Methodology Advisory Committee
June 2001

Part A
Cost and Variance Modelling For the 2001 Redesign of the Monthly Population Survey

Katrina Hicks

Part B
Optimal Design For the 2001 Redesign of the Monthly Population Survey

Daniel Elazar

Authors: Katrina Hicks and Daniel Elazar
c/ ABS, PO Box 10, Belconnen, ACT 2615
email: katrina.hicks@abs.gov.au, ph: 6252 7629
email: daniel.elazar@abs.gov.au, ph: 6252 6962

Executive Summary

1. Variance models for the redesign of the Monthly Population Survey (MPS) have been calculated at the area type level for both employed and unemployed persons estimators. These models describe the relationship between the sample design of the MPS and the variance of each estimator under that design for a given area type. Each model is determined by calculating the variances for fifty different sample design options centred around the 1996 design, and using the relationship between the parameters from those fifty designs and the design variances, to fit a model using ordinary least squares regression. The model was fitted to variances across all stages of selection rather than being aggregated from stage-wise models. As it is proposed to select sample for the 2001 redesign using independent selections in each stratum, models have been fitted separately for stratum independent selections. However as variance constraint values in optimisation need to be based on the current selection method, variance models have also been fitted for samples selected independently at the state level.

2. The regression analysis produced strong model fits with most parameter estimates significantly different from zero at the 0.05 significance level and high R^2 values. While every effort has been made to reflect the MPS procedures and methodologies as closely as possible in producing the variance models, there are still a couple of sources which may result in differences between variances predicted by our model and actual MPS variances. Firstly the 1996 Census labour force variance structure may not necessarily reflect that of the current MPS. Secondly there are difficulties in blocking CDs from census data as they would be in practice according to PSO procedures. In addition current variance models do not use post-stratified estimation in producing individual sample design estimates.

3. Cost models for the redesign of the MPS have also been calculated at the area type level. Each model describes the relationship between the sample design of the MPS and the cost of enumerating a sample under that design on the given area type. The cost models are determined by using detailed pay information for interviewers for a month of the MPS, to break the total enumeration costs down into components on each area type. The data used for this purpose more comprehensively covers MPS workloads and area types than in previous redesigns and gives detailed information on the costs associated with telephone interviewing for the first time. The cost parameter estimates based on this data are very robust with relative standard errors generally less than 10%. Problems with the quality of the cost data will nevertheless impact on the quality of these cost models for describing the enumeration costs for the MPS. Some of these problems include the quality of optical character recognition scanning of interviewer pay forms, the fact that only one month of data has been used to reflect the entire design period, and the difficulty of breaking costs down accurately into cost components.

4. As in previous redesigns, the aim of the current redesign is to produce a sample design that meets certain specific variance requirements and minimises cost. The approach to optimisation has changed from redesign to redesign as the method for controlling state variance relativities has changed. Prior to 1996, the state skips determining sample size were set outside the optimisation process and were based on making adjustments to a set of historically specified acceptable state sample sizes. In 1996 the state skips were determined through the optimisation process by including an explicit constraint to preserve adjusted historic state sample size relativities. The approach being considered for the current redesign is to incorporate explicit constraints on the relativities between state relative variances directly. This would have the advantage of allowing for direct control over state relative variances rather than determining state allocations largely through adjustments to state skips. The envisaged approach is made feasible by the development of models at the state by area type level as part of the cost and variance modelling investigations. The main potential problems associated with using this alternative over the 1996 method are firstly the impact of using state by area type models instead of the more robust area type level models and secondly some current uncertainty as to whether numerical methods for solving the optimisation problem are globally convergent at a sufficiently fast rate.

Discussion Points For MAC

5. The main issues of concern for MAC are as follows:

- Is the variance modelling methodology appropriate, particularly with respect to modelling the three stage process directly?
- Is collapsing area types into broader groups before modelling likely to produce significant gains in variance model quality?
- Is a linear cost model still the most appropriate approach to modelling enumeration costs?
- Is the proposed alternative optimisation approach for the 2001 redesign feasible?

Part A
Cost and Variance Modelling For the 2001 Redesign Of The Monthly Population Survey

1. Introduction

6. Initial cost and variance models have been determined for the 2001 redesign of the Monthly Population Survey. While these models may undergo further finessing, time permitting, they are of sufficient quality to provide information about the expected cost and variance structure under different sample design options. This paper discusses the methodology behind the cost and variance modelling. Section 2 discusses the background of the sample design for the Monthly Population Survey, while Sections 3 and 4 deal with the methodology for the variance and cost modelling respectively, and the quality of the models produced.

2. Background

7. The Monthly Population Survey (MPS) has a multistage sample design with selections undertaken in a number of stages. In most cases, there are three stages of selection, with the first stage involving the selection of a number of census Collector's Districts (CDs) followed by the second stage selection of blocks, generally only one, from within each selected CD, and the third stage selection of one cluster of dwellings from within each selected block. While there is often an extra stage of selection before the selection of CDs in more sparsely populated geographic areas, this situation is not explicitly accounted for in the modelling process.

8. The sample design for the MPS is specified by state skips and cluster sizes. The state skips determine the overall number of clusters to select from each state through the three stage selection process. For example, the state skip for NSW under the 1996 design is 300, which means that 1 in 300 clusters are selected from within NSW for the life of the 1996 design. The design cluster sizes are used primarily to assign size measures to CDs and blocks, in order to determine which particular CDs and blocks are selected in the first two stages on the way to selecting the specified number of clusters.

9. The design cluster size for a particular CD and the population census counts of the number of dwellings in the population in that CD, are used to calculate the number of clusters in the CD. This number is used as a size measure to select CDs within each state systematically by probability proportional to size (pps) without replacement. Each selected CD is then split up into blocks on the ground, and an exact, up to date, count of dwellings in each block is undertaken. The dwelling count and the design cluster size are then used to determine the number of clusters in each block within the CD. A number of blocks, usually just one, are selected from the blocks in the CD systematically pps without replacement. Once a block has been selected, a cluster of more or less equally dispersed dwellings is selected from within the block systematically, by running a skip through the block equal to the number of clusters in that block.

10. The CDs throughout Australia are split into a number of groups called area types on the basis of geographical region (e.g. state capital (met) vs rest of state (ex-met)) and population density. The very densely populated CDs in inner city Melbourne and Sydney are grouped together to form one area type, as are the CDs from all states with an average of less than 0.06 dwellings per square kilometre. On the other hand, Hobart and Darwin each define a single area type by themselves. Full definitions of the sixteen possible area types are given in Appendix A.

11. As cost and variance are often affected more by population density over an area than, say, state, the design parameters affecting cost and variance, the number of clusters selected and the cluster size within each selected CD, are set at the area type level. Let m_i denote the number of selected clusters, or equivalently, blocks, in area type i and let q_i denote the cluster size in area type i. The Australia level sample design incorporating the selection of m_i clusters and a cluster size of q_i in area type i will be referred to as an ($\underline{m}$,$\underline{q}$) sample design and the vectors $\underline{m} = (m_1, m_2, ..., m_{16})$ and $\underline{q} = (q_1, q_2, ..., q_{16})$ will be referred to as the design parameters.

12. The purpose of cost and variance modelling is to determine explicit area type level cost and variance functions that describe the relationship between the choice of design parameters m_i and q_i and the cost and accuracy associated with producing estimates for that area type under such a sample design. Once such functions have been fully specified, they can be used to algebraically determine the best ($\underline{m}$,$\underline{q}$) sample design to implement in order to meet particular cost and variance requirements for the MPS.

13. The MPS has the same set sample design for a period of approximately five years. The redesign of the sample is timed to coincide with the five yearly population census, so that dwelling counts obtained during the enumeration of the census can be used to calculate the size measure used in the pps selection of CDs. The design implemented at the beginning of the five year period is chosen so as to be optimal, or nearly optimal, for meeting the desired cost and variance requirements. The design becomes less optimal for meeting these requirements over time, however, as population characteristics, including variance structure, response rates, and cost relativities change. Consequently, during the sample redesign, more current information about cost and variance structure is used to calculate relevant cost and variance models and then update the sample design of the MPS.

14. The last redesign of the MPS sample was undertaken in 1996 and implemented over an eight month period from September 1997 to April 1998. Information is now being used to undertake the 2001 redesign of the MPS, with the resulting new sample design due to be implemented for the MPS from September 2002 to April 2003, at the time of writing.

3. Variance Modelling

3.1 2001 Variance Modelling Methodology

3.1.1 The Form of the Variance Model

15. Consider a sampling scheme in which m first stage units are selected pps with replacement and q second stage units are selected as a simple random sample without replacement from each selected first stage unit. The equation for the population variance of an estimate of population total, such as total employed or unemployed persons, is given by

$$Var\left(\hat{Y}\right) = \frac{1}{m}\sum_{j=1}^{M} P_j\left(\frac{Y_j}{P_j} - Y\right)^2 + \sum_{j=1}^{M}\frac{1}{mP_j}\frac{N_j^2}{q}\left(1 - \frac{q}{N_j}\right)S_j^2$$

where:

M is the total number of first stage units in the population

N_j is the total number of second stage units in the jth first stage unit

P_j is the pps probability of selecting the jth first stage unit

Y_j is the total number of employed persons (unemployed persons) in the jth first stage unit

S_j^2 is the variability in the employment (unemployment) characteristic in the jth first stage unit

Y is the total number of employed persons (unemployed persons) in the population.

This formula can be rearranged to have the form

$$Var\left(\hat{Y}\right) = \frac{k_1}{m} + \frac{k_2}{mq}$$

where

$$k_1 = \sum_{j=1}^{M} P_j\left(\frac{Y_j}{P_j} - Y\right)^2 + \sum_{j=1}^{M}\frac{N_j}{P_j}S_j^2$$

$$k_2 = \sum_{j=1}^{M}\frac{N_j^2}{P_j}S_j^2$$

are in terms of population values that are independent of the size of m and q. These values measure the variability between first stage units and within first stage units respectively.

16. The MPS selection process can be regarded as a two stage process in which the first stage is to select blocks pps without replacement, and the second stage is to select a cluster of dwellings systematically from within each selected block. This selection method is very similar to that described above and therefore the area type variance for the MPS employment and unemployment estimators is thought to have a similar form, with an additional constant factor representing efficiency gains in using without replacement sampling at the first stage. The variance of the employment and unemployment estimators at the area type level is therefore thought to have the form:

$$Var\left(\widehat{Y}_i\right) = v_{0i} + \frac{v_{1i}}{m_i} + \frac{v_{2i}}{m_i q_i}$$

where m_i is the number of blocks, or equivalently the number of clusters, selected in area type i, and q_i is the cluster size in area type i. Here v_{0i}, v_{1i} and v_{2i} are unknown constants.

17. The purpose of the variance modelling is to use information on the variances associated with a discrete number of ($\underline{m}$,$\underline{q}$) design options, to estimate the values of v_{0i}, v_{1i} and v_{2i} for each labour force variable at the area type level. This is achieved by fitting a model of the form:

$$Var\left(\widehat{Y}_i\right) = v_{0i} + \frac{v_{1i}}{m_i} + \frac{v_{2i}}{m_i q_i} + \varepsilon_i \qquad (1)$$

to the discrete number of (m_i,q_i) values associated with the ($\underline{m}$,$\underline{q}$) design options using ordinary least squares regression.

