Small Area Estimation Models for Disability

Methodology Advisory Committee

21 November 2003

I would like to express my thanks to Centrelink, FACS and, in anticipation, AIHW and disability administrators for assisting in making auxiliary data available for this empirical study. In addition I would like to express my appreciation to the project instigators / mentors Ken Tallis and Jon Hall for the support and advice they have given; to Robert Clark and Stephen Carlton for their discussions and technical advice and to Luke Samy, Sarah Vincent and Joanne Baker for their work on the project.

Main Questions for the Methodology Advisory Committee (MAC)

Motivation

1. This paper discusses possible methodological approaches for an empirical study on modelling small area estimates (SAE's) of disability using the ABS Survey of Disability, Ageing and Carers (SDAC) together with other external administrative data sources as auxiliary variables. External users are interested in small area estimates for five categories of disability (sensory, intellectual, physical, psychological and other) at the Statistical Local Area (SLA) level, which are mainly either Local Government Areas (LGAs) or parts thereof. Of all SLA's across Australia, 62% contain dwellings selected in SDAC.

2. Small area estimation is a new area and important part of Analysis Branch's work program. Many questions are exercising us at the moment, just a few of which are listed below. Any light MAC can shed on any of these, or other related areas would be gratefully received. But in particular, do MAC have views on:

<u>Broad Issues</u>

1. Does the broad approach seem reasonable?

2. Does the prioritisation of our shopping list of desirable features for a model appear reasonable? Where can we draw the line that would give us 85% of the efficiency gains?

3. We assume in the models that propensities to belong to each disability category are independent of each other. How can we guard against breaches of this assumption?

   Although each person is coded to only their main disability, it is possible for masking to occur in the case of multiple disabilities. To give a hypothetical example, persons with an intellectual disability due to a head injury might be more likely to report head injury for fear of stigmatisation. This phenomenon, if it occurs, is important as it leads to a violation in the assumption of independence between categories.

Choice of Model

4. Can MAC comment on the pros and cons of using generalised linear models (GLM's) with an underlying Poisson distribution for counts of disability as opposed to a multinomial logit regression?

5. How much trade-off between reliability and model interpretability is there likely to be as a result from using a generalised linear mixed model (GLMM) with underlying Poisson, say?

6. Are there any other model alternatives to the ones suggested in this paper worth considering?

Ecological Fallacy

7. Privacy requirements have prevented us obtaining unit level auxiliary data from other government departments. Broadly, what sacrifice in efficiency is likely to result from using area level rather than unit level models? Related to this question is how significant is the ecological fallacy relating to models at different levels likely to be?

Random Effects

8. Is it worth considering GLMM models incorporating random effects in addition to GLM's with fixed effects only? If we are evaluating both mixed and fixed effects models, what is the best way of comparing their goodnesses of fit.

Model Unidentifiability

9. MAC comment is sought on insights into how to deal with lack of model identifiability when trying to apply highly complex models under either a frequentist, empirical Bayes (EB) or hierarchical Bayes (HB) approach.

Model Validation / Validation of Output

10. Generally, after testing model fit, how can we best validate the modelled SAEs? ie check they look sensible. We plan to gather comments from disability administrators on how well the estimates correspond with their practical knowledge. But another possibility is to simulate a population and then compare modelled SAE's against known totals. What are the drawbacks with such an approach?

Design Informativeness

11. There is of course the issue of informativeness of the sample design when fitting these models. One approach is to include all design variables (state, CD measure of size, block size, area type etc) as explanatory variables in the model, however this may reduce parsimony and impact adversely on goodness of fit. Pfeffermann and Shverchkov (1999) give a framework for handling informativeness in the estimation of model parameters, although it applies only to single stage designs. This framework provides a more preferred approach, but needs to be extended to multi-stage designs where theoretical work is underway. Do MAC members have any suggestions on this issue?

Choice of Priors

12. We plan to use HB as an alternative to maximum likelihood methods of estimating model parameters. To test the robustness of the model to choice of priors we intend trying out a range of informative and non-informative priors. Does MAC have any suggestions on choosing reasonable informative priors for model parameters?

Background

3. The Small Area Estimation (SAE) Practice Manuals Project commenced in February 2003 with a view to increasing ABS capability in satisfying the growing demand for small area statistics by:

- expanding ABS knowledge and understanding of SAE methods,
- better melding the theoretical knowledge of SAE techniques with the practical issues of maximising accuracy, subject to:
    - affordability,
    - ease of implementation,
    - interpretability and
    - explainability to clients.
- provide a framework, in the form of the SAE Practice Manuals, for:
    - promulgating SAE methods and practices,
    - capturing the growing experience and intelligence of SAE techniques and processes as it applies to the ABS context, and
    - standardising and focusing the ABS' whole approach to meeting user demand for SAE's.

4. The SAE Practice Manuals will be targeted at a broad audience of technical and non-technical ABS employees. The manuals will include chapters on how to:

- discern clients' real data needs as opposed to requests and whether SAE's are in actual fact required,
- advise clients on the fitness for use of SAE's and the assumptions underpinning the methodology,
- assess the required level of quality for SAE's and determine what techniques are therefore most appropriate.
- find out what auxiliary data is potentially available and their associated quality and limitations
- choose from the array of SAE techniques available, the assumptions involved in each and what quality requirements and data contexts they are suitable for.
- validate the quality of SAE's during modelling and the clearance process prior to releasing the data.

5. A key aspect of the project will be two empirical studies into SAE methods the results of, and experience with, which will feed into the manuals. The topics for those studies are disability and retail sales/trade. This MAC paper will only focus on the disability empirical study.

