

A GUIDE TO SMALL AREA ESTIMATION - EXECUTIVE SUMMARY

1. Introduction

This paper is an executive summary of the small area practice manual developed by the Australian Bureau of Statistics (ABS, 2005). An earlier version of this paper was tabled as a room document to APEX 2 (McEwin and Elazar, 2006). This manual aims to provide a guide to the production, uses, quality and validation of small area estimates for survey practitioners and users of small area data. It is directed towards potential users of small area data as well as officers in national statistics centres and regional office client service areas of the ABS.

Small area estimation is a methodology for producing estimates for a more detailed level of geography than can be reliably obtained from direct survey estimates. In Australia, outputs are often sought for Local Government Areas (LGAs) where information is needed to assist in planning and delivering services to people and businesses. These typically range in size from 10,000 to 200,000 people. Conceptually similar to these are *small domain estimates*, which are disaggregated to fine classification levels (eg industry, income group or labour force status). Although the techniques outlined here have been used for small areas, we would expect that they could also be used in small domain applications.

The techniques to calculate small area estimates outlined in the practice manual combine the use of survey data and auxiliary data sources such as administrative data (data collected by organisations as part of their ongoing business operations) to produce estimates for a more detailed level of geography than can be reliably obtained from direct survey estimates. That is, the small area estimates produced by these techniques are new statistics that are not otherwise available from survey or administrative data sources. Some administrative data also can be used to produce statistics for small areas and ABS encourages this wherever possible as a cost effective means of obtaining data which reduces respondent load, improves scope and coverage and increases the availability of longitudinal and small area data.

ABS acknowledges the demand for small area data to support planning, decision making and service delivery at a local area level. ABS also recognises the role that analytical methods may have for producing official statistics. In doing so, there is a need to ensure methods and assumptions are described for users and the validity of the modelled estimates is assessed.

The ABS small area practice manual aims to:

- increase knowledge and understanding of small area estimation techniques and ensure greater consistency in their application;
- provide a guide for choosing the best method to apply for a particular situation; and
- advise on the trade-off between the complexity of methods and the cost required to produce the modelled estimates.

2. Assessing user requirements

Undertaking a small area modelling exercise is likely to involve significant time and effort. It is therefore important to ensure that this effort is well justified in terms of the importance of the use and the likely quality and fitness for purpose of the estimates derived. The following questions for users will assist with this decision. Some questions are directed more to understanding the needs and uses of the data with a view to establishing the importance and priority of the work. Others provide valuable information to assist in approaching the task and developing a modelling framework.

- a) *What are the nature and context of the key planning or funding decisions that require small area data?*

It is essential at the outset to establish then articulate the data problem. What are users trying to find out and why? How will local area data be used in their decision making processes? It is also important to ensure at this stage of the project that users are made aware of what might be feasible so that their expectations are appropriately managed.

- b) *What variables or indicators are needed to meet these decision making requirements? What disaggregations of these are important and why? What level of geography is needed?*

Users often request a range of variables and classifications they would like to have but which are not all equally important for making decisions. It is useful to determine the minimum amount of data and geographic detail required to meet the most important needs. Prioritising user requirements can assist in dividing the project into phases so that work can focus initially on a limited set of the most important items. Users should also be made aware of the trade-off between the size of the desired geographic area and the degree of cross classification required.

- c) *What level of accuracy is needed and which small area estimates have the greatest priority in terms of accuracy?*

If the small area estimates are to be used for decisions such as funding of key programs, then a higher level of quality assessment and validation, in consultation with users, will be required to ensure that results are of an acceptable level of quality. However, if, for example, small area estimates are to be used as a guide to indicate areas of unmet demand, then a simpler process may be adequate.

- d) *What theory is available to help identify the models which can be used to produce small area estimates?*

Expert advice on the factors likely to be good explanators for the variable(s) of interest is vital (eg disability is known to increase with age). Such information may be available from statistical subject matter experts, from academics in the

relevant field or from policy advisors in government agencies. Data exploration may identify correlations between variables to confirm these relationships.

- e) *What auxiliary data are available to support the modelling process? How are the data collected, for what purpose are they used, and how accurate are they likely to be?*

The success of small area estimation usually depends critically on the existence of census or administrative data that can be related to the variable(s) of interest. For example, data on the age structure of an area obtained from a census might be used to model disability for that area, or administrative data on the number of people receiving unemployment benefits might be used to model unemployment rates.

As administrative data are collected for a range of different purposes, there are invariably quality issues that need to be considered when choosing auxiliary data. For example, the scope may not include the entire population of interest; classifications used may differ; or insufficient quality control or editing processes may result in errors.