18. As a result of the systematic nature of the MPS selection process, it is not possible to effectively use historic survey data to estimate the design variances. In particular, as the MPS sample consists of one cluster of systematically selected dwellings from each block, the survey data does not provide information about the variability between clusters within blocks. As a result, the variance models for the 2001 MPS redesign were determined using information from the 1996 population census. The variance models were determined by calculating the impact of changing sample design on the variances of the labour force variables derived from 1996 census data for the census population in scope of the main monthly population survey, the Labour Force Survey (LFS). Variance models were determined for two variables, the employed persons variable, and the unemployed persons variable, as defined by 1996 census data.

3.1.2 Calculating Individual Design Variances

19. The first step of the variance modelling process was to calculate the area type level variance for each variable for each of fifty design options, that is for fifty particular sets of ($\underline{m}$,$\underline{q}$) values. These fifty design options were centred around the optimal 1996 design, as the optimal design for 2001 is not expected to differ significantly from the 1996 design. The fifty options were generated by combining one of ten different sets of state skip options for selecting clusters with one of five different cluster size options. For all area types, each of the ten fixed sets of state skips was a multiple of the 1996 values, the multiples ranging from one half to two. For each area type, five cluster sizes ±1 and ±2 either side of the 1996 optimal value for that area type were considered. More information on the designs considered for each area type can be found in Appendix B.

20. The methodology adopted for calculating the population variance of a variable under a particular ($\underline{m}$,$\underline{q}$) design option, was to calculate the variance of the sampling distribution of estimates under that design option. Consequently, all possible samples of the in-scope census population under each given design had to be identified. To undertake this task, it was necessary to have all CDs broken down into blocks and all blocks split into clusters.

21. Splitting of CDs into blocks and blocks into clusters is an operation carried out by Population Survey Operations (PSO) officers specifically for those CDs and blocks selected in the MPS. Blocking, in particular, is performed out in the field and involves partitioning a selected CD into a number of bounded groups of dwellings, called blocks, of a similar size. The blocks of dwellings are usually defined by obvious geographic boundaries such as roads, rivers or fence lines, so that it is easy to identify in the field which block any existing or new dwelling belongs to. The size of the blocks depends on the area type to which the CD belongs and is chosen to reach a compromise between cost, which is minimised by having small blocks that require minimal travel to move around, and variance, which is minimised by having large blocks of well spread dwellings. In metropolitan area types, for example, blocks are generally constrained to be within 4 to 8 clusters, according to the design optimal cluster size.

22. As blocking and forming clusters only occurs for CDs and blocks selected in the MPS, information about blocks and clusters is essentially non-existent for the census CDs. It was therefore necessary to implement a procedure that could create realistic blocks and clusters for all census CDs in order to complete the selection of all possible samples under a given design option. Information about the distribution of number of clusters per block for blocks selected in the MPS in a given area type under the 1996 design was used to generate a probability distribution specifying the probability of achieving each block size, in terms of number of clusters, in that area type. For example, if 10% of selected blocks in a particular area type consisted of 4 clusters, according to the 1996 design cluster size, then the probability of a block being allocated a size of 4 clusters was 1 in 10. For a given design, this distribution was then used to allocate the CD dwellings into blocks. For example, if the first block randomly generated according to the distribution was to have 4 clusters and the cluster size under the design was q_i, then the first $4q_i$ dwellings in the CD according to census order were placed in block one. The procedure then continued until the dwellings of the CD were completely partitioned into blocks. Blocks were then split into clusters by running a skip through the block equal to the number of clusters assigned to that block.

23. Once this procedure was complete for a given design, the entire population of Australia had been split into clusters of dwellings, with the number of dwellings in each cluster depending on the design cluster size parameter for each area type. As the MPS selection process is also equivalent to performing systematic selections on the set of clusters in the population, all possible samples were identified for a design by identifying the set of all possible systematic samples of clusters under that design. Under previous designs of the MPS, samples have been selected independently only at the state level rather than at the stratum level. Hence the stratification has been ignored for selection purposes, with only one random start being chosen to specify the sample of clusters coming from the strata within one state. Selecting the state sample by selecting samples from each stratum independently is expected to produce estimates with lower variance when the strata consist of CDs with similar labour force characteristics. It is therefore intended to change the MPS selection methodology to incorporate independent stratum level selections and this selection methodology was therefore incorporated into the variance modelling. As a result, the area type level variance under a given design were calculated by calculating and summing stratum level variances.

24. The number of clusters to select from each stratum is determined by the design state skip option. If there are M_h clusters in stratum h of size q_h ($=q_i$ where stratum h lies in area type i) and the state skip for the state in which stratum h lies is k_h, then the number of clusters selected from the stratum under that design should be M_h/k_h. As M_h/k_h is not likely to be a whole number, an adjustment is made to ensure that the number of clusters to be selected from the stratum is in fact a positive integer. This is achieved by adjusting the cluster size on stratum h, so that the number of clusters in the stratum is a positive integral multiple of the state skip, ensuring that the number of clusters selected is exactly a whole number. This means that the adjusted cluster size for a design must be used to split the CDs in a stratum up into blocks and clusters. Consequently, the adjusted optimal cluster size is calculated for each stratum under each design early on in the variance modelling programs, and it is the adjusted cluster size which is used to partition CDs into blocks and clusters as described earlier. The detail of the adjustment can be found in Appendix B.

25. Once all the samples in each stratum of an area type had been identified under a particular design, the estimates associated with each sample could be calculated. As the number of samples identified across all designs was extremely large, the estimation method was simplified in order to reduce computing time. Instead of replicating the Labour Force Survey estimation methodology by using post-stratification to calculate each of the sample estimates, Horvitz-Thompson estimates of population total were calculated. These estimates were calculated by weighting the cluster totals by the state skip, which is the inverse of the selection probability associated with each cluster. That is, each stratum level estimate had the form

$$\hat{Y}_h = \sum_{j \in s} k_h y_{hj}$$

where y_{hj} is the cluster total for cluster j in sample s from stratum h, and k_h is the design option state skip for the state to which stratum h belongs. The stratum level variance for each of employed persons and unemployed persons under a particular design was then calculated by measuring the variability of the set of all possible sample estimates of employed persons and unemployed persons respectively. The area type level variance for each estimator under a particular design was therefore given by

$$\begin{aligned} Var\left(\hat{Y}_i\right) &= \sum_{h \in i} Var\left(\hat{Y}_h\right) \\ &= \sum_{h \in i} \frac{1}{N_h(s)} \sum_s \left(\hat{Y}_{h_s} - \frac{1}{N_h(s)} \sum_s \hat{Y}_{h_s} \right)^2 \end{aligned}$$

where:

$N_h(s)$ is the number of samples in stratum h

$\widehat{Y}_{h_s}$ is the estimate of total for stratum h based on sample s.

3.1.3 Fitting the Model

26. As fifty design options were considered, there was a set of fifty points for each variable consisting of design values m_i and q_i and the area type variance associated with that design option. The relationship suggested by these fifty points was then used to estimate the values of v_{0i}, v_{1i} and v_{2i} by fitting a regression model of the form (1) to the fifty points. The estimates for v_{0i}, v_{1i} and v_{2i} that are produced from this regression fit are in some sense the best choice of values for explaining the relationship between m_i, q_i and variance for the fifty design options considered.

3.2 Results of Fitting Variance Models

27. The values of the v_{0i}, v_{1i}, and v_{2i} regression parameter estimates calculated for employed persons and unemployed persons at the area type level can be found in Appendix C. As these estimates measure different degrees of variability over different area types, an attempt has been made to standardise the parameter estimates for comparison across area types below. The values in Tables 3.2.1 and 3.2.2 are calculated by dividing the variance model parameter estimates for a particular variable on a given area type by the variance for that variable on that area type under a particular simple random sample. That is, the estimates in the tables are given by:

$$v'_{kiL} = \frac{v_{kiL}}{\frac{N_i^{\,2}}{n_i} P_{iL}(1-P_{iL})}$$

where:

v_{kiL} is the kth regression parameter estimate (k=0,1,2) on area type i for the variable L (=Employed or Unemployed Persons)

N_i is the population number of persons in area type i

n_i is the number of persons selected in area type i under the 1996 optimal design

P_{iL} is the population proportion of people in area type i with characteristic L (Employed or Unemployed).

The ratio v'_{2i}/v'_{1i} (= v_{2i}/v_{1i}) is also given in Tables 3.2.1 and 3.2.2 to aid in comparison of the area type models. This ratio indicates the level of within block variation to between block variation described by the models across the different area types.

28. The values highlighted in bold correspond to regression parameter estimates that are not significantly different from zero at the 0.05 significance level. A table of RSEs on each regression parameter estimate can be found in Appendix D.

Table 3.2.1
Standardised Variance Model Parameter Estimates For Variable "Employed Persons" By Area Type

Area Type	v'_{0i}	v'_{1i}	v'_{2i}	v'_{2i}/v'_{1i}	Adj-R^2
1. Inner City Melbourne/Sydney	-0.3151	121.349	1,076.34	8.87	0.9408
2. Inner City	**-0.0821**	83.997	2,722.21	32.41	0.9883
3. Settled Area	-0.2405	1,133.198	24,569.11	21.68	0.9971
4. Outer Growth	-0.2306	1,112.216	14,902.70	13.40	0.9948
6. Met Rural	-0.3680	124.745	1,645.92	13.19	0.9888
7. Large Town	-0.5167	1,010.123	14,466.46	14.32	0.9894
8. Small Town	-0.4093	318.635	4,919.76	15.44	0.9952
9. Ex-met Rural SRA	-0.4883	269.884	4,369.09	16.19	0.9966
10. Urban sampled	-0.1755	176.083	3,333.10	18.93	0.9947
11. Rural sampled	-0.4232	391.110	5,738.57	14.67	0.9975
12. Sparse	-0.3713	113.542	879.94	7.75	0.9832
13. Indigenous	-2.4681	47.954	**104.43**	2.18	0.8517
14. Growth	**-0.4725**	**1.672**	106.75	63.86	0.8072
15. Hobart	**-0.1282**	103.483	1,720.27	16.62	0.8868
16. Darwin	**-0.3170**	**33.257**	927.41	27.89	0.8433

Table 3.2.2
Standardised Variance Model Parameter Estimates For Variable "Unemployed Persons" By Area Type

Area Type	v'_{0i}	v'_{1i}	v'_{2i}	v'_{2i}/v'_{1i}	Adj-R^2
1. Inner City Melbourne/Sydney	-0.0898	**7.606**	806.01	105.97	0.9801
2. Inner City	0.0362	16.275	1,498.88	92.10	0.9953
3. Settled Area	-0.0337	130.571	14,713.56	112.69	0.9984
4. Outer Growth	-0.0319	123.900	9,860.39	79.58	0.9977
6. Met Rural	**-0.0409**	9.243	913.74	98.85	0.9863
7. Large Town	-0.0419	165.724	8,920.25	53.83	0.9978
8. Small Town	**-0.0343**	33.422	3,427.99	102.57	0.9964
9. Ex-met Rural SRA	-0.1264	46.963	2,771.43	59.01	0.9947
10. Urban sampled	**-0.0138**	29.081	2,251.06	77.41	0.9964
11. Rural sampled	-0.1006	63.202	3,936.61	62.29	0.9971
12. Sparse	**-0.0772**	15.431	510.89	33.11	0.9844
13. Indigenous	**-0.7143**	54.920	**31.35**	0.57	0.9425
14. Growth	**-0.3461**	**0.958**	52.72	55.01	0.8266
15. Hobart	-0.1612	37.037	1,069.10	28.87	0.9644
16. Darwin	**0.0512**	**11.267**	397.15	35.25	0.8281

29. The area type level variance models for each of employed and unemployed persons are very good model fits, in the sense that they describe the relationship between (m_i,q_i) and variance very well for the fifty available data points. This can be seen by the size of the adjusted R^2 goodness of fit measure associated with each regression model. In particular, the area type level models for area types 1 to 12 are particularly strong, adjusted $R^2>0.9$, with the remaining area types still exhibit good model fits, adjusted $R^2>0.8$. As the variances for different area types have been modelled over wide and varying ranges of values, potentially affecting the usefulness of the adjusted R^2 as a measure of goodness of fit, a second measure of robustness has been calculated in Appendix E. This measure compares the overall contribution of the model residuals over the range of variances used to fit each area type model. The values of this ratio are under 5% for most area types and around 10% for the worst area types, area types 13, 14 and 16. These small values suggest that the strength of the adjusted R^2 is not simply explained by having fit the models over a large range of values. All models are consequently regarded as being very reliable for the purpose of undertaking optimisation. This is particularly significant given that all data points were used to fit each model and no individual design (m_i,q_i) variance values were considered unusual enough to warrant removal from the modelling process as outliers.