6. Small area estimates of the incidence of disability based on the ABS Survey of Disability Aging and Carers (SDAC) were produced in 1994 and 2000. The 1994 work was based on 1988 and 1993 SDAC data. The consultancy provided in 2000 was based on the 1998 SDAC and used demographic benchmarks as auxiliary variables.

Purpose

7. MAC comment is sought on which models give the most accurate small area estimates of disability counts in light of practical considerations such as cost effectiveness, ease of implementation, interpretability and explainability to clients. We are still at the initiation phase of the current disability empirical study. This paper therefore focuses on assessing potential methodological approaches and identifying issues that need to be addressed rather than reporting results.

8. For the remainder of this paper, we firstly outline the data sources we intend to use for the small area models, then draw up a shopping list of desirable features for our repertoire of small area models. Four potential models are then discussed.

The Data

SDAC 1998 data

9. SDAC is the main source of data on disability collected by the ABS. It will be the source of response variable data in the modelling of small area estimates. SDAC is a multi-stage household survey which, at first stage, selects a sample of census collectors' districts (CD's) with probability proportional to size (PPS). Each selected CD is formed into blocks of approximately equal size (based on permanent landmarks such as roads and rivers etc). At second stage a block is selected, again PPS. In the third stage, a systematic sample of dwellings (referred to as a cluster) is selected throughout the block.

10. All in-scope persons (broadly, permanent residents aged over 15 not in the permanent defence forces) are surveyed from each dwelling in the selected cluster. SDAC uses the "any responsible adult" (ARA) methodology whereby the ARA who comes to the door is asked to respond for other in-scope persons in the dwelling.

11. Close to 43,000 people throughout most of Australia (excluding remote and very remote areas) were surveyed in SDAC 1998, comprising 37,000 from the private dwelling (PD) component and 6,000 from specific non-private dwellings (referred to as special dwellings (SD's)) such as homes for the aged and retirement homes. The design for the SDAC PD component is referred to as a "half cluster design" as it includes half the CD's selected in the Monthly Population Survey (MPS) whereas the SD component is a 8 cluster design which surveys 8 times the MPS sample for those types of SD's in-scope of SDAC.

12. The data published from SDAC includes estimates for each disability type by state, level of severity, age and sex.

13. Appendix 2 shows boxplots of RSE's for post-stratified estimates of types of disability by Statistical Local Area (SLA). Appendix 3 shows boxplots of RSE's for types of disability, this type by the broader geographic region, Statistical Sub-Division (SSD). These post-stratified estimates use demographic benchmarks at the level of State by capital city/non-capital city by age by sex (these are at a broader level than SLA or SSD). Such benchmarking might lead to biases if the SLA (or SSD) sample has different demographic characteristics (relative to its population) than that of the part of the State it is benchmarked to.

14. In short, only 819 of the 1332 SLA's in Australia are represented in the SDAC sample. Of these, over 75% have RSE's for the variable "any impairment" of greater than 25%. Hence the need for small area estimation models to provide reliable estimates for all SLA's.

Auxiliary Data Sources

15. The following table summarises the main sources of auxiliary data we are considering for the disability empirical study. The main issues concerning each data source are also discussed briefly. An overview of each data source collection can be found at Appendix 4.

Data Item	Level of Data	Source	Issues
receiving home assistance	SLA by age by sex by disability type by level of severity	AIHW Commonwealth State/Territory Disability Agreement (CSTDA) - Minimum Dataset	● Most detailed auxiliary data. ● Data used to allocate government money for programs and focus / coordinate service provision between providers to areas of greatest need. ● does not cover over 65 age group well ● does not cover many remote areas well ● AIHW Ethics Committee to decide whether to release data to ABS on 5 November 2003.
Disability Support Pension (DSP)	postcode by disability type	Centrelink / FACS	● Every non-zero cell has a count of < 20 and has been censored with a "<20" label. ***** ● The data has been extracted from Centrelink payments system. The usefulness of the DSP data depends upon definitional differences with those of SDAC. DSP mainly defines disability in terms of ability to work whereas SDAC defines it in terms of a wider range of activities. ● does not cover over 65 age group well
Home and Community Care (HACC)	yet to be advised	Department of Health and Ageing	● is a good source of disability data for over 65 age group ● is said to be a better source of disability data for very remote areas of NT. ● quality of data unknown

Remoteness	SLA	ABS Australian Standard Geographic Classification	● currently contemplating whether to use remoteness as an explanatory variable, as a parameter in the model (& hence a data item on all the data sources referred to here) or as a level in the model so that random effects can be derived at level of remoteness. ● should be a good indicator of disability levels: disability counts tend to be low for the non-indigenous population in remote areas as these people tend to move to less remote areas where better services can be provided; disability counts may be high for the indigenous population in remote areas due to health issues and a lower tendency to relocate to non-remote areas.
Number of Cared Accommodation beds/rooms	SLA	ABS Special Dwelling Framework	● Includes Homes for the Aged, Retirement Homes and Homes - Other ● Homes - Other is a mix of welfare & disability cared accommodation. Can't separate the two. ● Occupancy levels are averages across the year and not very reliable.
Socio-economic Indexes for Areas (SEIFA)	SLA	ABS SEIFA publication based on Census data	● will be useful if disability levels are correlated with socio-economic status, either in terms of occupation or social disadvantage. ● currently contemplating whether to use SEIFA as an explanatory variable, as a parameter in the model (& hence a data item on all the data sources referred to here) or as a level in the model so that random effects can be derived at

			the SEIFA level. ● based on 1996 census but may change slightly over time.
Population Benchmarks	Currently State by capital city/non-capital city by age by sex.	ABS Demography Section	● Size and age by sex profile of the SLA may be a good predictor of disability. ● ABS Demography section do not produce benchmarks at SLA level ● Need to estimate SLA benchmarks from broader level Demography benchmarks and MPS estimates. ● Benchmarks also needed at SLA level if unit level predictor models are used, so that SLA estimates can be derived.

