimation errors for each of the LADs in the Design Group variants that were highlighted in **Table 2**. These variants are those that have the largest undercount differences between LADs in the Design Group. The box and whisker plots will provide information about how variable the precision is across the estimates that are produced. All of the remaining plots for the other Design Group variants are contained in Annex B. The red circles indicate outliers and the stars represent extreme values for the distribution.

Diagram 1.1 – Distribution of Ln RRMSEs for each Local Authority District in West Yorkshire version 3 (CY is Calderdale and CZ is Kirklees)

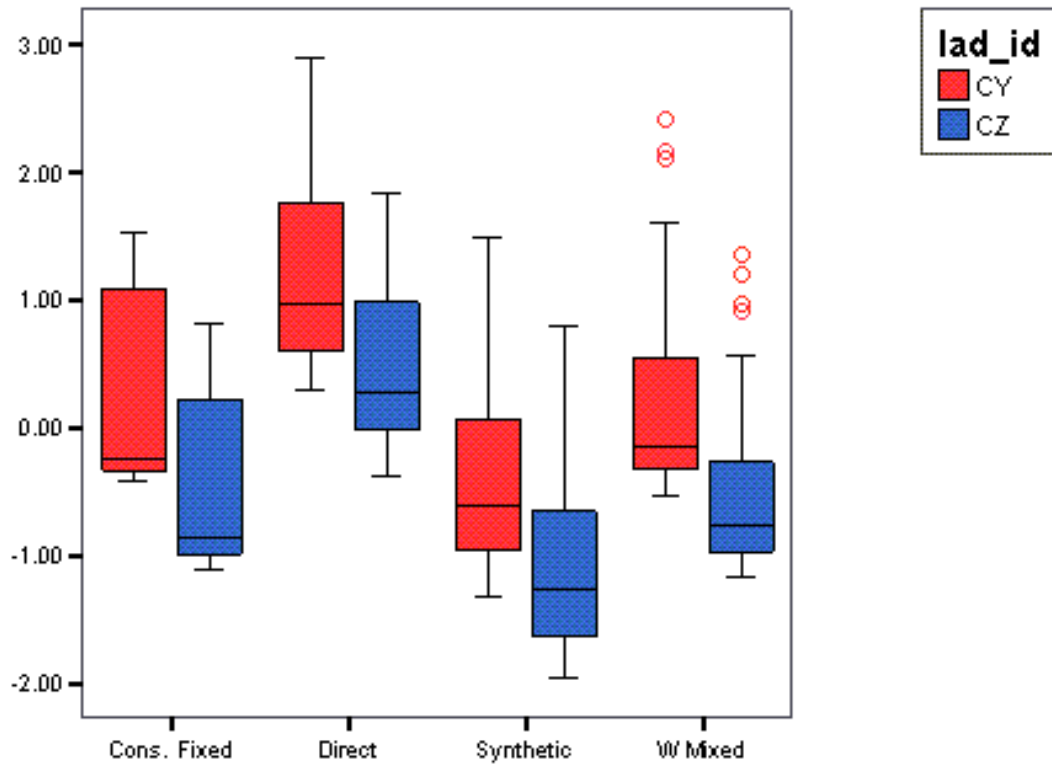
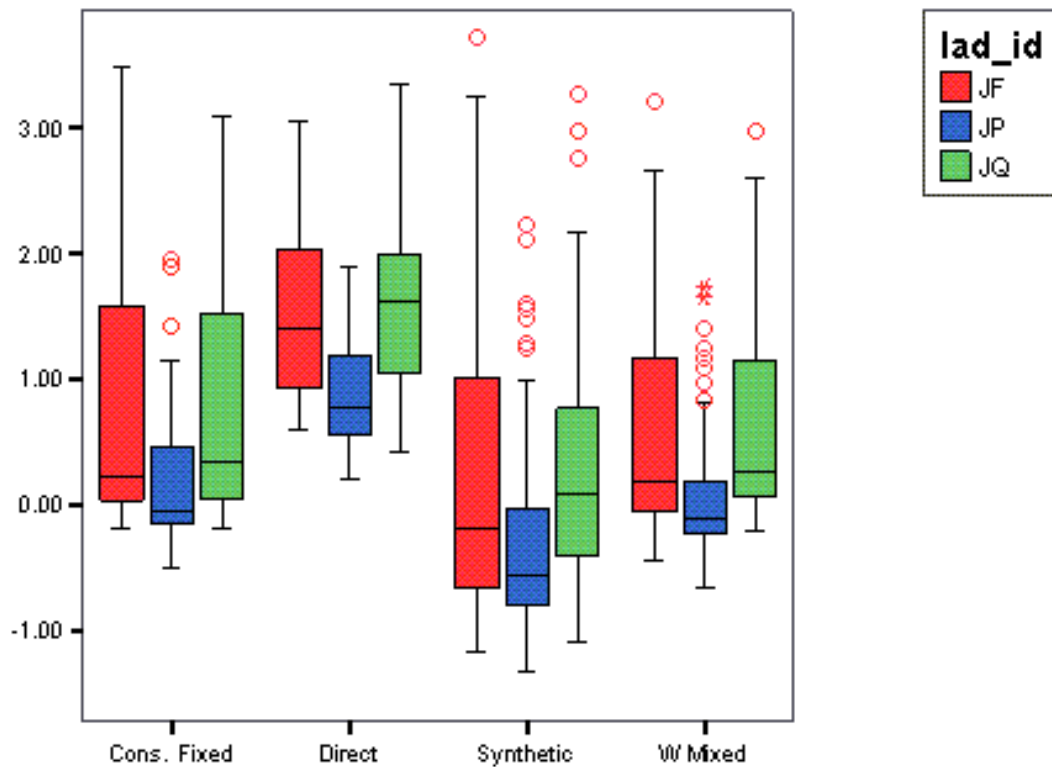