30. The above tables indicate that a number of model parameter estimates were not statistically different from zero at the 0.05 significance level. While the proportion of non-significant estimates is quite small, it nevertheless suggests that it may be worthwhile investigating the impact of collapsing some of the area types down to a broader level before modelling, to see if this gives rise to more reliable parameter estimates.

3.3 Variance Models For Optimisation

31. The main purpose for undertaking the variance modelling is to specify functions that describe the variance structure of the 2002 MPS population for the purposes of determining the optimal design for meeting certain cost and variance requirements. One of the variance requirements is to achieve the same national level variance as that achieved when the 1996 design was first implemented.

3.3.1 Hybridising

32. The variance modelling produced models describing the variance structure for the employed persons estimator and the unemployed persons estimator. To ensure that the 2001 sample design will produce acceptable variances for both estimators, both variables need to be considered in the optimisation process. This is achieved by forming a new variance model for the optimisation process, which is a weighted linear combination of the variance models for employed persons and unemployed persons. In order to sensibly combine these variances, which are measures of variability of estimates of different sizes, it is necessary to first adjust them relative to the size of the employment and unemployment totals respectively, to put them on a standard basis. The new hybridised model is therefore a measure of relative variance and is produced by weighting the relative variance models for each of employed and unemployed persons.

33. The appropriate weighting of the employment and unemployment relative variances is a complex issue that would require extensive consultation and analysis to resolve. For reasons of timing, the decision was made to adopt the same weights for the 2001 redesign as were chosen for the 1996 redesign. The weights used were 0.9 for the employment relative variance and 0.1 for the unemployment relative variance. These values were chosen in 1996 as they increased the importance of the employment variable in specifying the optimal design, whilst producing an efficient design for both variables. The investigation of the weighting issue and more detailed reasons behind the choice of these parameters for the 1996 design can be found in Section 5.4 of Clark (1997).

34. The hybridised relative variance model parameter estimates at the area type level are given in Table 3.3.1 below.

Table 3.3.1
Hybridised Relative Variance Model Parameter Estimates For Each Area Type
x 10^{-10}

Area Type	v_{0i}	v_{1i}	v_{2i}
1. Inner City Melbourne/Sydney	-1,372	337,979	8,209,462
2. Inner City	-8	711,899	35,903,341
3. Settled Area	-10,364	46,805,171	1,865,841,594
4. Outer Growth	-6,369	29,222,503	804,090,939
6. Met Rural	-905	287,871	7,009,786
7. Large Town	-9,568	22,209,389	592,127,343
8. Small Town	-2,269	1,848,329	63,937,421
9. Ex-met Rural SRA	-1,879	898,215	26,877,992
10. Urban sampled	-905	1,059,825	37,075,067
11. Rural sampled	-2,361	1,933,268	53,569,343
12. Sparse	-245	67,765	872,147
13. Indigenous	-278	11,851	11,953
14. Growth	-6	17	1,020
15. Hobart	-221	82,015	1,831,623
16. Darwin	-28	9,473	296,810

3.3.2 Calculating Variance Constraints

35. The function describing the national level of hybridised relative variance is obtained by aggregating the above area type level models to the national level and it is constrained to meet a particular value as part of the optimisation process. This constraint is the relative variance achieved by weighting the variances for employed and unemployed persons that were achieved upon implementation of the 1996 design, by 0.9 and 0.1 respectively. The resulting design for 2001 then has the desirable property that is preserves, in some sense, the level of variance that was achieved on the employment and unemployment estimators when the 1996 design was first implemented. While sample estimates of variance produced from the monthly LFS are available, the true population variances on the estimators upon implementation of the 1996 design are actually unknown. Consequently, these values also need to be calculated in order to calculate the hybridised relative variance constraint and undertake the optimisation process.

36. The 2001 variance models are based on 1996 census data and therefore capture information on population variance structure for a period of time very close to the time at which the 1996 design was first implemented. These models do not reflect the selection and estimation methodology of the MPS under the 1996 design however, as the 1996 design incorporated independent state level selections and the variance models produced for the 2001 redesign incorporate independent stratum level selections. Consequently, it was necessary to produce a set of variance models from 1996 census data to describe the variance structure of the employed and unemployed persons estimators under the 1996 MPS selection methodology. The methodology for determining these models was identical to that already described for the 2001 variance modelling, except that the samples were selected at the state level, and sample estimates and design option variances were calculated at the state by area type level instead of stratum level. The national level variances for each estimator were then calculated by summing the area type level variances evaluated from these new models on the 1996 optimal design values of $\underline{m}$ and $\underline{g}$.

37. The area type variance models under state based selections are given in Appendix F. Appendix F also shows the resulting hybridised relative variance models and the relative variance constraint arising from these models.

3.4 Variance Model Quality

3.4.1 Improvements Over 1996 Variance Modelling Methodology

38. The methodology for the 2001 variance modelling has been significantly improved over the methodology for the 1996 modelling. The major differences are as follows:

- Models were fitted separately for each of the three stages of selection in 1996 and aggregated to form a model of the form (1). In 2001, a model of the form (1) was fitted directly to the total design variances. This approach is expected to give a more stable model, as the errors associated with modelling variances from different stages of selection might not be offsetting or independent.

- 1996 first and third stage models were fitted on the basis of between four and six data points after outliers were removed and second stage models on the basis of two data points. The 2001 models were fitted on fifty data points and no outliers were removed.

- 1996 models were only produced for area types 1 to 9, 15 and 16. In 2001 models have been produced for all area types and so all area types can have cluster sizes determined through the optimisation process.

- 2001 models have been calculated on the basis of independent stratum level selections instead of state level selections.

39. Given that the 2001 models have been based on a much larger number of data points and have been modelled in one single stage to reduce the potential contribution from compounding model error, the variance models calculated for the 2001 redesign are expected to be of a much higher quality than those produced for the 1996 redesign.

3.4.2 Comparing Census Variances and LFS Variance Estimates

40. In order to asses the quality of the census based models for explaining LFS variance structure, relative variances were compared between the 1996 census variance models and LFS estimates of relative variances for the month in which the 1996 design was first fully implemented, April 1998.

41. The variance models are calculated assuming a 100% response rate. That is, the area type variance of a variable under an ($\underline{m}$,$\underline{q}$) sample design will be the variance associated with the response of $m_i q_i$ dwellings selected in each area type under the given three stage design of m_i CDs, one block per CD and one cluster per block. In practice when such a design is implemented in the MPS, not all $m_i q_i$ dwellings will respond as some dwellings will be lost through demolition or removal, while some dwellings will not be contactable or will refuse to respond. As approximately 15% of selected dwellings fall into this "sample loss" category on average over Australia, it is important to ensure model and LFS relative variances are compared for comparable responding sample sizes.

42. Jackknife estimates of variance were calculated from the April 1998 LFS for the estimators of employed and unemployed persons. Census based estimates of variance were derived by evaluating the 2001 variance models at the 1996 optimal design values of m_i and the achieved values of $m_i q_i$ derived from the April 1998 LFS data. On comparison of the national level relative variances it was found that the employment relative variance coming from census data was 1.57 times the calculated relative variance for employment coming from the LFS. The census derived unemployment estimate of relative variance was almost exactly the same as that derived from the LFS at a ratio of 0.98 times the size.

43. While the relative variances are much closer than expected on the basis of historical comparisons between census model variances and LFS estimates, the size of difference still represents a reasonable degree of discrepancy, particularly for the employment variable. The fact that the variances for the LFS are only estimates, based on the available sample, of the true variance on the employed and unemployed persons estimators may be a minor contributing factor to this discrepancy. However, there are a number of methodological issues associated with the variance modelling that are likely to result in the calculated census population variances not fully reflecting the LFS variance structure for both the employment and unemployment estimators. These issues are discussed below, as are ways in which the variance modelling could be improved to take account of them.

3.4.3 Modelling Using LFS Data

44. The variance models produced are intended to describe the variance structure of the major labour force variables for the population on which the 2001 MPS design is to be implemented. The variance models have been derived using 1996 census data and using the census questions and derivation of the variables of employed persons and unemployed persons. The census population will be more than five years old before the 2001 MPS design is implemented. Furthermore, the census labour force variables are derived from different questions (both in number and phrasing) using a different collection methodology than the LFS. Therefore it is possible that the variance structure of the employed and unemployed persons variables from census data may not accurately reflect the variance structure of the employed and unemployed persons variables for the LFS over the new design period.

45. Census data is used so that population variances can be generated for different (m,q) design options. Another possible approach being considered for future investigation, is to use the LFS sample and calculate variances for (m,q) design options that can be drawn from within this sample by sub-sampling. These variances could then be used to generate a model that could be extended to the larger design options. This would have the advantage of reflecting the variance structure of the LFS labour force variables at a point in time far closer to the implementation of the new design. It would also better capture PSO field practices such as blocking (see below). The disadvantage of this method is that there will be a significant amount of error associated with sub-sampling from a fixed LFS sample. There may also be a significant amount of model error associated with extrapolating this model to larger design options.

3.4.4 Differences in Estimation And Selection Methodology

46. One major discrepancy between the selection and estimation methodologies for the variance modelling and the MPS is in the estimation methodology used to calculate sample estimates. Post-stratification is used to calculate sample estimates for the MPS. However, in order to simplify an already computer intensive and complex task, the series of sample estimates used to calculate the individual (m,q) design option variances for the variance modelling have been calculated using Horvitz-Thompson estimators. Using Horvitz-Thompson estimators in the variance modelling is expected to underestimate the size of the v_1 parameter, and therefore the total variance, associated with the post-stratified estimation process. This is expected to be more pronounced for the employment estimator because the gains associated with using post-stratification are greater and will therefore explain part of the large discrepancy between the variance model and LFS relative variances for employment. One area for further investigation is to attempt to quantify the impact of incorporating post-stratification into the variance modelling process and thereby produce variance models that better reflect the variance structure of the post-stratified MPS estimators.