*** We plan to impute for censored DSP cells by constructing an offset-function (Lee and Eltinge, 1999) between the DSP and CSTDA distributions for > 20 and then extrapolate this to generate the <20 tail of the DSP distribution.**

Small Area Models to be Considered for Empirical Study

16. In this section we discuss issues in selecting models. We also propose a couple of candidate models without going into the mathematical detail. Bear in mind that the main objective of this empirical study is to apply a range of appropriate models and estimation procedures in order to develop an understanding of how they perform, not just in terms of accuracy but also in terms of ease of implementation, production costs and model interpretability. It won't necessarily be just one model that we want to take from this exercise but several, ranging from simple to more complex, to suit a range of client requirements or data contexts. We hope the understanding gained from the empirical study will be a starting point in extending this knowledge to other small area problems and data contexts.

17. We start by putting together a shopping list of desirable features we would like in SAE models for disability:

<u>Shopping List of Desirable Model Features</u>

a. the model should be appropriate for rare count data. eg GLM with underlying Poisson distribution (Greene (2000) p880) or multinomial logit.

b. Best use is made of unit level SDAC data and area level auxiliary data to give the most efficient model for prediction purposes.

c. a model can be found that gives sufficient quality, robust estimates and is relatively simple to apply and run as part of a potential SAE production system. (as the US Bureau of Labour Statistics has recently done (email from John Eltinge))

d. the model is parsimonious, is relatively easy to interpret and to explain to users of the data.

e. the model incorporates the best available predictors of disability and any appropriate interaction terms.

f. the model accommodates the multivariate nature of the dependent variable: disability type (head injury, physical, psychological, intellectual, sensory) by level of severity (mild, moderate, severe, profound). A multivariate model is preferred so that we can account for any correlations between categories.

g. the model is a mixed effects model in case we need to account for SLA level random effects. (The Hausman test can be used to test for the necessity of a random effects term, (Cameron & Trevedi (1998), p293)

h. the model takes account of the clustered multi-stage nature of the survey design

i. if necessary, the model can take account of correlated sampling errors. That is for each variable (category) correlations between sampling errors of different SLA estimates.

j. model ensures additivity of predicted SAE's to reliable estimates at a broader region level. (Pfeffermann and Bleuer (1993))

k. Model estimation takes account of any informativeness, if present, in the sample design (Pfeffermann and Sverchkov (1999))

l. If using an underlying Poisson, potential over-dispersion, if present, is allowed for.

plus the usual features of no multi-collinearity, homoscedasticity, normally distributed error terms

18. If the incorporation of many of the more complex features into the one model is warranted, it may be necessary to use Hierarchical Bayes methods (Monte Carlo Markov Chain) to estimate the model.

<u>The Basic Fay-Herriot Model (Rao, 2003)</u>

19. All the potential models A - D we consider below, are extensions of the basic small area level Fay-Herriot model. We briefly discuss this model before giving reasons for using this model as the foundation model for those models we consider.

20. Assume that the population statistic $\theta_i = g\left(\overline{Y}_i\right)$ for small area i, is a function $g(.)$ of population means for variable Y. Also assume that θ_i can be modelled as:

$$\theta_i = \mathbf{z_i^T}\boldsymbol{\beta} + b_i v_i, \qquad i = 1,\ldots\ldots\ldots,m \qquad \ldots\ldots\ldots\ldots\ldots\ldots\text{(i)}$$

where

$b_i, \quad i = 1,\ldots\ldots\ldots,m$	known positive constants
$\boldsymbol{\beta} = \left(\beta_1,\ldots\ldots,\beta_p\right)^T$	(px1) vector of regression coefficients
$z_i = \left(z_{1i},\ldots\ldots,z_{pi}\right)^T$ $i = 1,\ldots\ldots\ldots,m$	(px1) vector of small area specific auxiliary data
$v_i, \quad i = 1,\ldots\ldots\ldots,m$	small area specific random effects assumed to be iid with $E_M\left(v_i\right) = 0, \quad V_M\left(v_i\right) = \sigma_v^2\left(\geq 0\right),$ wrt the model M

21. Next, we assume that survey estimates $\hat{\theta}_i$ for θ_i can be modelled as

$$\hat{\theta}_i = \theta_i + e_i, \qquad i = 1,\ldots\ldots\ldots,m \qquad \ldots\ldots\ldots\ldots\ldots\ldots\text{(ii)}$$

where

e_i's	are sampling errors assumed to be iid with $E_P\left(e_i \mid \theta_i\right) = 0, \quad \mathrm{V}_P\left(e_i \mid \theta_i\right) = \psi_i,$ wrt the population and the conditional variances ψ_i assumed known.

22. Inserting (i) into (ii) we obtain the overall model

$$\hat{\theta}_i = \mathbf{z_i^T}\boldsymbol{\beta} + b_i v_i + e_i, \qquad i = 1,\ldots\ldots\ldots,m \qquad \ldots\ldots\ldots\ldots\ldots\ldots\text{(iii)}$$

where the random effects v_i and sampling errors e_i are assumed independent. The model given by (iii) is commonly referred to as the Fay-Herriot model.

Why the Fay-Herriot Model?

23. The main reasons for choosing Fay-Herriot as the underlying form for the models below are:

- models of the Fay-Herriot form are widely used in the small area estimation literature.

- Fay-Herriot models incorporate synthetic estimation models as a special case. Synthetic estimation models, which have been used widely in previous SAE work in the ABS, can be easily obtained from the Fay-Herriot model by removing the random effects term (Rao, 2003).