Diagram 1.2 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire group A



version 3 (JF is Eastleigh, JP is Southampton and JQ is Test Valley)

Diagram 1.3 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire group B version 2 (JG is Fareham, JH is Gosport, JK is Havant and JM is Portsmouth)

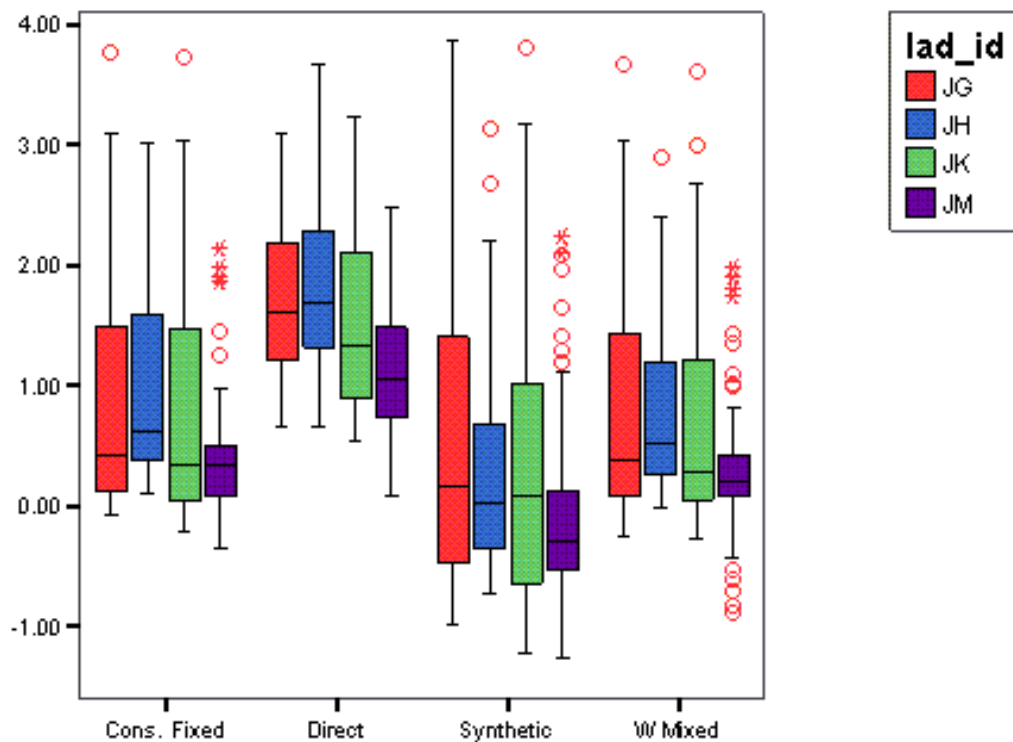


Diagram 1.4 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire group B version 3 (JG is Fareham, JH is Gosport, JK is Havant and JM is Portsmouth)