47. A further problem results from the fact that the census CDs must be split into blocks and clusters before the samples required for the calculation of individual ($\underline{m}$,$\underline{q}$) design option variances can be identified. In the MPS, selected CDs are blocked by PSO interviewers according to a certain set of basic principals. These blocking patterns could not be replicated on the census CDs and instead an automated method was applied that generated a somewhat artificial partition of CDs into blocks for each design. As the block sizes were generated randomly for each design, the number of blocks allocated to each CD and the number of clusters allocated to each block will vary randomly from design to design. Furthermore, the number of dwellings allocated to a given block will depend on the cluster size for that design. The failure to reflect PSO blocking practices and the introduction of this random element to the blocking from one design to the next, may impact on the fitting of variance models by increasing the amount of "noise" in the data being modelled. Another issue to investigate with respect to the variance modelling is therefore the impact of fixing the blocking across as many designs as possible.

48. The final major discrepancy in the selection and estimation methodologies relates to the method of selection in the less densely populated area types. In the MPS, sampled and sparse area types undergo a fourth, PSU, stage of selection before the selection of CDs. This PSU stage of selection results in the selection of a group of CDs which are close geographically and from which CDs, blocks and dwellings are then selected in a three stage process. As including this additional stage of selection reduces the area from which the sample can be drawn, thereby increasing variance, the current three stage process used in the variance modelling would be expected to produce variance models which underestimate the level of variance in these area types. One final issue to investigate would therefore be whether a PSU stage of selection can be incorporated into the variance modelling process, so that sampled and sparse variance models can be produced which better reflect the variance structure of the MPS selection process on these area types.

4. Cost Modelling

49. The purpose of cost modelling is to describe the relationship between different sample designs and the cost of enumerating the MPS under those different designs. As with variance modelling, cost modelling is undertaken at the area type level because population density is expected to have a significant bearing on enumeration costs. Hence cost functions are determined that describe the relationship between the number of area type cluster, or equivalently, block selections, m_i, the area type cluster size, q_i and the cost of enumerating that area type in the MPS. These cost models are then used in conjunction with the variance models to specify the optimal values of m_i and q_i for each area type that will meet certain cost and variance requirements.

4.1 2001 Cost Modelling Methodology

4.1.1 The Form of the Cost Model

50. The cost of enumerating one month of the MPS under a particular design will depend on the number of blocks selected and the number of dwellings selected within each selected block. The model used for the 2001 redesign to describe the relationship between the number of blocks selected in area type i, m_i, the cluster size in area type i, q_i, and the cost of enumeration on area type i, c_i, is given by:

$$c_i = c_{0i} + c_{1i}m_i + c_{2i}m_i q_i \qquad (2)$$

where

c_{0i} is the overhead cost associated with enumerating area type i

c_{1i} is the cost of enumerating a block in area type i

c_{2i} is the cost of enumerating a dwelling within a block in area type i

4.1.2. Cost Data

51. The data used to undertake the 2001 cost modelling consisted of information on interviewer workloads supplied by PSO. A workload generally consists of the blocks to be enumerated by one interviewer and is a group of selected blocks that are geographically close together. Workloads are not constrained to fall within just one area type and in May 1999 for example, only 40% of workloads were contained within one area type.

52. The first form, called an AP10 form, lists every instance of every activity the interviewer carried out during their enumeration of the MPS for a given month, and the start time, duration and distance travelled, for each one of those instances. Examples of activities include travel to and from workloads, travel between blocks and interviewing. The AP10 form also indicates which block the instance of an activity took place in (if relevant) which can be used to identify the area type to which that instance consequently relates.

53. Using the AP10 form, interviewers calculate the total time spent on each AP10 activity over the enumeration period and the total distance travelled over the enumeration period for all those activities, and transfer the totals to a second from, called an AP10x form. These forms are then submitted to PSO and are used as a basis for calculating the interviewer's salaries. Interviewers are paid $17 an hour for the time components recorded on the form and, on average, 50 cents per kilometre for the distance component given on the form. The only activity for which interviewers are not paid exactly for the time they have recorded on the AP10x form is the activity of interviewing. Instead, interviewers are paid on the basis of the number of interviews they conduct using an average interview time calculated by PSO and called the "mean assessed time".

54. The AP10 forms contain valuable information about the breakdown of workload costs to the area type level that is not available from the AP10x forms. PSO therefore arranged for the interviewers to hand their AP10 forms in with their AP10x forms for one month, May 1999. The AP10 forms were then scanned in using Optical Character Recognition (OCR) so that the data could be analysed. As some interviewers were not required to submit their AP10 forms, and other forms were of insufficient quality to be of use in the analysis, AP10 data was only available for 80% of workloads.

4.1.3 Derivation of Cost Model Parameters

55. In order to produce cost models, it is necessary to split the costs associated with enumerating an area type down into the three components referred to in Equation (2). It was found that there was insufficient data at the activity level to produce reliable estimates of these components, and hence reliable cost models, for some of the smaller area types. Area types that were considered to have similar cost characteristics were therefore grouped together. Area types 4 and 14 (outer growth and growth) were grouped together, as were area types 6 and 9 (MET rural and Ex-met rural SRA), and area types 11, 12, and 13 (rural sampled, sparse and indigenous). The remaining area types had sufficient data to be modelled individually. These 11 area type groups are described more fully in Appendix G. The methodology for estimating each of the cost model parameters at the grouped area type level is described below.

Overhead Cost - c_0

56. The cost model parameter, c_0, is the overhead component of the cost model. It includes travelling costs (distance and time) that the interviewer incurs travelling to and from the workload and any advice/counselling the interviewer receives from their supervisor. While the overhead costs will clearly depend on the total number of workloads, or equivalently, number of interviewers, the number of workloads is not directly proportional to either the number of blocks or dwellings enumerated. As it is not possible to accommodate costs specific to the number of workloads in the model given in Equation (2), and as the number of interviewers employed is assumed to vary insignificantly over the range of $\underline{m}$ and $\underline{q}$ values likely to be considered in the optimisation, these cost are incorporated into the c_0 parameter of the model.

57. Workloads can, and often do, cover more than one area type. As AP10 data does not provide sufficient information to assign all the overhead costs for a workload to the blocks, and hence area types, covered by the workload, it is necessary to find a way to apportion the total workload overhead costs across area types. The total overhead cost for a workload is calculated by adding the cost associated with the time spent on the three overhead activities for the workload (at $17 an hour) to the distance related costs associated with those activities for the workload (at 50c per kilometre, on average). A regression model is then fitted to the workload overhead costs using the proportion of the blocks of the workload belonging to each area type as the explanatory variables. The model has the form:

$$c_{0w} = \alpha_1 p_{1w} + \alpha_2 p_{2w} + \alpha_3 p_{3w} + ... + \alpha_{11} p_{11w} + \varepsilon_w \quad (3)$$

where

α_j is the cost coefficient for broad area type group j (j=1,2,...,11)

p_{jw} is the proportion of blocks contained in area type group j for workload w

c_{0w} is the total overhead cost for workload w

ε_w is the residual term for workload w.

58. Given the regression estimates of the α_j, the term $\alpha_j p_{jw}$ can be thought of as the component of the total workload overhead cost for workload w that is attributable to area type group j. The total workload cost for area type group j could be calculated by summing the component costs, $\alpha_j p_{jw}$, for area type group j across the workloads covered by the AP10 data. However, as the AP10 data only covers 80% of workloads, this would not give the total overhead cost for each area type. Instead, proportions p'_{jw} are calculated for all workloads in Australia using information from the Labour Force Survey (LFS) data file for May 1999. The area type overhead cost component is then calculated by summing the component costs, $\alpha_j p'_{jw}$, for area type group j across all workloads in Australia using these proportions.

Per Block Cost - c_1

59. The cost model parameter, c_1, is the cost per block associated with enumerating the blocks in the given area type. It includes between block travel costs (distance and time) and checklisting costs. After some debate (see Section 4.3.3), it has also been decided to include the cost of blocklisting in c_1. Blocklisting is the procedure that occurs at the beginning of a new design period, when each CD in the newly selected sample is broken down into blocks and the dwellings in the blocks are listed (as distinct from checklisting, in which the interviewer walks around a newly selected block and updates the listing of the dwellings in the block so that clusters can be formed appropriately).

60. Estimates of blocklisting costs for the 1996 design are available at the area type level (see Appendix H). These are given as costs per CD over the five year life of the design, which is equivalent to costs per block as the MPS is essentially a one block per CD design, and includes the cost of blocklisting new CDs rotated into the sample. A monthly per block cost associated with blocklisting can therefore be calculated for each area type by dividing this cost through by 60 (5 years of 12 months). It only remains to calculate the per block costs associated with the remainder of the block related activities and sum the two together to obtain c_1.

61. The remaining block related activities, travelling between blocks and checklisting, only occur for blocks in which at least one dwelling is enumerated under face to face interviewing (a block completely enumerated under telephone interviewing in a given month is not visited in the field). As each instance of each of these activities has a block recorded against it on the AP10 form, the total AP10 block related costs for each activity can be calculated at the area type level. As the AP10 data only provides information on the costs associated with 80% of workloads in Australia, these costs are then scaled up using the AP10x summary data supplied for all workloads in Australia. That is, inflation factors are applied to the time cost components for each activity and the total distance cost component, so that the total Australian costs derived from the AP10 data are equal to those derived from the AP10x data. As AP10x data does not provide a breakdown of costs by area type, the same Australia wide inflation factor is used for all area types in each case. The scaled time and distance costs are then summed together on each area type and divided by the number of blocks in the area type to obtain the per block cost for these activities. The final per block cost is then obtained by adding on the per block cost for blocklisting.

Per Dwelling Cost - c_2

62. The cost model parameter, c_2, is the cost per dwelling associated with enumerating the dwellings in the given area type. It includes the cost of interviewing and between dwellings travel cost (time and distance) for dwellings that are in the same block (as travel between blocks has already been accounted for). The per dwelling cost is an average over all dwellings enumerated in the MPS, regardless of whether an interview is successfully carried out or not, as even the sample loss dwellings incur costs that are accounted for on the AP10 form in the form of between dwelling travel.

63. Each month, PSO pay interviewers a fee for interviewing on the basis of the number of interviews they conduct, ($17 per hour*number of interviews*mean assessed time). In May 1999 the mean assessed time was 23 minutes per interview, but this was found to be higher than average. For modelling purposes, the average interview time was therefore taken as the average mean assessed time over a period of two years from October 1997 to September 1999. This average value was 20 minutes per interview and therefore cost model interview costs were calculated assuming a cost of $5.67 per interview ($17/60*20). The May 1999 LFS data file was used to determine the total number of dwellings in which interviews were undertaken for each area type and the total area type interview cost was then calculated assuming a cost of $5.67 per interview.

64. The between dwellings travel cost (time and distance) was calculated in a similar manner as the block related costs discussed earlier. That is, the AP10x summary time and distance data was used to adjust the AP10 data to derive an estimate of total between dwellings travel cost, for all workloads, on each area type. This area type cost was then added to the interview cost and an average per dwelling cost was then calculated by dividing by the number of dwellings selected in May 1999. Note that this ensures that the average is over all dwellings selected, including those sample loss dwellings for which no interview is conducted.