- f) *What previous studies have been undertaken that are similar to the current small area problem?*

Previous small area studies can provide valuable insight for addressing the current problem. This can be useful in identifying what approaches worked well but also what did not work so well and why. Previous studies can also be used for comparative analysis or validation. Previous small area estimation projects undertaken by the ABS are tabled in an appendix to this paper.

3. Small area estimation techniques

Each small area problem needs to be carefully assessed to ensure that the approach taken and techniques applied suit the particular problem at hand.

The choice of small area method depends on the availability of auxiliary data and the relationship between these data and the variables of interest at the small area level. In essence, we are looking to "borrow strength" from these auxiliary data to increase the accuracy of the estimates. Small area models range from the simple to the more complex, the latter requiring considerably more time, effort, technical skill and available data. A range of quantitative and qualitative diagnostics should be used to choose the best model for the given data. The modelling framework (see Figure 1) can be summarised as follows:

- (i) direct estimator - this is a standard method for ABS surveys where estimates are obtained by applying survey weights to those sample units selected in each small area. However, since most ABS surveys are designed to provide reliable estimates only at national or state levels, sample sizes are often too small at this level to produce reliable direct estimates. Furthermore, there are often situations where there is no sample in the areas of interest (eg not all LGAs are sampled in every survey).

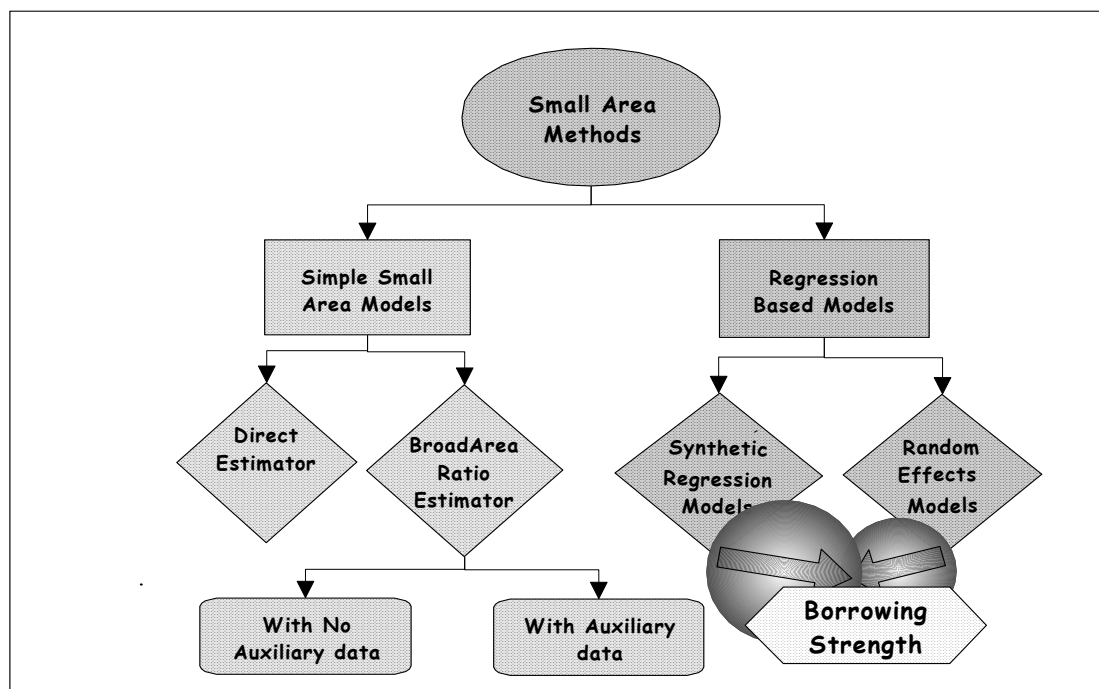
Small area estimation methods are pursued when direct estimation fails due to inaccuracy or inadequate sample. When undertaking a small area modelling exercise, it is often useful to produce the direct estimates to compare with modelled estimates. They can then contribute to quality diagnostics of bias and additivity.

(ii) broad area ratio estimator (BARE) - this is one of the simplest types of small area models. It is calculated by applying the rate for a broad area obtained from a survey (eg disability rate or unemployment rate or poverty rate) to the small area populations (available from say a population census or demographic estimates). The success of the BARE hinges largely on the choice of the broad area. The broad area (eg urban/rural) needs to be large enough to allow for a reliable direct survey estimate but small enough so that the small areas within the broad area can be assumed to be homogenous in the characteristic of interest. This is a strong assumption and users need to be made aware of it. As with direct estimation, BARE can be used to validate more complex approaches.

(iii) broad area ratio estimator with auxiliary data - this uses information that is correlated with the variable of interest and is available at the small area level to derive an estimate that adjusts for compositional differences in small areas. For example, it can be demonstrated from surveys that disability is correlated with age. This means that those areas with a larger proportion of older people would be expected to have a higher rate of disability. The estimator applies age specific disability rates to a small area population classified by age group. It is a deterministic model that assumes that disability only varies by age and does not allow for other effects. It can be applied in association with a broad area ratio estimate. Like the BARE, the underlying assumption of homogeneous broad areas is still quite strong.