- The Best Linear Unbiased Predictor under the Fay-Herriot model can be shown to take the form of a composite estimator (Pfeffermann, 2002). The composite estimator, in this context, is a weighted average of the direct survey estimator and a synthetic estimate based on a generalised linear model fitted to the observed data at a broader area.

A. Multivariate Fay-Herriot Linear Area Level Model (Rao, 2003 p81)

Summary of Extensions of Model A from the Basic Fay-Herriot Model	
1	Model A is multivariate in $\boldsymbol{\theta}_i$ with a covariance structure for the sampling error and random effects terms. By taking account of the correlations between variables, Model A should therefore give efficiency gains over fitting separate univariate models for each variable.

Notation for Model A		
Descriptor	**Index**	**Range**
Small Areas: SLA	i	i = 1,.....,m
Disability Categories	r	r = 1,.....,R
Auxiliary Variables		$x_{i1},.....,x_{ip}$

24. In this approach, SDAC survey estimates for R disability categories, represented in vector notation as $\hat{\boldsymbol{\theta}}_i = \left(\hat{\theta}_{i1,.....}, \hat{\theta}_{iR}\right)^T$, are modelled for each SLA i where the $\hat{\theta}_{ir}$ are functions of SLA means. In our case these are post-stratified estimates using demographic population counts.

25. Then a linear random effects area level model takes the form:

$$\hat{\boldsymbol{\theta}}_i = \boldsymbol{\theta}_i + \mathbf{e}_i, \qquad i = 1,....,m \quad \text{..... (1)}$$

where

- $\mathbf{e}_i = \left(e_{i1,.....}, e_{iR}\right)^T$ are the sampling errors, distributed as independent r-variate normal

$$N_r\left(\mathbf{0}, \boldsymbol{\Psi}_i\right),$$

- $\boldsymbol{\Psi}_{\mathbf{i}}$ are known covariance matrices (conditional on $\boldsymbol{\theta}_{\mathbf{i}}$) which can be calculated from SDAC data.

26. The unknown true disability counts $\boldsymbol{\theta}_{\mathbf{i}} = \left(\theta_{i1,\ldots\ldots\ldots}, \theta_{iR}\right)^{T}$ are then modelled according to:

$$\boldsymbol{\theta}_{\mathbf{i}} = \mathbf{X}_{\mathbf{i}}\boldsymbol{\beta} + \mathbf{v}_{\mathbf{i}}, \qquad i = 1,\ldots,m \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \quad (2)$$

where

- $\mathbf{X}_{\mathbf{i}}$ is a (R x Rp) matrix of auxiliary variables with r'th row given by $\left(\mathbf{0}^{\mathbf{T}},\ldots\ldots,\mathbf{0}^{\mathbf{T}},\mathbf{x}_{\mathbf{ir}}^{\mathbf{T}},\mathbf{0}^{\mathbf{T}},\ldots\ldots,\mathbf{0}^{\mathbf{T}}\right)$.

  {Note if r=1 indicates sensory disability type then for SLA i, $\mathbf{x}_{\mathbf{i1}}^{\mathbf{T}}$ will have as its elements: the DSP sensory count, CSTDA sensory count, HACC sensory count, SD occupancy count, remoteness indicator, etc. giving a total of p auxiliary variates.}

- $\boldsymbol{\beta}$ is the Rp -vector of regression coefficients, and

- $\mathbf{v}_{\mathbf{i}}$ are the SLA level random effects, independent $N_r\left(\mathbf{0},\boldsymbol{\Sigma}_{\mathbf{v}}\right)$.

Comments on Model A:

1. incorporates SLA specific random effects $\mathbf{v_i}$ as well as sampling effects $\mathbf{e_i}$.

2. We would prefer to fit the $\hat{\boldsymbol{\theta}}_\mathbf{i}$ with a Poisson distribution which we suggest under model B. The Poisson distribution is regarded as the "benchmark model for count data" (Cameron & Trivedi, 1998) and this seems appropriate considering the rare nature of the presence of disability, especially profound and severe levels of severity. It certainly will be informative to compare area level models A and B to see what difference an underlying Poisson makes.

3. Being a multivariate model, the estimated parameter vector $\boldsymbol{\beta}$ will provide Rp separate values, one for each response variable and covariate, all output from the one modelling process. This would be preferable to modelling each component of $\hat{\boldsymbol{\theta}}_\mathbf{i}$ as separate univariate models. If in applying Model A we find that some elements of $\boldsymbol{\beta}$ are not significant then the corresponding elements of $\mathbf{x}_\mathbf{ir}^\mathbf{T}$ can be made zero.

4. This model takes account of the known variance/covariance matrix $\boldsymbol{\psi}_\mathbf{i}$ (between the components of $\hat{\boldsymbol{\theta}}_\mathbf{i}$ for a given SLA i), thereby reducing model error.

5. Synthetic estimation models are a special case of this model after setting $\boldsymbol{\Sigma}_\mathbf{v} = \mathbf{0}$. Such models do not account for variation between SLA's other than that governed by the auxiliary variables. (Rao, 2003)

6. In our case, the $\hat{\boldsymbol{\theta}}_\mathbf{i}$ are post-stratified estimators which are a non-linear function of the component sample estimates. With the small sample size in some SLA's, the bias in the $\hat{\boldsymbol{\theta}}_\mathbf{i}$ may become more appreciable, thereby making equation (1) above invalid as the $\mathbf{e_i}$

account for sampling error only. Rao (2003) suggests replacing (1) with counterparts for the component estimates, but the resulting mismatch between (1) and (2) requires Hierarchical Bayes methods. An alternative approach would be to introduce another error term for the bias, and estimate this using say, Taylor series methods.