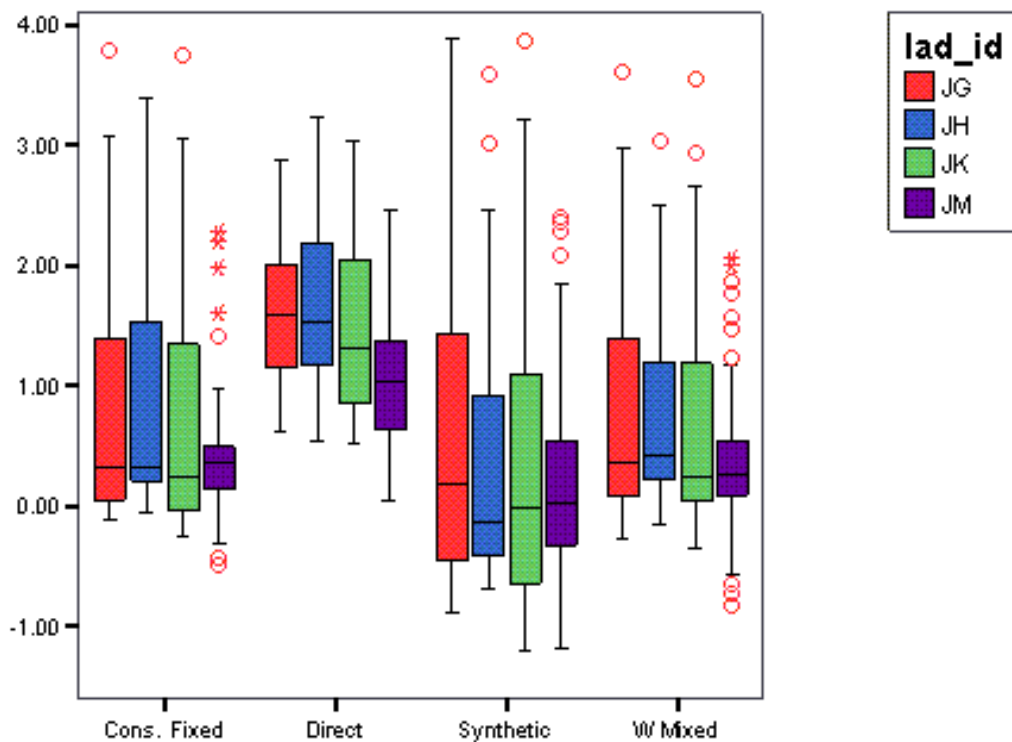
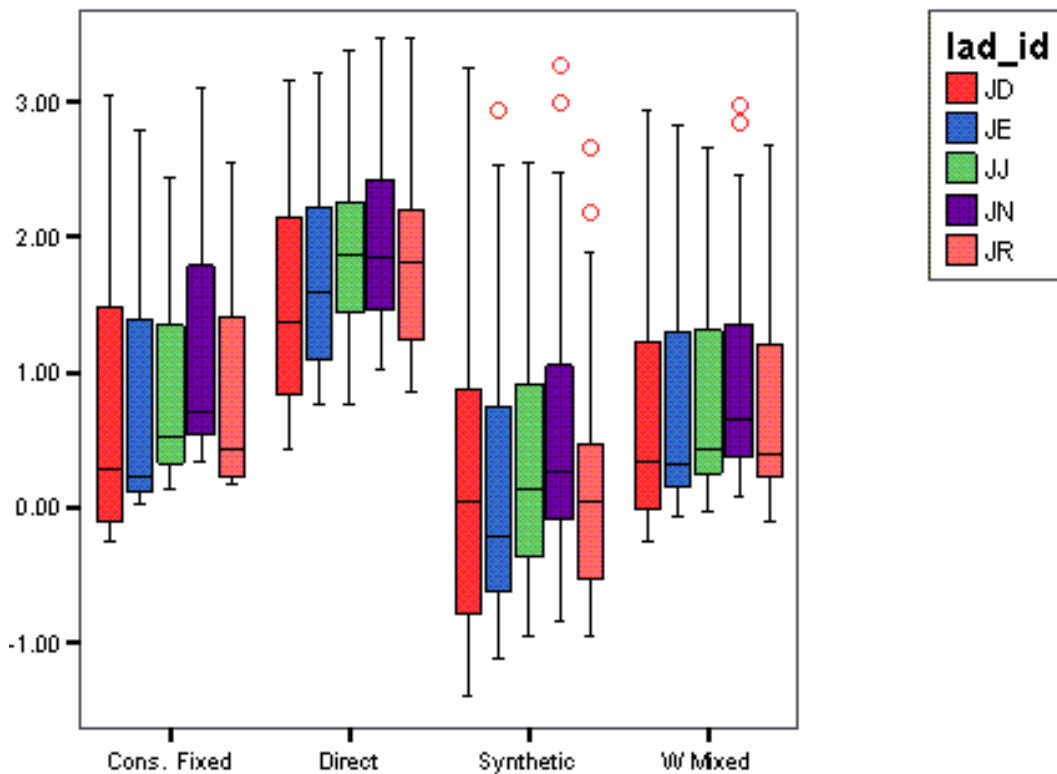
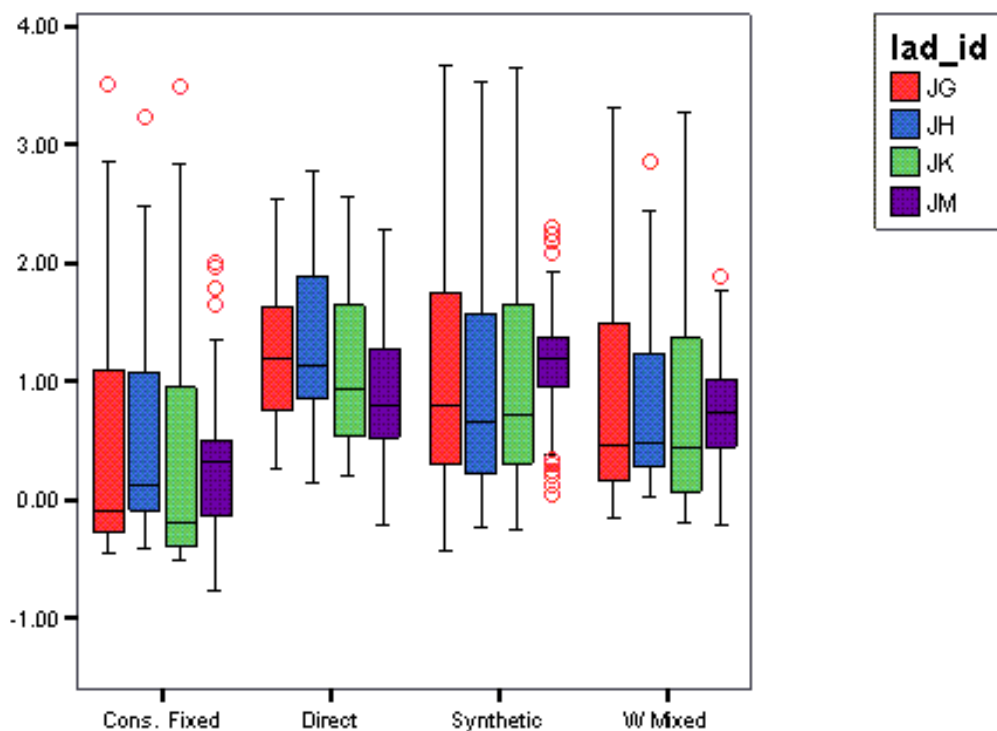


Diagram 1.5 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire group C version 2 (JD is Basingstoke, JE is East Hampshire, JJ is Hart, JN is Rushmoor and JR is Winchester)



- 9.7 The diagrams displayed on the previous pages show that while the synthetic estimator has the lowest mean, the distribution of the precision of the estimates has a larger spread than some of the other methodologies, such as the fixed alternative. However, the synthetic clearly produces more estimates of a higher precision than any of the alternatives – hence the tails toward the bottom of each of the plots. This feature is repeated throughout the box and whisker plots presented in Annex B.
- 9.8 The plots above represent those cases where there were large differences in the LAD census coverage. **Diagram 1.3** shows the results where there is a difference of around 5% between one of the small areas and the others. In this case, the synthetic approach is clearly not performing as well as in the other cases. Therefore, a further simulation was carried out to push the difference between the LADs further to assess the impact on the estimates. This additional research used Hampshire Group B version 4 data – this time with overall census coverages of 98.37%, 97.92%, 98.32% and 90.49% in the LADs - which has a difference of about 7.5% between one of the LADs and the rest.
- 9.9 The simulation was carried out on the same basis as before. **Diagram 2** shows the resulting distribution of the estimation errors. It shows clearly that the synthetic model has failed due to the assumption of no difference in coverage between LAD being violated. The conclusion drawn would be that the synthetic model is not particularly robust. This also seems the case for the model including the random LAD effect. However, the constrained fixed type approach has the greatest precision in this case, and would seem to be a sensible model to use if a large LAD effect were expected in any Design group.