Level of Modelling

65. The methodology described above was used to calculate cost model parameter estimates for each of the 11 area type groups described in Appendix G. As it was necessary to have cost models at the finer area type level for optimisation purposes, these group level area type parameter estimates were used to derive area type models at the finer level. As the fine level area types were collapsed (where necessary) with area types that had similar cost characteristics, it seems reasonable to assume that the per block and per dwelling costs are essentially the same for the area types grouped together. The grouped area type c_1 and c_2 values calculated above were therefore used for the c_1 and c_2 parameter estimates for all area types within the group. Hence area types 4 and 14, which were grouped together, will have the same c_1 and c_2 parameter estimates.

66. The c_0 component for a grouped area type represents the total overhead cost across all area types within the group and therefore has to be apportioned down to the fine area type level. This was achieved using the regression estimates, α_j, determined from the model in Equation (3). Proportions p'_{iw} were calculated for all sixteen fine level area types from the May 1999 LFS data. The total workload cost for area type i belonging to area type group j was then calculated by summing the component costs, $\alpha_j p'_{iw}$, for area type i across all workloads.

4.2 Results of Fitting Cost Models

67. The values of the c_{0i}, c_{1i}, and c_{2i} parameter estimates calculated for each area type are given below in Table 4.2.1.

Table 4.2.1
2001 Cost Model Parameter Estimates By Area Type

Area Type	c_{0i}	c_{1i}	c_{2i}
1. Inner City Melbourne/Sydney	1,300.73	6.24	4.36
2. Inner City	2,238.49	5.96	4.68
3. Settled Area	12,658.17	6.97	5.01
4. Outer Growth	11,149.31	6.38	5.10
6. Met Rural	2,101.47	15.14	4.56
7. Large Town	10,803.91	11.56	4.91
8. Small Town	5,538.30	17.86	4.43
9. Ex-met Rural SRA	5,048.50	15.14	4.56
10. Urban sampled	6,563.77	16.72	4.58
11. Rural sampled	7,430.07	22.08	4.42
12. Sparse	1,940.53	22.92	4.42
13. Indigenous	1,121.87	22.92	4.42
14. Growth	85.27	6.38	5.10
15. Hobart	1,888.18	9.02	4.89
16. Darwin	861.60	6.89	5.10

68. Estimates of Relative Standard Error (RSE) were calculated for the parameter estimates in Table 4.2.1 and can be found in Table I.1 of Appendix I. A description of the method for calculating these RSEs is also given in Appendix I. The c_{0i} and c_{2i} parameter estimates were found to be particularly stable, with RSEs generally smaller than 5%. In the former case, this was related to the quality of the fit of the regression model in Equation (3). This model had an adjusted R^2 of 0.7204 and all α_j estimates were significant at the 0.05 significance level.

69. Standard error estimates for c_1 indicated that c_1 was a little more volatile on each area type, RSEs generally between 5% and 10%. PSO confirmed that c_1 is likely to be volatile. If a workload is enumerated inefficiently or if the selected blocks within a workload are far apart, then block related costs can increase considerably as a result of increased between block travel costs. Furthermore, with the implementation of telephone interviewing, there can be considerable variation in the number of dwellings within a block requiring a personal visit, either for face to face interviewing or follow-up, and the distances between them. Hence interviewers can visit a large number of blocks that each require only a small number of interviews. Prior to the introduction of telephone interviewing, there was effectively a limit to the number of blocks that could be visited in a day, which had the effect of constraining between block travel costs. While this heuristic argument suggests the c_1 parameter may be more volatile than in previous designs because of the introduction of telephone interviewing, there is no concrete evidence available to either confirm or deny this to date. This is an area that could be investigated in the future.

4.3 Cost Model Quality

4.3.1 Improvements Over 1996 Cost Modelling

70. The major improvements in the cost modelling methodology relate to the availability of significantly more detailed cost data in 2001. While the 1996 cost modelling was based on a sample of AP10x data for 11% of workloads over a period of three months, the 2001 modelling was based on AP10 data for 80% of workloads and AP10x data for all workloads, for one month. The main differences in methodology were as follows:

- Cost models were only produced for area types 1 to 9 and 15 and 16 in 1996. In 2001 models have been produced for all area types and so all area types can have cluster sizes determined through the optimisation process.
- As there was no breakdown of data to the area type level, regression analysis was used to produce area type level estimates for all parameters in 1996. In 2001, AP10 data has been used as the basis for assigning c_1 and c_2 costs to the area type level.
- The 1996 cost modelling used pilot study data to estimate the costs associated with telephone interviewing. The 2001 cost modelling has made use of information on the established costs associated with implementing telephone interviewing in the MPS.

71. Given that the 2001 models have been based on data with detailed information about area type level costs and the costs associated with telephone interviewing, the 2001 cost models should better reflect the cost structure of the MPS at the area type level under a combined face to face/telephone interviewing methodology.

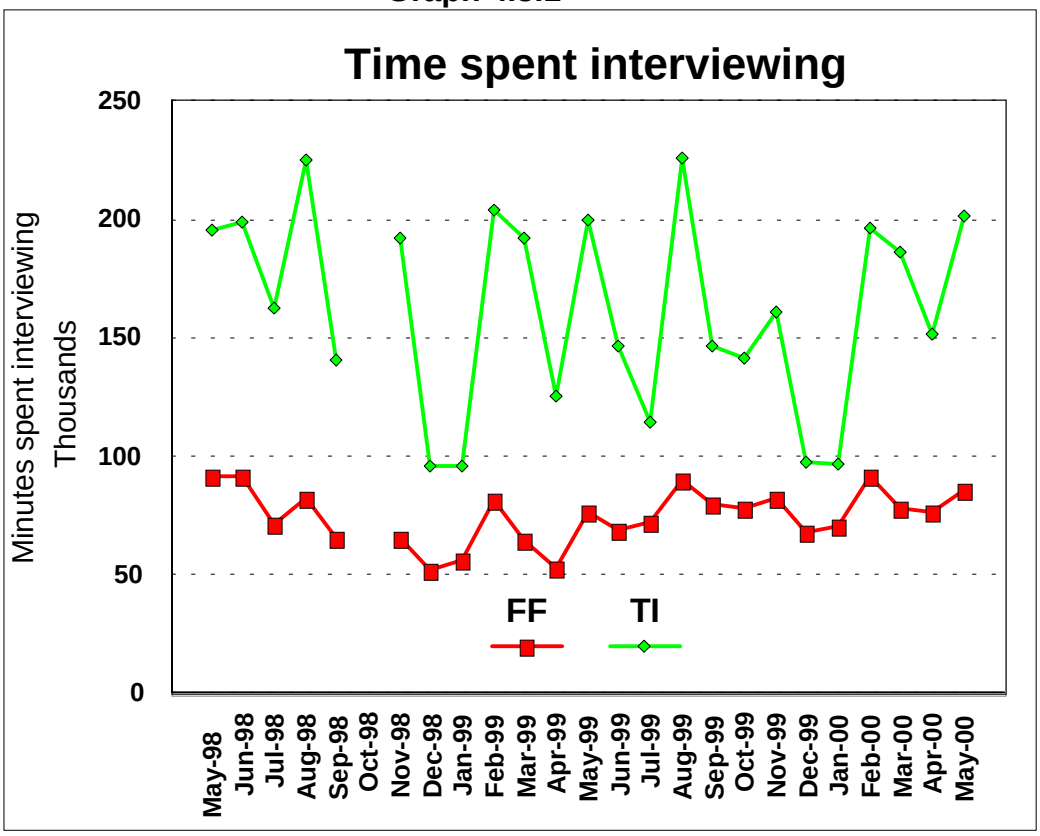
4.3.2 Data Quality

72. Whilst significantly more detailed data was available for the 2001 cost modelling, there were a large number of problems with the AP10 data used to estimate the cost model parameters. Many of these problems were related to the quality of OCR scanning of the AP10 forms. Lines were missing or incomplete, or numbers were scanned in incorrectly. Despite extensive editing work undertaken independently by both PSO and the Statistical Support Section to correct many of these problems, a number of data quality problems that are more difficult to resolve still remain. These data quality issues could affect the quality of the parameter estimates. A detailed list of the data qualify problems encountered can be found in Appendix J.

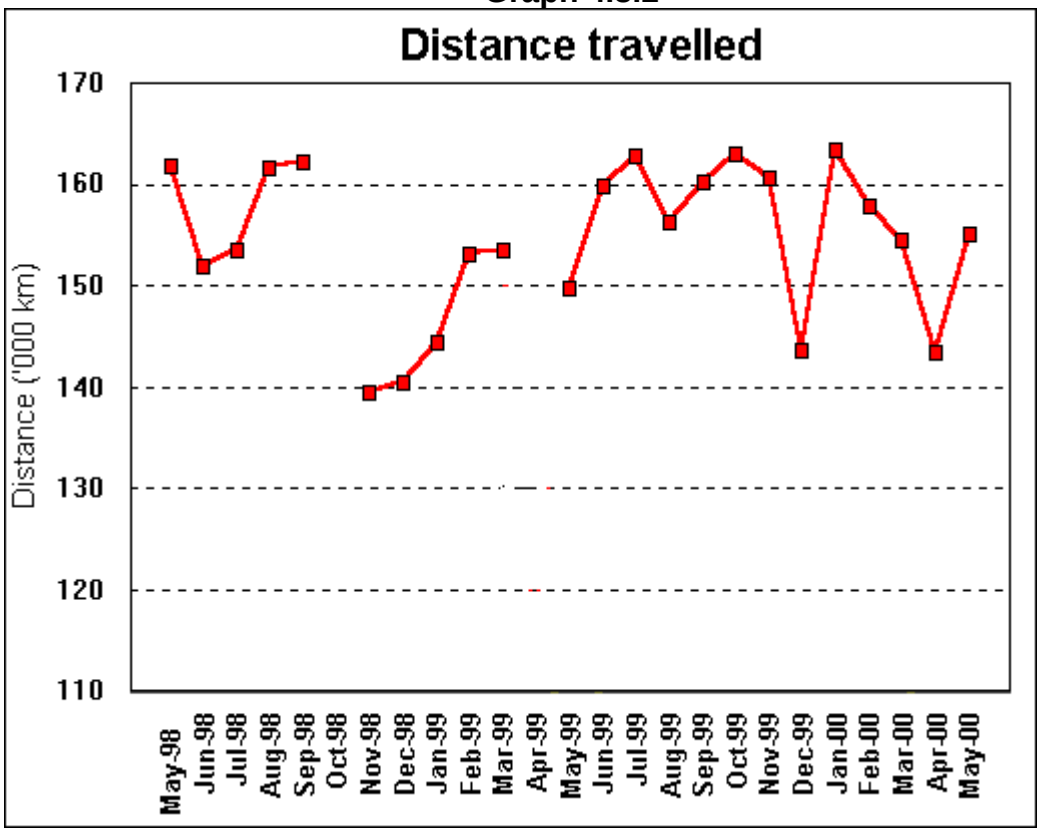
4.3.3 Timing

73. The cost model parameters have been estimated on the basis of only one month's worth of data, May 1999. To check whether May 1999 is representative of other months, time series based on AP10x data were produced for the two years from May 1998 to May 2000. Graph 4.3.1 displays the total time spent on interviewing (the major time related contributor to cost) broken down by face to face and telephone interviewing, while Graph 4.3.2 displays the total distance travelled for all activities. These graphs show that May 1999 is roughly average for time spent face to face interviewing, higher than average for time spent telephone interviewing, and slightly lower than average for distance travelled.