(iv) regression based models - these allow for the small area predictions to be based on a number of different variables. A multivariate analysis can be applied at an area level (eg LGA), a unit level (eg person or household) or as a combined model that has elements of both. The model form will need to reflect the underlying data (eg continuous, count, or categorical). Models can include a random effect which allows for differences between areas to be included in the model. This can be particularly beneficial when the small areas contain some sample on which the area effect can be based. Models that do not allow for an area effect are known as synthetic models.

Figure 1: Modelling Framework



4. What makes for a successful small area study

It is useful to identify the ingredients for success of any small area study. This will depend on a number of interrelated factors:

User commitment and client interaction - the ability to work closely with users who have a clear idea of how the small area estimates will be used, can prioritise their requirements and can assist in validation. The approaches described in this paper are not appropriate for general purpose statistics production. Rather, they rely on tailoring output to specific needs and uses.

Variable(s) of interest - the variables of interest should be a reasonably common population characteristic. Rare characteristics (less than 10% of the population) are difficult to model and result in less reliable small area estimates. It follows from this that disaggregation (eg by age, sex, type etc) of such variables should be kept to a minimum.

Population size of the small area - when small areas contain some sample, even if inadequate for accurate direct estimation, the modelled estimates will be more reliable. In practice, this ideal may not be achieved because small areas need to be meaningful in the user context. Generally, the accuracy of small area estimates decreases along with the size of the small area. As a working rule, researchers suggest a minimum population size of 20,000. In practice, many small area exercises deal with areas that are smaller than this. In such cases, there needs to be even stronger relationships with auxiliary data to ensure success. Some areas may need to be excluded from the output if they are judged to be too small, as the modelled estimates will be too unreliable.

Auxiliary data - the availability of administrative, census or other survey data with a significant relationship to the variable of interest is crucial. These data needs to exist for each small area and be accurately collected and maintained. Where relationships established in a survey context are to be applied to the auxiliary data, similar scope and definitions are necessary.

5. Assessing the quality of small area estimates

Small area estimates are usually obtained by fitting statistical models to survey data and then applying these models to auxiliary information available for the small area population of interest. Often a number of potential or candidate models are considered involving various combinations of the auxiliary variables. The most reliable of these candidate models is then chosen as the final model, on the basis of:

- plausibility of the model in light of previous studies or accepted wisdom;
- how well the model fits the observed data; and,
- accuracy of the small area estimates predicted from the model.

In light of this, there is a need to examine various quality diagnostics to determine which of the candidate models to use. Having chosen a model, it is then necessary to provide users with an assessment of its quality as well as the quality of the small area estimates produced from it. In doing so, a range of diagnostics are used to assess the accuracy, validity and consistency of the small area estimates. These include:

- a bias test that compares the small area predictions with direct estimates;
- testing whether model assumptions are met and that the model is a good fit;
- checking that small area estimates add to published state or national estimates;
- local knowledge and expert advice on the spread of estimates across small areas; and,
- relative root mean squared errors (RMSE) - in modelling these are analogous to sampling errors calculated for survey estimates.

Although these diagnostics are crucial in terms of assessing the relative performance of competing small area models, they have to be supported by good judgement from practitioners and expert advice from users.

6. Communicating quality to users

Trewin (1999) encouraged national statistical agencies to make greater use of small area estimation methods to generate statistical output. However, in doing so, he emphasised that:

- o the estimates need to be branded differently from other official statistics (the methods and the assumptions should be described in any releases);
- o their validity needs to be assessed to provide user confidence;
- o the underlying models need to be described in terms that users can understand and the validity of the underlying assumptions should be discussed with the key users;
- o their quality should be described in quantitative terms as far as possible; and
- o there should be peer review of the models by an expert as the models are very complex and the choice of methods is considerable.