Diagram 2 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire group B version 4 (JG is Fareham, JH is Gosport, JK is Havant and JM is Portsmouth)



9.10 Analysis of the agegroup estimates will also provide information on the performance of the estimators for these subgroups. **Diagrams 3.1, 3.2 and 3.3** display the mean Ln RRMSE across the hard to count index for each age-sex group within the Hampshire Group A version 2 LADs. Note that agegroups 1 to 18 are the male five year age groupings, and 19 to 36 are the female five year age groupings. The graphs clearly show that the synthetic estimator has the best precision across the age-sex estimates for each of the LADs within the design group. However, the accuracy varies between the subgroups quite a lot.

Diagram 3.1 - Mean Ln RRMSE by age-sex group for Eastleigh District of Hampshire Design Group A version 3

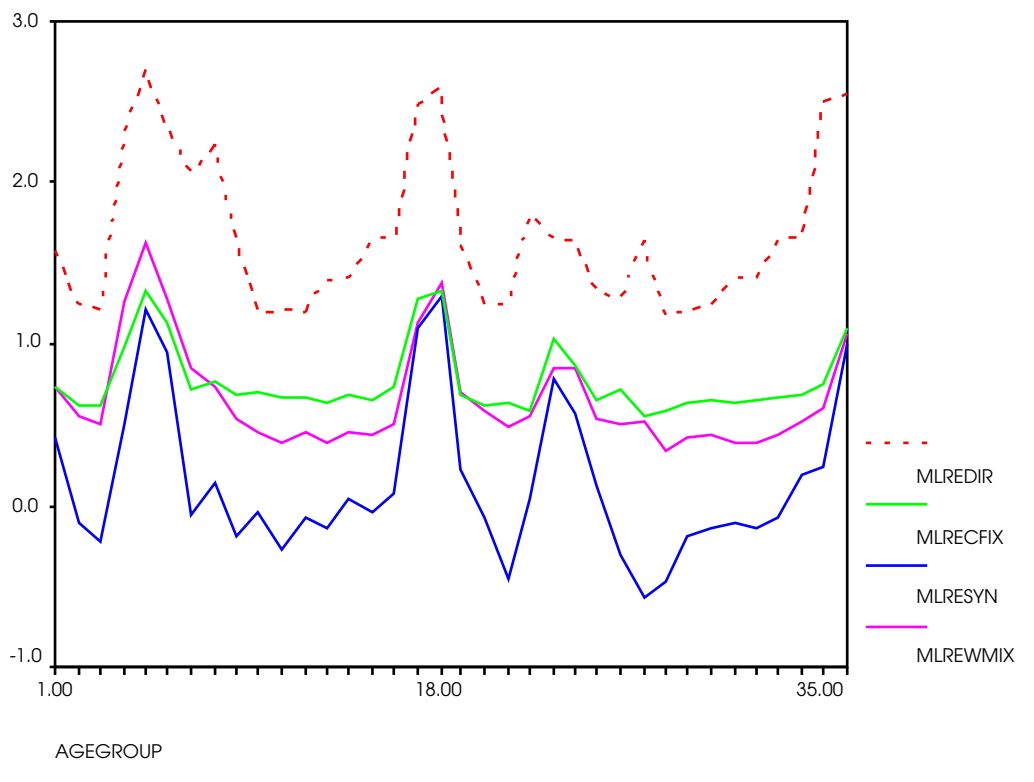


Diagram 3.2 - Mean Ln RRMSE by age-sex group for Southampton District of Hampshire Design Group A version 3