Graph 4.3.1


Time spent interviewing
Minutes spent interviewing
Thousands
250
200
150
100
50
0
FF
TI
May-98
Jun-98
Jul-98
Aug-98
Sep-98
Oct-98
Nov-98
Dec-98
Jan-99
Feb-99
Mar-99
Apr-99
May-99
Jun-99
Jul-99
Aug-99
Sep-99
Oct-99
Nov-99
Dec-99
Jan-00
Feb-00
Mar-00
Apr-00
May-00

Graph 4.3.2



74. Ideally, cost data similar to the May 1999 AP10 data should have been collected for a number of months to ensure a representative picture of the breakdown of costs into their constituent components. The large costs involved in collecting this data precluded this for the current design however.

4.3.4 Splitting Data Into Components

75. One of the major difficulties with the cost modelling is accounting for all the costs associated with the sample design of the MPS and splitting them into the correct cost components. For instance, there were a number of costs that were not included in the modelling because of a lack of data. Examples of c_0 costs not included are car/plane rental, training costs, line rental for telephones and furniture costs. The cost of making phone calls for telephone interviewing were not accounted for and could be added to c2 costs.

76. The cost that proved the most difficult to assign to a particular component was the cost associated with blocklisting. Blocklisting is a procedure that happens at the beginning of the design period when the new sample of CDs is selected. It has two components, setting up the CD by forming and listing blocks and listing the dwellings in one or two of the blocks in the CD (the remainder of the blocks are generally listed when they are needed). A single figure for the cost of blocklisting a CD was supplied by PSO for a number of broad area type groupings. For example, the cost of blocklisting in metropolitan areas was given as $150 per CD, whereas the cost of blocklisting in sparsely populated areas was given as $400 per CD. These costs are given in Appendix H. An average cost per month could then be calculated for each area type to be incorporated into the model.

77. The blocklisting costs consist of a number of components such as plane hire, car hire, fuel expenses, interviewer costs, etc. While some of these costs will depend on m_i, the number of selected CDs (or equivalently, number of selected blocks in a one block per CD design), the majority will not depend on m_i in a proportional manner, while some could perhaps be better interpreted as set up or overhead costs. For example, in remote areas, the cost of hiring a plane is not really proportional to the number of CDs that need blocklisting. Adding in a few extra CDs will not change the cost significantly, although generally speaking, the more CDs that are selected, the more blocks will need to be listed and the longer it will take. As the blocklisting cost could not be broken down into components, the entire cost was added in under the c_1 component. If the blocklisting cost components do not all change in proportion to m_i, this could affect the quality of the parameter estimates for explaining the cost structure of the MPS.

78. A final issue relates to the manner in which interview cost data was used in the 2001 cost modelling. In keeping with PSO practices, costs for interviewing were calculated using the number of interviews and a measure of average interview time called the mean assessed time. This mean assessed time is calculated in a given month as the average interview time for all interviews across Australia in that month. The current cost models therefore do not take account of any area type differences in interviewing times that contribute to overall costs. For example, if the interview times are significantly longer in one area type, then the average interview time and resulting interview cost can be reduced by undertaking less interviews in that area type. A further improvement to the cost modelling could therefore be made by using the interviewer's actual interview times to calculate interview costs at the area type level.

79. Table 4.3.1 provides a summary of the AP10 activities that contribute to each cost model parameter.

Table 4.3.1
Costs That Feed Into Each Cost Component

Inputs	**c_0**	**c_1**	**c_2**
Included in the calculation of cost parameter estimates	● Travel to and from the workload by the interviewer. ● Advice/counselling from supervisor.	● Travel between blocks by the Interviewer ● Checklisting ● Blocklisting	● Interview time using "mean assessed time" ● Between dwelling travel costs
Could be included in the calculation of cost parameter estimates for the next design	● Car/plane/telephone rental ● training ● furniture costs ● OHAS costs ● Car, plane hire and clerical costs associated with blocklisting	● Time and travel costs associated with blocklisting,	● Actual interview times ● Cost of telephone calls

4.3.5 Form of the Cost Model

80. A simple linear cost model has been used as the preferred cost model for describing the cost structure of the MPS since the 1986 redesign. A more complicated model was used prior to 1986. Assuming one block is selected per CD, as is usually the case, this model has the form

$$C_i = C_{0i} + C_{1i}\sqrt{m_i} + C_{2i}m_i + C_{3i}m_i q_i .$$

81. The result of using a cost model of the above form is that there is no direct algebraic solution for the optimal design parameters, m_i and q_i. While numerical methods can be used to solve for the optimal design, a historical lack of computing power has made this undesirable. Furthermore, some experts believe that this model offers no significant gain over the linear model given in (2). However, given advances in computer technology and changes in the type of cost data available, it is possible that the cost structure for the MPS could be better explained using this, or some other, non-linear model. This is an area that could be investigated in the future.

References

Clark (1997), "Optimal Clustering in Multi Stage Samples With Application to the 1996 Monthly Population Survey Redesign", an unpublished ABS internal paper for the Methodology Advisory Committee, June 1997.

Appendix A

82. The following table contains definitions of the sixteen area types used in he design of the MPS. The region for area type just indicates whether the area type is a metropolitan area type, or an extra-metropolitan area type. The terms SRA (Self Representing Area), sampled and sparse indicate the level of population density. In SRA areas, 50 dwelling selections per 4000 square kilometres are expected. Remaining area types are either sampled or sparse, depending on whether they have a density of more or less than 0.06 dwellings per square kilometre, respectively. The indigenous area type consists of CDs which are in sparse areas and in which 70 percent or more of the population are indigenous. The growth area type consists of CDs that are identified by PSO as expecting high dwelling growth over the period of the redesign.

Table A.1
Definition of Area Types

Area Type	Region	Definition
1	Met	Inner city Melbourne/Sydney
2	Met	Inner city (pop density >3125/square kilometre)
3	Met	Settled area with increase in private dwellings < 10% since 1991
4	Met	Outer growth area with an increase in private dwellings of at least 10% since 1991
5	Met	*Other Urban**
6	Met	Rural
7	Ex-met	Large town with population of at least 8000
8	Ex-met	Small town with population < 8000
9	Ex-met	Rural SRA
10	Ex-met	Urban sampled
11	Ex-met	Rural sampled
12	Ex-met	Sparse
13	Ex-met	Indigenous
14	Met	Growth
15	Met	Hobart
16	Met	Darwin

* No CDs in Australia fall into area type 5.

Appendix B

83. The fifty design options considered in the variance modelling were centred around the optimal 1996 design option and were generated by combining one of ten different set of state skip options for selecting clusters with one of five different cluster size options. For all area types, each of the ten fixed sets of state skips was a multiple, α, of the 1996 values. This ensures that the state skips are altered by the same proportion across all states for each option. The values of α used and the resulting ten sets of state skips are given below. The 1996 state skips, α=1, are highlighted in bold.

Table B.1
State Skip Options

State \ α	0.5	0.6	0.8	0.9	1.0	1.1	1.2	1.3	1.5	2.0
NSW	150	180	240	270	**300**	330	360	390	450	600
VIC	129	154	206	231	**257**	283	308	334	386	514
QLD	111	133	178	200	**222**	244	266	289	333	444
SA	74	88	118	132	**147**	162	176	191	221	294
WA	80	96	128	144	**160**	176	192	208	240	320
TAS	42	50	66	75	**83**	91	100	108	125	166
NT	43	51	68	77	**85**	94	102	111	128	170
ACT	43	51	68	77	**85**	94	102	111	128	170

84. The cluster sizes considered for each area type were centred ±1 and ±2 either side of the 1996 optimal values for that area type. The five cluster size options considered for each area type are given below. The 1996 optimal cluster sizes, q_{i3}, are highlighted in bold.

Table B.2
Cluster Size Options

Area Type	q_{i1}	q_{i2}	q_{i3}	q_{i4}	q_{i5}
1. Inner City Melbourne/Sydney	3	4	**5**	6	7
2. Inner City	4	5	**6**	7	8
3. Settled Area	5	6	**7**	8	9
4. Outer Growth	3	4	**5**	6	7
6. Met Rural	6	7	**8**	9	10
7. Large Town	6	7	**8**	9	10
8. Small Town	6	7	**8**	9	10
9. Ex-met Rural SRA	8	9	**10**	11	12
10. Urban sampled	6	7	**8**	9	10
11. Rural sampled	8	9	**10**	11	12
12. Sparse	8	9	**10**	11	12
13. Indigenous	7	8	**9**	10	11
14. Growth	3	4	**5**	6	7
15. Hobart	6	7	**8**	9	10
16. Darwin	6	7	**8**	9	10

85. Each state skip multiple, α, and cluster size option, q_i, gives rise to a set number of area type first stage selections, m_i, and hence a single sample design option ($\underline{m}$,$\underline{q}$) as follows:

86. Let D_h denote the number of dwellings in the population in stratum h in area type i. Let q_h denote the cluster size in stratum h (=q_i where stratum h lies in area type i). The total number of clusters, M_h, in stratum h is given by

$$M_h = \frac{D_h}{q_h}.$$

87. Let k_h denote the 1996 state skip for stratum h, which will be the state skip for the state to which stratum h belongs. The number of clusters selected from stratum h when the state skip multiplier is α and the cluster size is q_h, denoted by m_h, is given by:

$$m_h = \frac{\text{Total number of clusters}}{\text{skip for stratum h}}$$

$$= \frac{D_h / q_h}{\alpha k_h}.$$

88. The number of clusters to select is not likely to be a whole number in practice, and this value is therefore rounded to the nearest positive integer to determine the number of clusters to select from each stratum for each sample under the given design option. This rounding operation is achieved by adjusting the design option cluster size to be used in each stratum. Let m'_h denote the rounded integral number of clusters to be selected from stratum h. That is

$$m'_h = Round\left(\frac{D_h / q_h}{\alpha k_h}\right)$$

89. If the proposed design option cluster size is q_h, the adjusted cluster size used in stratum h, q'_h, is defined to be the value for which

$$m'_h = \frac{D_h}{\alpha k_h q'_h}$$

ie

$$q'_h = \frac{D_h}{\alpha k_h m'_h}.$$

90. The adjusted cluster size q'_h is then the cluster size used to break the CDs in stratum h down into blocks and clusters for sample selection.

This rounding process occurs for every stratum in area type i under the given state skip and cluster size option. The number of cluster selections, m_i, from area type, i, is then given by

$$m_i = \sum_{h \in i} m'_h.$$

Appendix C

91. The tables below give the variance model parameter estimates at the area type level under an independent stratum level selection methodology. Tables C.1 and C.2 give the employment and unemployment variance model parameter estimates calculated by fitting a regression model of the form (1). The parameter estimates highlighted in bold are the estimates that are not significantly different from zero at the 0.05 significance level.