7. This model is not so readily amenable for taking account of clustering effects in the sample design. The two-fold nested error regression model (Sect 5.5.3 of Rao, 2003) would be more suited for this.

8. Does not accommodate correlated sampling errors between SLA's, otherwise known as a spatial error structure.

9. This multivariate model can be fitted using the SAS Procedure MIXED (See Verbeke and Molenberghs, 1997).

10. Unlike the unit level models discussed below, we don't have to go through the process of simulating unit level auxiliary records for each person selected in SDAC. Also, unlike the unit level generalised linear models discussed below, it is simpler to predict SLA estimates of disability from the model as well as interpret parameters and output.

B. Multivariate Fay-Herriot Area Level Model with Poisson Log Transform

Summary of Extensions of Model B from the Basic Fay-Herriot Model	
1	Model B is multivariate in $\boldsymbol{\theta}_{\mathbf{i}}$ with a covariance structure for the sampling error and random effects terms. By taking account of the correlations between variables, Model B should therefore give efficiency gains over fitting separate univariate models for each variable.
2	Model B uses a underlying Poisson distribution with log transform to take account of the fact that we are dealing with rare count data.

27. We would prefer to employ a Poisson regression model because we are dealing with count data and rare counts at that.

28. The same notation applies as in model A above.

29. Let the Poisson conditional density function for each $\hat{\theta}_{ir}$ be

$$f\left(\hat{\theta}_{ir} \mid \mathbf{Z_i}\right) = \frac{e^{-\mu_{ir}} \mu_{ir}^{\hat{\theta}_{ir}}}{\hat{\theta}_{ir}!}, \quad \hat{\theta}_{ir} = 0,1,2,....$$

$$i = 1,....,m \quad \mathrm{r} = 1,....,R$$

with mean parameters μ_{ir} obeying

$$\boldsymbol{\gamma}_{\mathbf{i}} = \mathbf{Z_i}\boldsymbol{\beta} + \mathbf{v_i} + \mathbf{e_i},$$

where

- $\boldsymbol{\gamma}_{\mathbf{i}} = \left(\log\left(\mu_{i1}\right),.....,\log\left(\mu_{iR}\right)\right)^T$
- $\mathbf{e_i} = \left(e_{i1,.....},e_{iR}\right)^T$ are the sampling errors, distributed as

independent r-variate normal

$$N_r\left(\mathbf{0}, \boldsymbol{\Psi}_{\mathbf{i}}\right),$$

- $\mathbf{v}_{\mathbf{i}}$ are the SLA level random effects, distributed independently as $N_r\left(\mathbf{0}, \boldsymbol{\Sigma}_{\mathbf{v}}\right)$.

Comments on Model B

1. Model B with its Poisson link function may be fitted using the SAS Procedure NLMIXED. ***However PROC NLMIXED cannot handle the multivariate aspect of the model***, in which case it may be necessary to fit univariate models one disability type (by severity if required) at a time.

C. Fay-Herriot Combined Unit/Area Level GLMM with Logistic Transform

Summary of Extensions of Model C from the Basic Fay-Herriot Model	
1	In Model C, person level response data is modelled against area level auxiliary data. (person level auxiliary data is not available)
2	A logistic transform with an underlying Bernoulli is used to take account of the binary nature of the response variable at person level.

30. This model makes use of the auxiliary and response variable data at the finest level of detail available: SDAC person level disability data for the response variable and SLA level count data for auxiliary variables. Thus every person in a given SLA cell will have, for example, the same Disability Support Pension (DSP) explanatory variable count value. If we receive the requested Commonwealth State/Territory Disability Agreement (CSTDA) data, then for this auxiliary variable, each selected person in SDAC will be associated with the corresponding SLA by disability type by severity by age by sex cell from the CSTDA data.

31. In models A and B, the dependent variable was a rare count for the SLA, and hence the Poisson distribution was appropriate. The dependent variable in model C is now at the person level. In SDAC a person is assigned to only one disability category so each disability category response variable can only be 0 or 1.

Notation for Model C		
Descriptor	**Index**	**Range**
Small Areas: SLA	i	i = 1,.....,m

Persons	j	j = 1,.....,n_i
Disability Categories	r	r = 1,.....,R
Auxiliary Variables		$x_{i1},.....,x_{ip}$

32. Our goal is to model the disability status y_{ij} of person j, $j = 1,....,n_i$ within SLA i (our small area) $i = 1,....,m$ and then use this model to predict disability statuses for non-sampled units, thereby producing estimates of disability $\hat{\theta}_i$, $i = 1,....,m$. We assume the y_{ij}'s to be independent Bernoulli(p_{ij}) variables with conditional probability density function:

$$f\left(y_{ij} = 1 \mid p_{ij}\right) = p_{ij}$$
$$f\left(y_{ij} = 0 \mid p_{ij}\right) = 1 - p_{ij}$$

33. Then the logit of the Bernoulli parameters p_{ij} can be modelled thus:

$$\theta_{ij} = \text{logit}\left(p_{ij}\right) = x_i^T \beta + v_i + u_{ij}$$
$$i = 1,.....,m \quad j = 1,.....,n_i$$

where

- the residual errors u_{ij} are distributed $u_{ij} \overset{iid}{\sim} N\left(0, \sigma_u^2\right)$
- the SLA level random effects v_i are distributed $v_i \overset{iid}{\sim} N\left(0, \sigma_v^2\right)$ independently of the u_{ij}.

Comments on Model C

1. Note model C involves area level explanatory variables x_i^T, only (although these values would in practice be replicated for each

person in a given SLA), SLA level random effects v_i, and person level error terms u_{ij}.