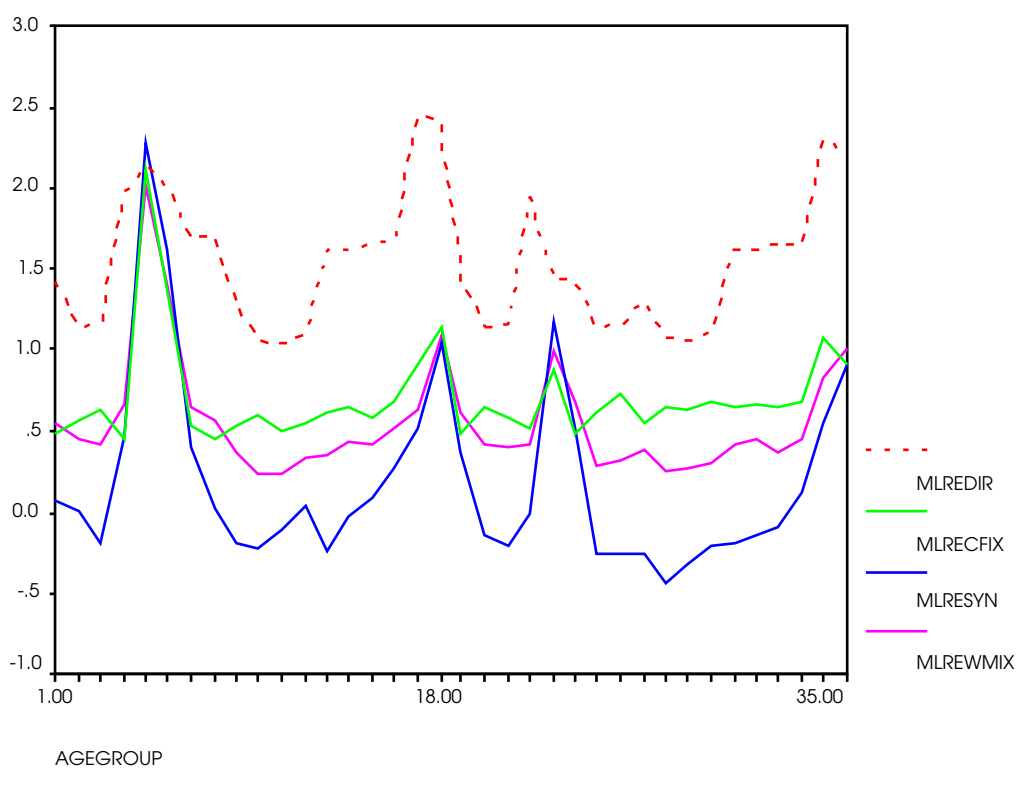
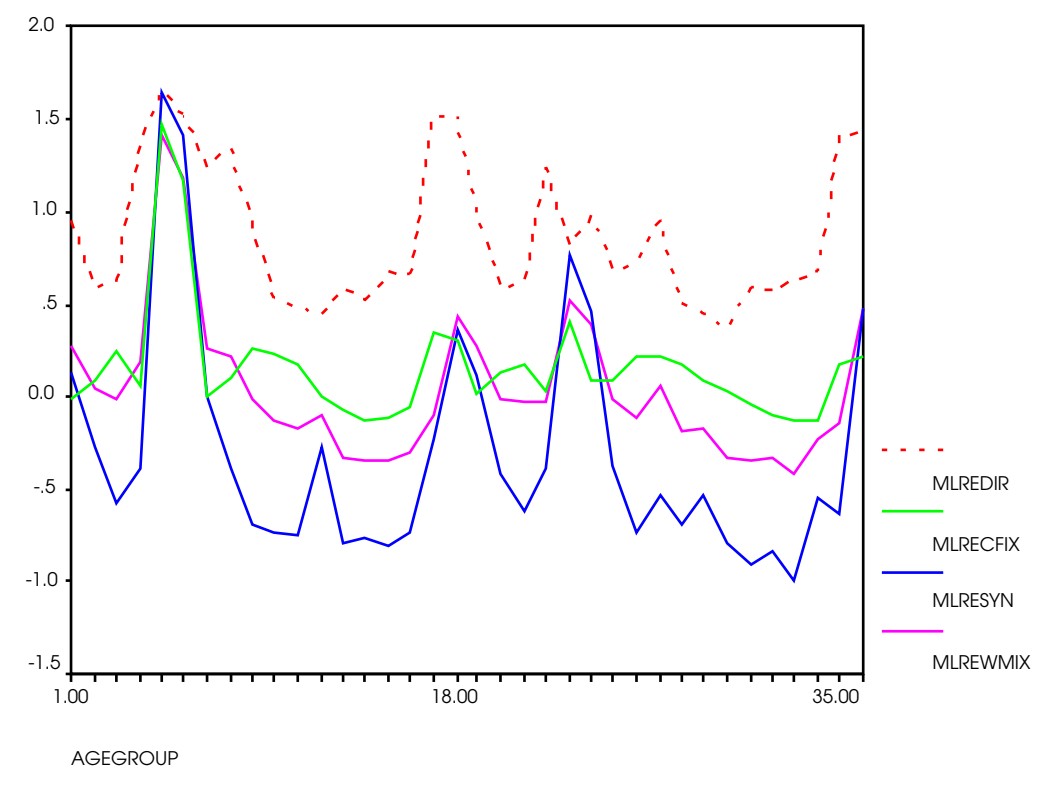


Diagram 3.3 - Mean Ln RRMSE by age-sex group for Test Valley District of Hampshire Design Group A version 3



9.11 In terms of finding a robust estimator, the diagrams do show that the fixed approach has the least variability around its mean for all age-sex groups. This is a desirable quality, as it indicates that the estimates for the different agegroups are likely to have a similar level of error associated with them. Similar patterns are repeated across the other design groups, although the graphs are not presented in this paper.

10 CONCLUSIONS AND RECOMMENDATIONS

- 10.1 The results clearly show that for the majority of Design Groups, a simple synthetic estimator provides results with the greatest precision. Furthermore, while there is no great LAD effect, it is consistently better across the age-sex estimates for each LAD. However, when a significant LAD effect is present, this type of estimator does not become clearly optimal. The point at which the synthetic fails is when the overall census coverage between the LADs differs by about 6% to 7%. Therefore this estimator is clearly not particularly robust.
- 10.2 The random effects approach also runs into some problems when a large LAD effect is introduced. This may be because of the assumption that the LAD specific random effects are normally distributed. When there is a single LAD with a very low coverage, this is likely to produce a skewed distribution and hence the estimator does not perform so well.
- 10.3 The estimators that are robust under all circumstances simulated in the study are the direct ratio estimator that does not borrow strength and a model including a fixed LAD effect. As expected the direct approach is not as efficient as all the alternative models. As the number of LADs within

the design group increases and hence the sample size within each LAD decreases, the direct estimates become unstable and unreliable. Therefore the direct approach is not considered further.

10.4 Although the fixed alternative is not the most efficient of the options tested, it is robust to different census coverage patterns and provides estimates of a similar error level across age-sex groups and LADs. The robustness of this model is the most important aspect as there are likely to be many different conditions in 2001. In particular, there are a number of 2001 design groups where it is expected that the underenumeration will differ between LADs. While this is not desirable, grouping rural and urban areas together has been unavoidable due to the geographical spread of the population. Therefore the LAD specific fixed estimator is approach that will be able to be applied with confidence for the 2001 ONC estimation strategy.

10.5 It is therefore recommended that a ratio model including a LAD specific fixed effect be used in the production of the One Number Census Local Authority Population estimates.

10.6 Further work may be required to investigate whether gains in precision can be made by using a weighted combination of the fixed and synthetic approaches. The calculation of variance estimates will also need to be considered.