Table C.1
Variance Model Parameter Estimates For Variable "Employed Persons" By Area Type

Area Type	v_{0i}	v_{1i}	v_{2i}	Adj-R^2
1. Inner City Melbourne/Sydney	-4,567,489	1,758,811,044	15,600,214,849	0.9408
2. Inner City	**-3,004,876**	3,074,148,084	99,628,266,024	0.9883
3. Settled Area	-49,168,860	231,702,047,470	5,023,579,200,000	0.9971
4. Outer Growth	-29,487,622	142,221,095,829	1,905,634,500,000	0.9948
6. Met Rural	-4,569,073	1,548,946,229	20,437,197,164	0.9888
7. Large Town	-48,331,944	94,478,827,512	1,353,076,800,000	0.9894
8. Small Town	-11,460,463	8,920,985,818	137,741,066,584	0.9952
9. Ex-met Rural SRA	-6,828,881	3,774,496,906	61,104,568,652	0.9966
10. Urban sampled	-4,741,546	4,756,638,768	90,039,117,072	0.9947
11. Rural sampled	-9,397,265	8,685,268,057	127,434,810,594	0.9975
12. Sparse	-1,094,102	334,567,174	2,592,872,414	0.9832
13. Indigenous	-1,027,979	19,972,757	**43,495,831**	0.8517
14. Growth	**-12,713**	**44,978**	2,872,143	0.8072
15. Hobart	**-335,363**	270,771,761	4,501,217,974	0.8868
16. Darwin	**-297,012**	**31,158,362**	868,882,300	0.8433

Table C.2
Variance Model Parameter Estimates For Variable "Unemployed Persons" By Area Type

Area Type	v_{0i}	v_{1i}	v_{2i}	Adj-R^2
1. Inner City Melbourne/Sydney	-363,972	**30,828,822**	3,266,802,049	0.9801
2. Inner City	274,638	123,473,496	11,371,628,202	0.9953
3. Settled Area	-1,387,461	5,371,175,258	605,259,010,880	0.9984
4. Outer Growth	-920,633	3,580,044,660	284,911,885,858	0.9977
6. Met Rural	**-95,368**	21,533,541	2,128,685,546	0.9863
7. Large Town	-1,008,371	3,984,504,305	214,469,790,410	0.9978
8. Small Town	**-239,450**	233,357,501	23,934,712,334	0.9964
9. Ex-met Rural SRA	-445,348	165,463,766	9,764,477,010	0.9947
10. Urban sampled	**-79,471**	167,101,501	12,934,891,602	0.9964
11. Rural sampled	-483,811	304,027,193	18,936,593,955	0.9971
12. Sparse	**-39,324**	7,859,151	260,195,585	0.9844
13. Indigenous	**-64,405**	4,952,138	**2,826,880**	0.9425
14. Growth	**-2,094**	**5,798**	318,991	0.8266
15. Hobart	-95,591	21,967,061	634,102,128	0.9644
16. Darwin	**11,585**	**2,547,843**	89,806,217	0.8281

Appendix D

92. The tables below show the volatility of the variance model parameter estimates at the area type level for each of employed and unemployed persons. It gives the relative standard error (RSE) of each parameter as a percentage of the size of the parameter estimate.

Table D.1
Relative Standard Error % For Variance Model Parameter Estimates For Variable "Employed Persons" By Area Type

Area Type	RSE%(v_{0i})	RSE%(v_{1i})	RSE%(v_{2i})
1. Inner City Melbourne/Sydney	46%	13%	9%
2. Inner City	66%	15%	3%
3. Settled Area	13%	6%	2%
4. Outer Growth	17%	5%	2%
6. Met Rural	22%	8%	5%
7. Large Town	16%	9%	5%
8. Small Town	12%	6%	3%
9. Ex-met Rural SRA	12%	6%	4%
10. Urban sampled	26%	7%	3%
11. Rural sampled	12%	4%	3%
12. Sparse	47%	8%	12%
13. Indigenous	27%	18%	79%
14. Growth	92%	120%	12%
15. Hobart	139%	32%	17%
16. Darwin	87%	56%	18%

Table D.2
Relative Standard Error % For Variance Model Parameter Estimates For Variable "Unemployed Persons" By Area Type

Area Type	RSE%(v_{0i})	RSE%(v_{1i})	RSE%(v_{2i})
1. Inner City Melbourne/Sydney	45%	57%	3%
2. Inner City	47%	25%	2%
3. Settled Area	34%	18%	1%
4. Outer Growth	41%	15%	1%
6. Met Rural	81%	45%	4%
7. Large Town	39%	11%	2%
8. Small Town	61%	24%	2%
9. Ex-met Rural SRA	27%	18%	3%
10. Urban sampled	143%	19%	2%
11. Rural sampled	26%	14%	2%
12. Sparse	70%	19%	6%
13. Indigenous	55%	9%	155%
14. Growth	59%	98%	11%
15. Hobart	32%	26%	8%
16. Darwin	231%	72%	18%

Appendix E

93. The following table gives the contribution form the residuals as a percentage of the range against which those residuals are derived for each area type model. The second column gives the root mean square error on the employment model residuals as a percentage of the range of employment variances over which the model for employment was fitted. The third column gives the root mean square error on the unemployment model residuals as a percentage of the range of unemployment variances over which the model for unemployment was fitted.

Table E.1
Ratio of Residual Root Mean Square Error To Variance Range For Employed And Unemployed Persons Variables At Area Type Level

Area Type	Employment	Unemployment
1. Inner City Melbourne/Sydney	5.77%	3.67%
2. Inner City	2.75%	1.90%
3. Settled Area	1.37%	1.11%
4. Outer Growth	1.77%	1.29%
6. Met Rural	2.65%	3.20%
7. Large Town	2.65%	1.23%
8. Small Town	1.77%	1.63%
9. Ex-met Rural SRA	1.48%	1.92%
10. Urban sampled	1.86%	1.65%
11. Rural sampled	1.22%	1.47%
12. Sparse	3.23%	3.50%
13. Indigenous	10.39%	6.10%
14. Growth	11.75%	10.65%
15. Hobart	7.66%	4.55%
16. Darwin	9.50%	10.57%

Appendix F

94. The tables below give the variance model parameter estimates at the area type level under an independent state level selection methodology. Tables F.1 and F.2 give the employment and unemployment variance model parameter estimates calculated by fitting a regression model of the form (1). Table F.3 gives the parameter estimates of the hybridised relative variance model to be used to calculate the relative variance constraint. The fourth column of Table F.3 shows the relative variances achieved under the 1996 optimal design at the area type level and the resulting total national relative variance constraint to be used in the optimisation.

95. The parameter estimates highlighted in bold in Tables F.1 and F.2 are the estimates that are not significantly different from zero at the 0.05 significance level.

Table F.1
Variance Model Parameter Estimates For Variable "Employed Persons" By Area Type Under State Based Selections

Area Type	v_{0i}	v_{1i}	v_{2i}	Adj-R^2
1. Inner City Melbourne/Sydney	-4,567,489	1,758,811,044	15,600,214,849	0.9408
2. Inner City	**-1,347,520**	2,525,265,328	102,246,567,188	0.9727
3. Settled Area	**-29,111,067**	192,788,073,507	5,119,229,100,000	0.9806
4. Outer Growth	**-12,453,951**	184,093,659,727	1,648,380,100,000	0.9769
6. Met Rural	-8,776,502	1,969,616,124	19,851,662,894	0.9715
7. Large Town	-39,133,896	76,584,293,081	1,434,885,400,000	0.9627
8. Small Town	-11,806,807	9,248,714,180	130,224,845,697	0.9818
9. Ex-met Rural SRA	-9,019,855	4,091,730,021	63,560,097,499	0.9915
10. Urban sampled	**-2,304,332**	4,712,088,187	89,289,587,025	0.9698
11. Rural sampled	-10,090,979	8,775,198,497	124,421,481,137	0.9844
12. Sparse	-2,976,963	388,875,129	2,379,611,356	0.9651
13. Indigenous	-1,027,979	19,972,757	**43,495,831**	0.8517
14. Growth	**-12,713**	**44,978**	2,872,143	0.8072
15. Hobart	**-190,000**	**123,263,709**	5,524,082,550	0.7390
16. Darwin	**-297,012**	**31,158,362**	868,882,300	0.8433

Table F.2
Variance Model Parameter Estimates For Variable "Unemployed Persons" By Area Type Under State Based Selections

Area Type	v_{0i}	v_{1i}	v_{2i}	Adj-R^2
1. Inner City Melbourne/Sydney	-363,972	**30,828,822**	3,266,802,049	0.9801
2. Inner City	**147,970**	**54,330,066**	11,899,388,666	0.9788
3. Settled Area	**-83,417**	**2,583,069,612**	615,876,165,740	0.9843
4. Outer Growth	**324,690**	3,157,678,547	278,882,774,198	0.9886
6. Met Rural	**-163,741**	**21,908,220**	2,178,982,101	0.9813
7. Large Town	**99,995**	4,284,103,057	205,020,367,404	0.9830
8. Small Town	**-429,112**	349,363,217	23,563,684,111	0.9868
9. Ex-met Rural SRA	**-225,963**	**77,450,745**	10,358,909,735	0.9868
10. Urban sampled	**-525,446**	231,461,358	12,831,057,841	0.9827
11. Rural sampled	**-80,488**	230,383,534	19,552,922,220	0.9862
12. Sparse	**-42,584**	7,040,012	277,185,349	0.9752
13. Indigenous	**-64,405**	4,952,138	**2,826,880**	0.9425
14. Growth	**-2,094**	**5,798**	318,991	0.8266
15. Hobart	**-145,297**	28,442,334	642,138,204	0.8402
16. Darwin	**11,585**	**2,547,843**	89,806,217	0.8281

96. Comparison of Tables F.1 and F.2 with tables C.1 and C.2, show that the employment and unemployment parameter estimates are identical for state based and stratum based selections for area types 1, 13, 14 and 16. This is because there is at most only one stratum contributing to each of these area types from each state and the two selection methodologies are therefore equivalent.

Table F.3
Hybridised Relative Variance Model Parameter Estimates By Area Type Under State Based Selections x 10^{-10}

Area Type	v_{0i}	v_{1i}	v_{2i}	1996 RV
1. Inner City Melbourne/Sydney	-1,372	337,979	8,209,462	15,581
2. Inner City	40	502,790	37,245,450	33,096
3. Settled Area	-4,851	35,660,600	1,899,788,253	206,477
4. Outer Growth	-1,448	35,255,629	752,009,832	131,070
6. Met Rural	-1,704	356,523	7,002,707	12,295
7. Large Town	-6,152	19,838,449	588,899,116	112,140
8. Small Town	-2,656	2,103,280	62,076,470	33,079
9. Ex-met Rural SRA	-1,851	796,257	28,309,871	18,189
10. Urban sampled	-1,287	1,164,679	36,773,126	27,958
11. Rural sampled	-1,771	1,819,585	54,155,329	25,739
12. Sparse	-555	75,117	867,255	4,349
13. Indigenous	-278	11,851	11,953	1,918
14. Growth	-6	17	1,020	22
15. Hobart	-284	69,445	2,010,955	2,435
16. Darwin	-28	9,473	296,810	1,031
Total				**625,378**

Appendix G

97. The sixteen fine area types were collapsed into (different) broad groupings of data in order to undertake the cost modelling for both the 1996 and 2001 redesigns. The following table contains definitions of the area type groups used under each design. Column 2 indicates the area type groupings for 1996 and Column 3 indicates the area type groupings for 2001.