2. Model C above has been shown in its univariate form. We would like to derive a corresponding multivariate model but we're not certain whether such as model is identifiable.

D. Fay-Herriot Unit Level GLMM with Logistic Transform

Summary of Extensions of Model D from the Basic Fay-Herriot Model	
1	Model D is fully person level for both the response and auxiliary variables. (person level auxiliary data is simulated in order to fit this model)
2	A logistic transform with an underlying Bernoulli is used to take account of the binary nature of the response variable at person level.

34. Due to departmental privacy considerations, it has not been possible to acquire any unit level auxiliary data. And so the only way a fully unit level model approach could be pursued is by simulating person level observations (using the Poisson density function conditional on each given cell count total). This is not our preferred option for producing SAE's, however if time permits during the empirical study, numerous simulations of person level auxiliary data could be used to compare the performance of the unit level model against that of the area level model. This may provide useful information on what reduction in efficiency might result from going with an area level model.

35. To simulate the population auxiliary data from the count data provided, we propose to fit a Poisson distribution conditional on the known counts and then convert the Poisson parameter to a proportion. This proportion will then be used as the distribution parameter for a Bernoulli process to generate person level auxiliary variable values. This exercise would have to be done separately for each of the CSTDA, DSP and HACC data sources.

36. Once this simulation has been carried out, Model D can then be applied. Model D is similar to that of Model C except that the simulated explanatory variables are now at person level:

$$\theta_{ij} = \text{logit}\left(p_{\text{ij}}\right) = x_{ij}^{T}\beta + v_i + u_{ij}$$

The notation and description of terms in this model are the same as those for Model C.

Generating SAEs from Unit Level Models C and D

37. Small area estimates need to be formed from the models that use unit level response variables. This can be done for each disability category by summing unweighted responses from the sample s_i and then adding to that the sum of the $\tilde{p}_{\text{ij}}$ across the non-sampled component of the population s_i^c. The

$\tilde{p}_{ij}$ are predicted from the respective models by estimating β and also generating a realisation of v_i from its underlying distribution. We then have:

$$\hat{y}_i^r = \sum_{j \in s_i} y_{ij}^r + \sum_{j \in s_i^c} \tilde{p}_{ij}^r$$

where

y_{ij}^r = the sample response for the r'th disability category from the j'th person in the i'th SLA

$\tilde{p}_{ij}^r$ = predicted distribution parameters for the r'th disability category of the j'th person in the i'th SLA

$\hat{y}_i^r$ = modelled count estimate for the r'th disability category in the i'th SLA

s_i = the sample of persons in the i'th SLA

s_i^c = the sample complement of the i'th SLA

38. An important issue is how do we ensure that the sum of the modelled count estimates across disability categories, $\sum_r \hat{y}_i^r$ agrees with the population benchmark for the i'th SLA (assuming we've included no disability as a category).

References

Cameron, A.C. and Trivedi, P.K. (1998), *Regression Analysis of Count Data*, Cambridge: Cambridge University Press

Greene, W. H. (2000), *Econometric Analysis,* NJ: Prentice-Hall

Ghosh, M., Natarajan, K., Stroud, T.W.F. and Carlin, B.P. (1998), Generalised Linear Models for Small-Area Estimation, *Journal of the American Statistical Association*, **93**, 273-282

Lee, S.R. and Eltinge, J.L. (1999), Confidence Bounds for Survey-Weighted Quantile Plots and Offset-Function Plots, *Sankhya, Series B*, 61, 106-132

Molenberghs, G. and Lesaffre, E. (1999), Marginal Modelling of Multivariate Categorical Data, *Statistics in Medicine*, **18**, 2237-2255

Pfeffermann, D. and Bleuer, S.R. (1993), Robust Joint Modelling of Labour Force Series of Small Areas, *Survey Methodology*, **19**, 149-163

Pfeffermann, D. and Sverchkov, M. (1999), Parametric and Semi-Parametric Estimation of Regression Models Fitted to Survey Data, *Sankhya*, **61**, 166-186

Pfeffermann, D. (2002), Small Area Estimation - New Developments and Directions, *International Statistical Review*, **70**, 125-143

Rao, J.N.K. (2003), *Small Area Estimation*, Hoboken, NJ: Wiley

Verbeke G. and Molenberghs, G. (Eds) (1997), *Linear Mixed models in Practice: A SAS Oriented Approach*, New York: Springer,