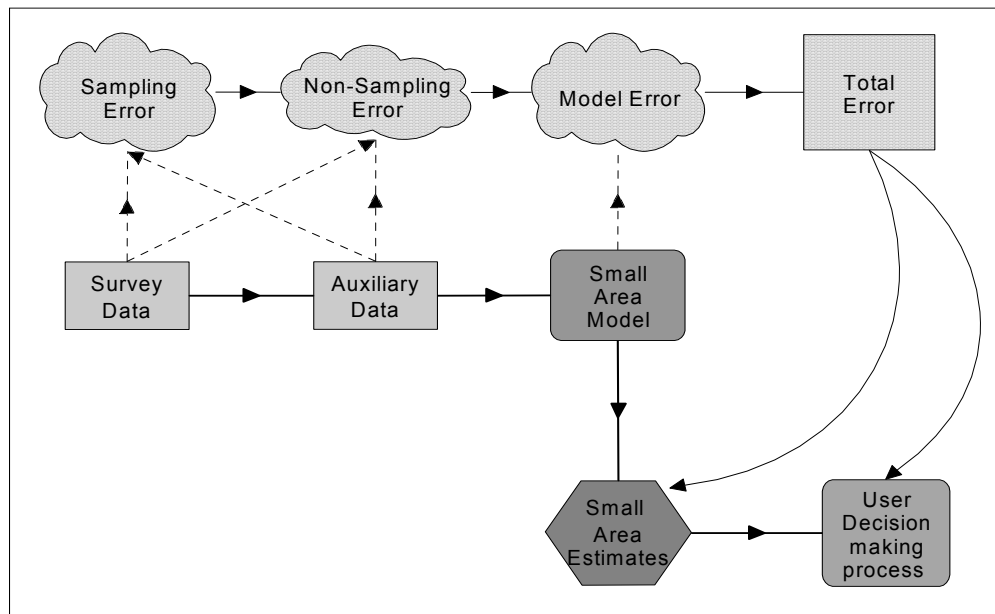
These points raise some important issues for small area practitioners.

a) Quantifying the quality of small area estimates

Understanding quality requirements should be an integral part of assessing user requirements and should be a part of user discussions from the outset. In practically all data contexts, the level of quality required for small area data can be determined from an understanding of how the data are going to be used to solve the decision making problem at hand.

Small area estimates can be subject to a number of different sources of error depending on the way in which they are produced. There are three broad types of error that may impact upon small area estimates: sampling error, non-sampling error and model error as shown in Figure 2. The contribution of these errors to a particular small area application depends upon the small area method and type of data being used.

Figure 2: Sources of Error in Small Area Estimates



b) Documenting results

When small area estimates are released to users, documentation in the form of explanatory notes should accompany these estimates to give users a clear understanding of the concepts and methods used in producing the small area estimates. Such documentation should include :

- a brief introduction of the underlying problem, scope and applicability of the estimates;
- an overview of the small area estimation procedure (the specific model used, variables included, main assumptions, etc);
- a quality declaration, which highlights and emphasises quality issues specific to different sets of small area estimates;
- broad guidelines on how to use the small area output including any limitations; and
- a summary of key issues and recommendations (eg, aggregation of small area estimates, the need for local knowledge, etc.).

The quality declaration may cover, but is not limited to, the following:

- the specific models used to produce each set of small area estimates plus an assessment of their plausibility, validity and goodness of fit;
- how each set of small area estimates performed against specific quality diagnostics; and
- any other quality issues users need to be aware of when applying or interpreting the small area data.

A more detailed technical report may also be produced.

7. Summary and Conclusions

Analytical methods such as small area estimation may be useful for producing statistics by official statistical agencies. Each small area problem needs to be carefully assessed to ensure that the approach taken and techniques applied suit the particular problem at hand.

It is important to work closely with users to choose the geographic areas and the key output variables carefully to ensure the results will be fit for purpose. This can be an iterative process. The availability of auxiliary data with a significant relationship to the variable(s) of interest is crucial to the success of the project.

In addition, it is important to ensure that the methods used, the assumptions underlying the models, and the quality of the outputs are clearly described for users.

References

- Australian Bureau of Statistics (2005), *A Guide to Small Area Estimation*, available online at <http://www.nss.gov.au/nss/home.NSF/pages/Small+Areas+Estimates?OpenDocument>, last accessed on 17/09/07.
- Brackstone, G. J. (2002) "Strategies and Approaches for Small Area Statistics", *Survey Methodology*, 28(2), 117-123.
- Elazar, D. (2004) "Small Area Estimation of Disability in Australia", *Statistics in Transition*, 6(5), 667-684.
- Elazar, D. and Conn, L. (2005) "*Small Area Estimation of Disability in Australia*", Presented to Australian Statistical Conference, 11-16 July 2004, Cairns, Australia (ABS Catalogue No1351.0.55.003).
- McEwin, M. and Elazar, D. "Small Area Estimation in Official Statistics", *Room document at Second Forum for Asia/Pacific Statisticians (APEX 2)*, available online at http://www.unescap.org/stat/apex/2/APEX2_S.3_SAE%20in%20Official%20Statistics_Ver_2.pdf, last accessed on 17/09/07.
- Saei, A and Chambers, R. (2003) *Small Area Estimation: A Review of Methods Based on the Application of Mixed Models*, Southampton Statistical Sciences Research Institute Methodology Working Paper M03/16.
- Trewin, D. (1999), Small Area Statistics Conference, *Survey Statistician*, 41, 8-9