References:

Brown, J. J., Diamond, I. D., Chambers, R. L., Buckner, L. J., and Teague, A. D. (1999)¹ A methodological strategy for a One Number Census. *Journal of the Royal Statistical Society A* **162**, 247-267.

Ghosh and Rao (1994), Small Area Estimation: An Appraisal, *Statistical Science*, Volume 9, No 1, pp55-93.

Heady, Smith and Avery (1994), 1991 Census Validation Survey: Coverage Report, OPCS.

ONS(ONC(SC))00/01 (2000) One Number Census Methodology.

ONS(ONC(SC))00/03A (2000) Estimation strategy for Design Group Estimates by age and sex from the Census Coverage Survey.

ONS(ONC(SC))00/10 (2000) Design Groups for 2001

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

ANNEX A – Description of terms used

a) Relative Root Mean Square Error

For a given population quantity such as the total T with estimator \hat{T} , one can measure the mean accuracy of the estimator using the relative root mean square error (RRMSE) defined as:

$$\text{RRMSE}(\hat{T}) = \frac{1}{T} \left\{ \sqrt{\frac{\sum (\hat{T} - T)^2}{n}} \right\} \cdot 100$$

where the summation is carried out over all the n observations of \hat{T} .

The RRMSE is a measure of the mean level of variability for a population total, relative to the population total being estimated.

b) Mean Square Error

For a given population quantity such as the total T with estimator \hat{T} , one can measure the mean accuracy of the estimator using the mean square error (MSE) defined as:

$$\text{MSE}(\hat{T}) = \frac{\sum (\hat{T} - T)^2}{n}$$

where the summation is carried out over all the n observations of \hat{T} .

The MSE is a measure of the mean level of variability for a population total. The MSE includes a measure of the bias for the population estimate.

ANNEX B

Diagram B1 – Distribution of Ln RRMSEs for each Local Authority District in West Yorkshire version 1 (CY is Calderdale and CZ is Kirklees)

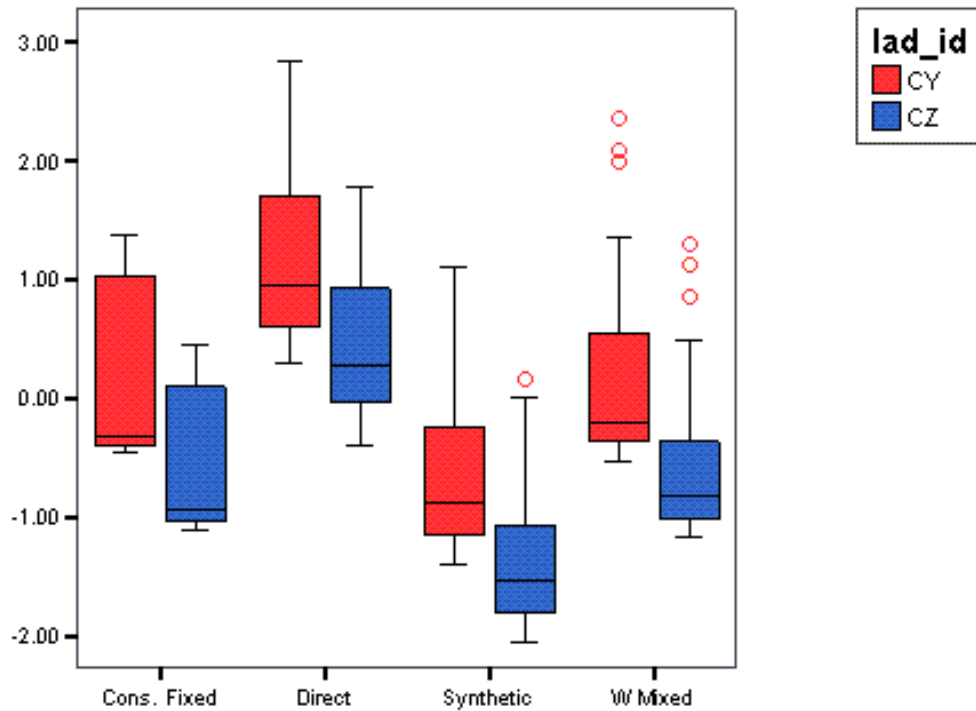


Diagram B2 – Distribution of Ln RRMSEs for each Local Authority District in West Yorkshire version 2 (CY is Calderdale and CZ is Kirklees)