Wolfinger, R.D., Fitting Nonlinear Mixed Models with the New NLMIXED Procedure

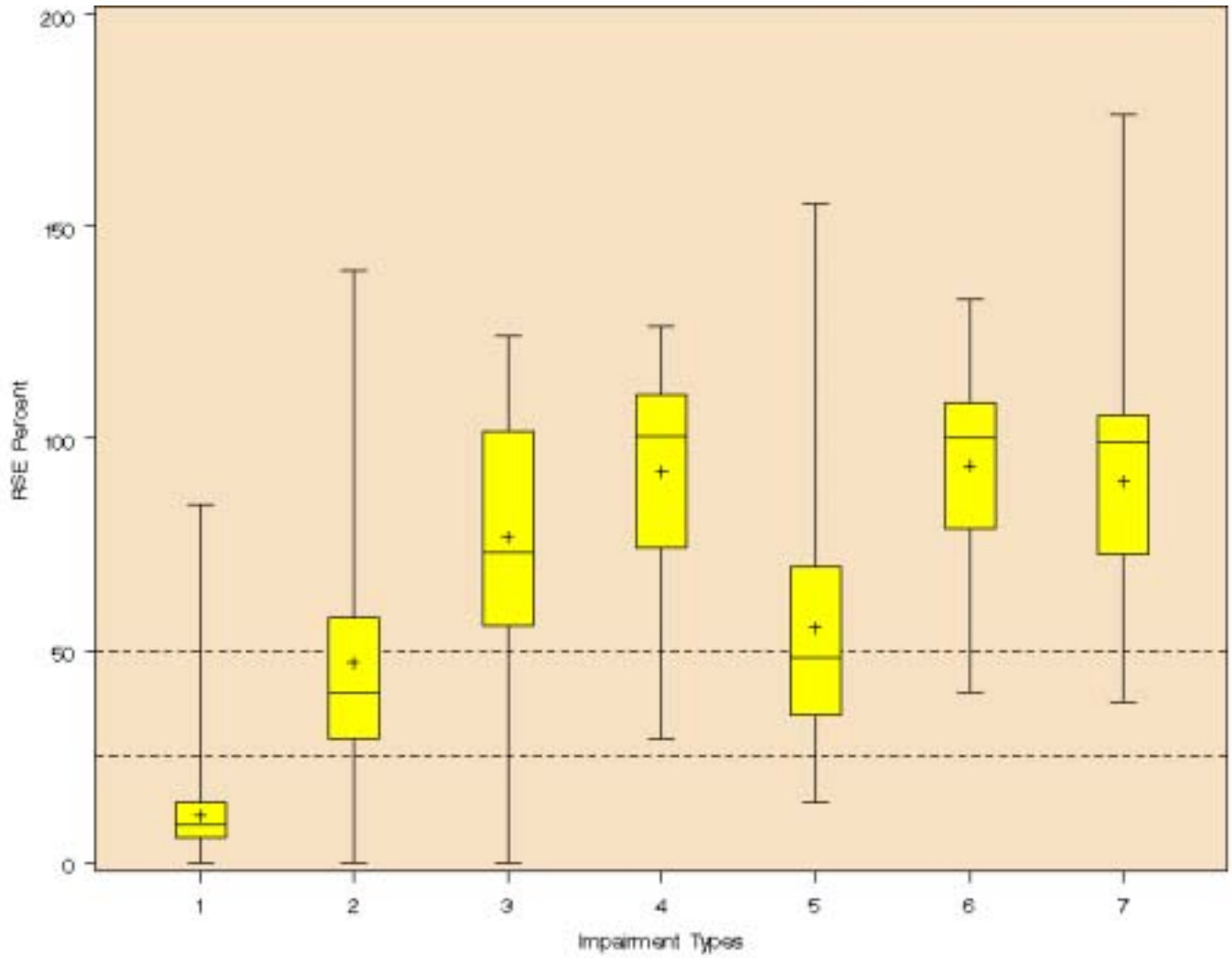
Appendix 1

SDAC 1998 Percentage Estimate of Persons by Disability Type

Disability Type	Percentage Estimate
1 = No disability	80.4%
2 = Any Impairment	19.6%
3 = Sensory	3.7%
4 = Intellectual	1.3%
5 = Physical	12.3%
6 = Psychological	0.8%
7 = Head Injury / Brain Damage	1.5%

Appendix 2

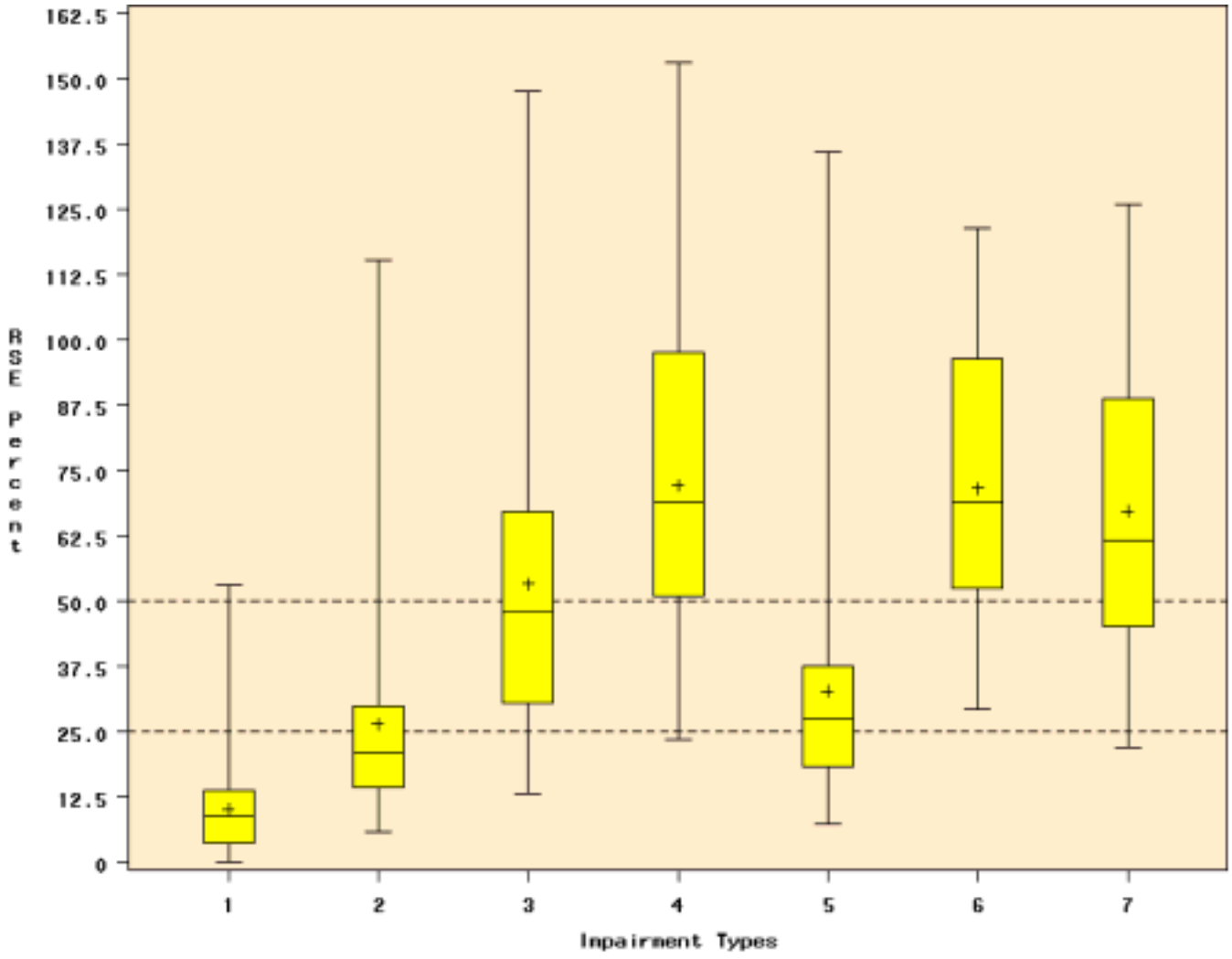
Boxplots of Jackknife RSE's for Statistical Local Areas (SLA)