Table G.1
Area Type Definitions and Groupings For 1996 and 2001 Cost Modelling

Area Type	1996 Groups	2001 Groups
1. Inner City Melbourne/Sydney	1	1
2. Inner City	1	2
3. Settled Area	1	3
4. Outer Growth	1	4
5. *Other Urban**	*n/a*	*n/a*
6. Met Rural	3	7
7. Large Town	2	5
8. Small Town	2	6
9. Ex-met Rural SRA	3	7
10. Urban sampled	n/a	8
11. Rural sampled	n/a	9
12. Sparse	n/a	9
13. Indigenous	n/a	9
14. Growth	n/a	4
15. Hobart	4	10
16. Darwin	5	11

* No CDs in Australia fall into area type 5.

Appendix H

98. The following table gives the blocklisting costs for a CD for the life of the 1996 design in each of a number of broad area type groupings.

Table H.1
Cost Associated With Blocklisting a CD

Cost Per CD Per Design Period in $	Area Type Group	Area Types Included
150	MET	1, 2, 3, 4, 6, 14, 15, 16
200	Ex-met Urban SRA	7, 8
250	Ex-met Rural SRA	9
300	Ex-met Urban sampled	10
350	Ex-met Rural sampled	11
400	Ex-met Sparse	12, 13

Appendix I

99. To evaluate the quality of the cost models, variances, and hence relative standard errors, were calculated for each of the cost model parameters. As different methodologies were used to estimate each of the cost model parameters, different methodologies were also required to calculate the variances.

Overhead Cost - c_0

100. The variance can be obtained for c_0 from the standard errors on the regression model parameters in Equation (3). Consider area type i in area type group j. Let α_j denote the regression parameter estimate for area type group j and let p'_{iw} denote the proportion of blocks from workload w in area type i, derived from May 1999 LFS data. Then the variance formula is as follows:

$$Var(c_{0i}) = Var(\sum_{w \in i} \alpha_j p'_{iw})$$
$$= \sum_{w \in i} p'^{2}_{iw} Var(\alpha_j)$$

where the sum is over all workloads in Australia falling into area type i. The variances on the c_{0i} estimates can therefore be calculated using the standard error output from the regression analysis.

Per Block Cost - c_1

101. The per block cost can be considered to be the average of the individual per block costs for the blocks covered by the AP10 data (although this is a simplification of the method of calculation actually used). The formula for measuring the volatility of c_1 can therefore be derived by assuming that c_1 is the estimator of average from a set of sample realisations of individual block values, the sample of blocks being defined by the 80% of workloads covered by the AP10 data. The variance on c_1 can be calculated using the formula:

$$Var(c_{1j}) = \frac{1}{B_j} \sum_{b \in j} (c_{1jb} - c_{1j})^2$$

where:

B_j is the number of blocks in area type group j
c_{1j} is the c_1 cost parameter for area type group j
c_{1jb} is the c_1 cost for block b in area type group j

102. A number of blocks are completely enumerated under telephone interviewing. The block related costs are constant for these blocks as there are zero between block travel costs, zero checklisting costs, and the same constant cost for blocklisting as is attributable to all blocks in an area type. As the variance component for these blocks is zero, the variance formula can be simplified by splitting the average per block costs between the blocks which are completely enumerated under telephone interviewing and the blocks for which at least one interview is conducted using face to face interviewing. First, c_{1j} can be expressed as:

$$c_{1j} = \frac{B_{j\_ff}}{B_j} c_{1j\_ff} + \frac{B_{j\_ti}}{B_j} c_{1j\_ti}$$

where:

B_{j_ff} is the number of blocks in area type group j that have at least one person enumerated by face to face interviewing

B_{j_ti} is the number of blocks in area type group j that are completely enumerated by telephone interviewing

c_{1j_ff} is the average c_1 cost for blocks in area type group j that have at least one person enumerated by face to interviewing

c_{1j_ti} is the average c_1 cost for blocks in area type j that are completely enumerated by telephone interviewing

Thus the variance become

$$Var(c_{1j}) = Var(\frac{B_{j\_ff}}{B_j} c_{1j\_ff}) + Var(\frac{B_{j\_ti}}{B_j} c_{1j\_ti})$$

$$= (\frac{B_{j\_ff}}{B_j})^2 \frac{1}{B_{j\_ff}} \left(\frac{1}{B_{j\_ff} - 1} \sum_{\substack{b \in j \\ b \in ff}} (c_{1jb\_ff} - c_{1j\_ff})^2 \right) + (\frac{B_{j\_ti}}{B_j})^2 \frac{1}{B_{j\_ti}} \left(\frac{1}{B_{j\_ti} - 1} \sum_{\substack{b \in j \\ b \in ti}} (c_{1jb\_ti} - c_{1j\_ti})^2 \right)$$

$$= (\frac{B_{j\_ff}}{B_j})^2 \frac{1}{B_{j\_ff}} \left(\frac{1}{B_{j\_ff} - 1} \sum_{\substack{b \in j \\ b \in ff}} (c_{1jb\_ff} - c_{1j\_ff})^2 \right)$$

where:

c_{1jb_ff} is the c_1 cost for block b in area type group j, where block b contains at least one person who is enumerated by face to face interviewing

c_{1jb_ti} is the c_1 cost for block b in area type group j, where block b is completely enumerated by telephone interviewing.

Per Dwelling Cost - c_2

103. The method for calculating the standard error on the c_2 parameter estimate is very similar to the method adopted for the c_1 parameter estimate. In this case the per dwelling cost, c_2, is considered to be the estimator of average from a set of sample realisations of individual dwelling costs. As individual dwelling costs can not actually be determined from the available data, an additional assumption is made that the costs at the dwelling level are the same within a block. The formula for calculating the variance on c_2 is then given by:

$$Var(c_{2j}) = Var(\frac{D_{j\_ff}}{D_j} c_{2j\_ff}) + Var(\frac{D_{j\_ti}}{D_j} c_{2j\_ti})$$

$$= (\frac{D_{j\_ff}}{D_j})^2 \frac{1}{D_{j\_ff}} \left(\frac{1}{D_{j\_ff} - 1} \sum_{\substack{d \in j \\ d \in ff}} (c_{2jd\_ff} - c_{2j\_ff})^2 \right) + (\frac{D_{j\_ti}}{D_j})^2 \frac{1}{D_{j\_ti}} \left(\frac{1}{D_{j\_ti} - 1} \sum_{\substack{d \in j \\ d \in ti}} (c_{2jd\_ti} - c_{2j\_ti})^2 \right)$$

$$= (\frac{D_{j\_ff}}{D_j})^2 \frac{1}{D_{j\_ff}} \left(\frac{1}{D_{j\_ff} - 1} \sum_{\substack{d \in j \\ d \in ff}} (c_{2jd\_ff} - c_{2j\_ff})^2 \right)$$

$$= (\frac{D_{j\_ff}}{D_j})^2 \frac{1}{D_{j\_ff}} \left(\frac{1}{D_{j\_ff} - 1} \sum_{\substack{b \in j \\ d \in ff}} d_{bj\_ff} (c_{2jb\_ff} - c_{2j\_ff})^2 \right)$$

where

D_j is the number of dwellings selected in area type group j

D_{j_ff} is the number of dwellings in area type group j that are enumerated under face to face interviewing

D_{j_ti} is the number of dwellings in area type group j that are enumerated by telephone interviewing

D_{jb_ff} is the number of dwellings in block b and area type group j that are enumerated by face to face interviewing

c_{2j_ff} is the average c_2 cost for dwellings in blocks in area type group j that have at least one person enumerated by face to interviewing

c_{2j_ti} is the average c_2 cost for dwellings in blocks in area type j that are completely enumerated by telephone interviewing.

c_{2jd_ff} is the c_2 cost parameter for a dwelling in area type group j that is enumerated by face to interviewing

c_{2jd_ti} is the c_2 cost parameter for a dwelling in area type group j that is enumerated by telephone interviewing

c_{2jb_ff} is the c_2 cost for block b in area type group j where b is a block in which there is at least one person that is enumerated by face to face interviewing

104. The variance between dwellings under telephone interviewing is again zero as the cost for telephone interviewing is constant for all dwellings in a block (there is zero block travel, and each interview is charged at the same flat rate for all dwellings). As with the calculation of the per dwelling estimate of c_2, all dwellings are included in the dwelling counts used in the calculation. That is, dwelling counts include both the dwellings for which an interview is conducted and those dwellings contributing to sample loss for which no interview is conducted.

105. Table I.1 shows the volatility of the cost model parameters at the area type level. It gives the relative standard error (RSE) of each parameter as a percentage of the size of the parameter estimate, ie

$$RSE\%(c_{ki}) = \frac{\sqrt{Var(c_{ki})}}{c_{ki}} * 100$$

for each parameter c_{kj} k=0,1,2. As with the parameter estimates themselves, the RSEs on the c_1 and c_2 parameters estimates are calculated at the grouped area type level and applied to all area types within the group.

Table I.1
Relative Standard Error % For Cost Model Parameters At Area Type Level

Area Type	RSE%(c_{0i})	RSE%(c_{1i})	RSE%(c_{2i})
1. Inner City Melbourne/Sydney	6.45	7.05	0.16
2. Inner City	4.26	6.21	0.04
3. Settled Area	0.66	2.30	0.04
4. Outer Growth	0.71	2.51	0.04
6. Met Rural	2.16	6.21	0.22
7. Large Town	0.81	4.41	0.10
8. Small Town	2.03	6.89	0.07
9. Ex-met Rural SRA	1.63	6.21	0.22
10. Urban sampled	1.29	10.29	0.13
11. Rural sampled	1.21	7.11	1.02
12. Sparse	2.78	6.85	1.02
13. Indigenous	3.80	6.85	1.02
14. Growth	6.36	2.51	0.04
15. Hobart	4.36	8.31	0.06
16. Darwin	13.24	16.98	1.14

Appendix J

106. Table J.1 summarises the main data quality problems that were identified with the AP10 data.

Table J.1
AP10 Data Quality Issues

Problem	Number of Workloads affected	Number of Workloads Corrected	Comments
Time incorrect/missing	approx. 100	most	Sometimes it was difficult to work out what the missing time was supposed to be, especially if it was missing for a few activities in a row.
Odometer reading incorrect/missing	approx. 100	most	About 10 workloads only provided distance measures for groups of activities and it was impossible to split the distance between individual activities.
Blank Lines	96	96	Blank lines cause unrealistic times and distances to be calculated (such as negative times and distances)
Interviewer's time or distance totals on AP10 did not match to AP10x	35 (only those with a large discrepancy were investigated)	Can not be fixed	When the data on the AP10 forms were added up, the totals were less (although sometimes more) than the provided AP10x totals. There were no apparent reasons for these differences.
Activity incorrect or missing	20	20	There were more that were fixed without being recorded.
Pages Missing	16	Can not be fixed	These pages were not scanned in and therefore could not be replicated.
State incorrect	14	14	This caused problems with either matching to AP10x or where one page had one state and another page had another state (as workloads lie within one state).
Workload number incorrect	12	12	This is when one page had one workload number, and the next page had another workload number.
Page number incorrect	12	12	Calculations of time and distances for an activity can carry over from one page to the next, if the pages are in

			the wrong order (due to numbering being incorrect) unrealistic times and distances can be calculated.
Extra Lines written in by interviewer.	at least 4 which were only picked up due to other errors for that workload.	4 have been corrected	There are probably more cases of this, but these were discovered as they caused unrealistic times or distances to be calculated

–