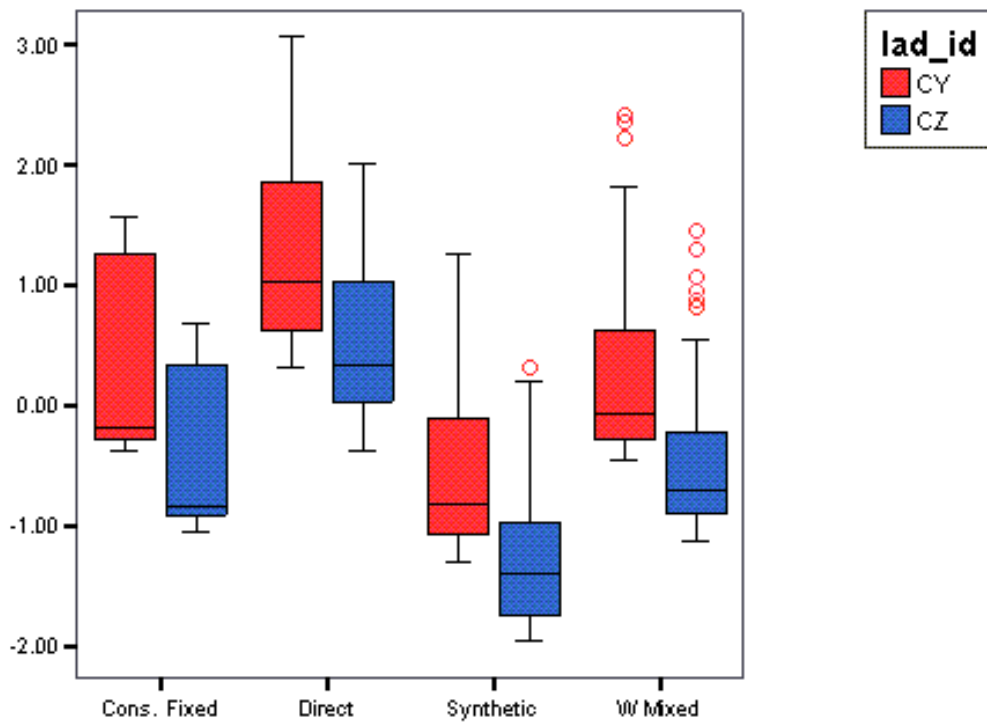


Diagram B3 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire group A version 1 (JF is Eastleigh, JP is Southampton and JQ is Test Valley)

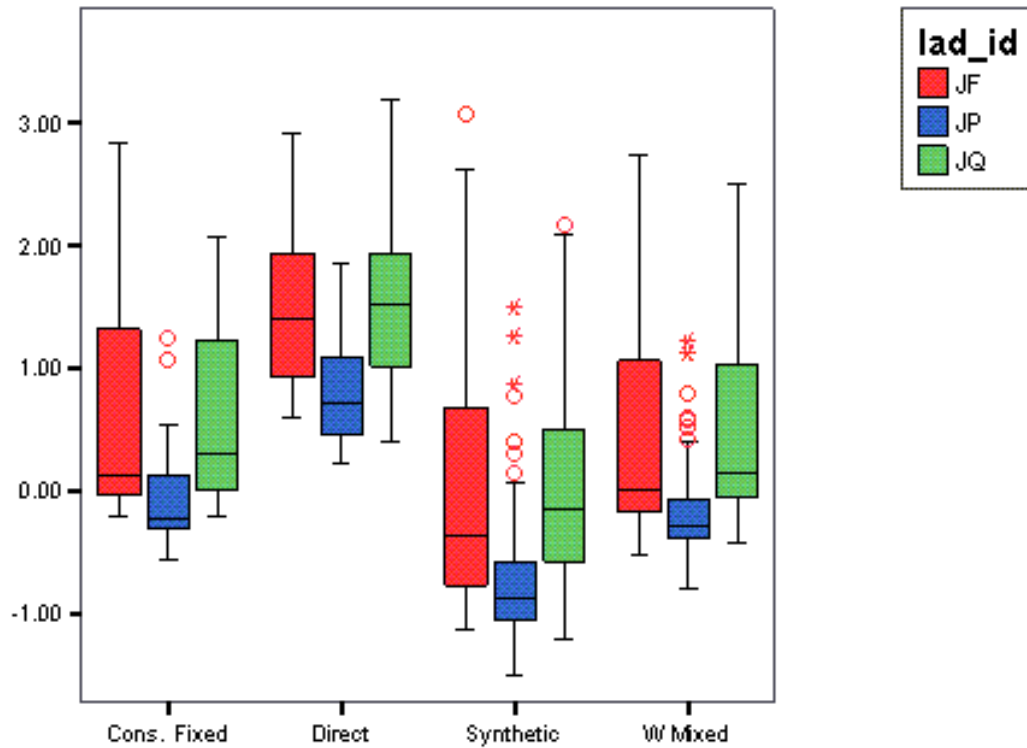


Diagram B4 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire group A version 2 (JF is Eastleigh, JP is Southampton and JQ is Test Valley)

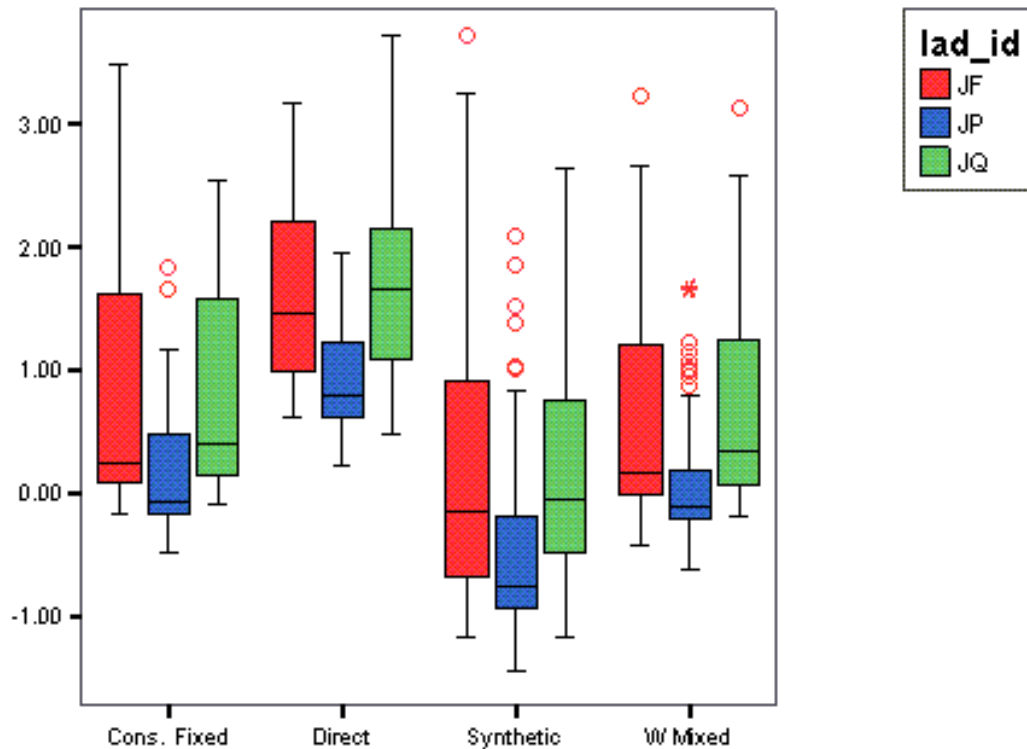


Diagram B5 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire group B version 1 (JG is Fareham, JH is Gosport, JK is Havant and JM is Portsmouth)