Impairment Types	1 = No impairment	2 = Any impairment	3 = Sensory	4 = Intellectual
	5 = Physical	6 = Psychiatric	7 = Head injury/ brain injury	

Appendix 3

Boxplots of Jackknife RSE's for Statistical Sub-Divisions (SSD)



Impairment Types	1 = No impairment	2 = Any impairment	3 = Sensory	4 = Intellectual
	5 = Physical	6 = Psychiatric	7 = Head injury/ brain injury	

Appendix 4

Auxiliary Variable Data Sources

1. *Commonwealth State/Territory Disability Agreement - Minimum Dataset - Consumer File (CSTDA MDS)*

   A combined dataset of persons throughout Australia receiving home assistance or support from service providers funded (fully or partly) by either Commonwealth, state or territory governments. This data, if released to the ABS, will be in the form of a table of person counts crossed with disability type and degree of severity at the SLA by age by sex level.

   We have asked for two data snapshots, one at 2003 and the other at 1999, which is the closest timepoint to 1998 where data is of suitable quality.

   The data is kept by the Australian Institute of Health and Welfare but responsibility for the data lies with disability administrators in each state and territory jurisdiction.

2. Disability Support Pension (DSP)

   Two datasets compiled from Centrelink records of persons in Australia receiving the DSP, one at 1999 and the other at 2003. The data we have been provided with consists of tables of counts of persons with a given disability type in each SLA. No split by age and sex or degree of severity was provided. Any cells with a count of less than 20 have been censored. A method of dealing with censored cells, prior to modelling, is suggested at Appendix Z.

3. Home and Community Care (HACC)

   A dataset of persons receiving support under the HACC program which provides basic support services to enable frail older people and younger people with disabilities to remain living in their home. Most of the people on the HACC dataset are in the over 65 age group which thereby covers the deficiency of this age group on the DSP and CSTDA datasets. Importantly the HACC dataset is a better source of disability data for the more remote areas of Australia. We are still in the process of obtaining this data.

4. Remoteness

   The ABS Australian Standard Geographic Classification (ASGC) contains a classification for remoteness of any locality throughout Australia. Every Collectors District (CD) in Australia is assigned an index value that identifies the level of remoteness from major or minor service centres. The remoteness index, which ranges continuously from 0 to 15, is based on ARIA Plus scores for each 1 kilometre square grid in Australia. These scores result from calculating the weighted average of road distances from the given 1 kilometre square grid centroid to the nearest service centre in each of five population size classes.

There is a belief that instances of more serious levels of disability in the non-indigenous population are lower in more remote areas of Australia, generally because people with such disabilities are more likely to move to regional centres where appropriate care and support is more readily available. Remoteness therefore may be a good predictor of the level of disability in an SLA.

5. Total occupancy of residential care special dwellings (SD's) in the SLA

   The number of beds/rooms in cared accommodation in the SLA may be an indicator for the number of persons with a disability in the SLA. A good source of this data is the household surveys SD framework, which is a comprehensive list updated from sources independent of the sample.

6. Socio-economic Indexes for Areas (SEIFA)

   SEIFA indexes are derived from census variables and are available at the SLA level. There are indexes for a range of socio-economic indicators which cover

   - economic advantage and disadvantage
   - education and occupation
   - economic resource

   Some or all of these variables will be included to determine whether they are appropriate predictors of disability levels at the small area level. For example it's possible that levels of disability in the under 65 population are higher in areas where employment in hazardous occupations is more common.

7. Population size

   This provides a measure of the population size as an explanatory variable. This is not expected to add much more predictive power in addition to the other auxiliary variables but is included here for comparison purposes with previous synthetic estimation models of disability produced by the ABS as a consultancy service. In these synthetic estimation models, demographic benchmark counts projected from the previous census, were the main (if only) auxiliary variables used. Population size may also be found to be collinear with remoteness or other variables.

   Population size will also be very useful for forming estimates of disability in the case of unit level models discussed below.

   Demography Section of the ABS does not produce estimated resident populations (ERP's) at the SLA level. However it's possible to derive SLA population estimates either using Labour Force Survey (LFS) estimates directly (which carry their own sampling error) or by prorating ERP's at a broader level by the LFS population estimate. I'd prefer using either depending on size of SLA. The LFS estimates for larger SLA's are more likely to be well correlated with the LFS population estimate for the broader area, hence giving a potentially lower RSE for the estimate of ratio. Smaller SLA's are more likely to have LFS population estimates less correlated with corresponding estimate at the broader area.