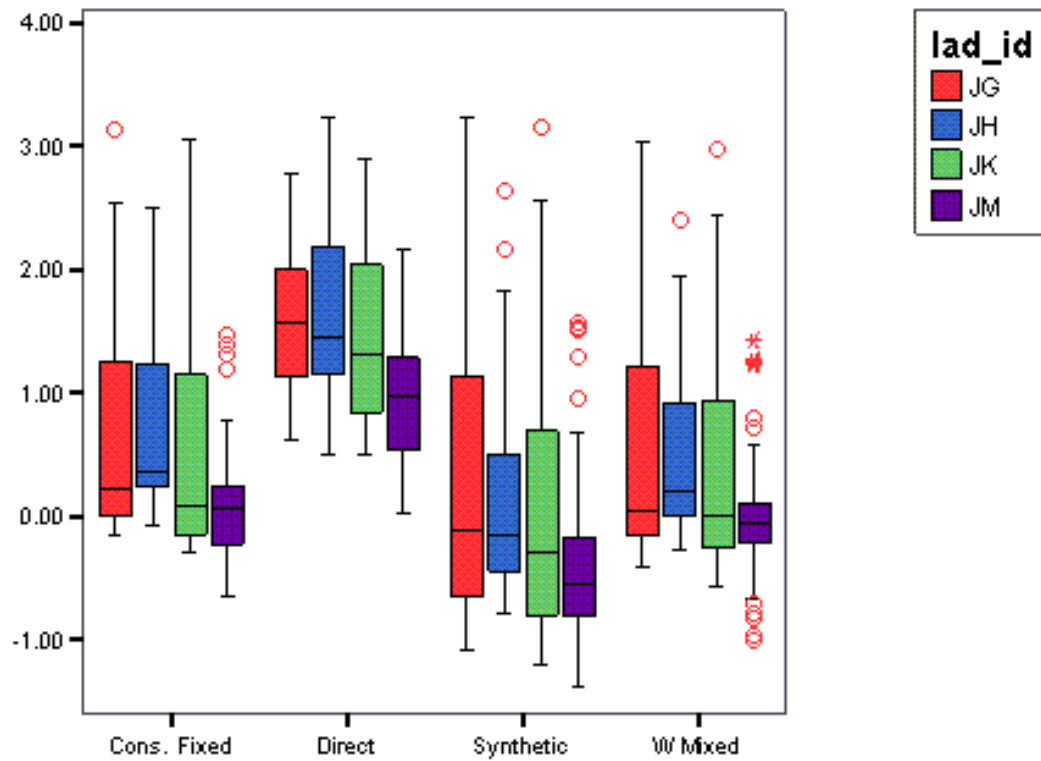


Diagram B6 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire group C version 1 (JD is Basingstoke, JE is East Hampshire, JJ is Hart, JN is Rushmoor and JR is Winchester)

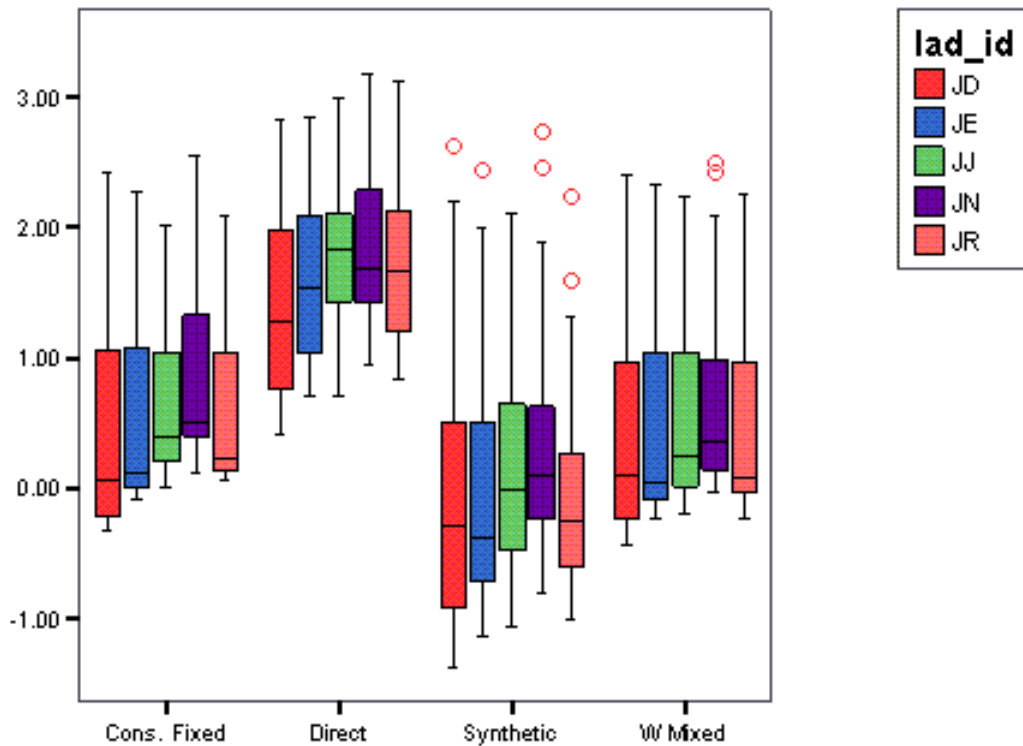


Diagram B7 – Distribution of Ln RRMSEs for each Local Authority District in Hampshire group C version 2 (JD is Basingstoke, JE is East Hampshire, JJ is Hart, JN is Rushmoor and JR is Winchester)

