Available online at www.isas.org.in/jisas
JOURNAL OF THE INDIAN SOCIETY OF AGRICULTURAL STATISTICS 66(1) 2012 203-212

# Practical Guidelines for Design and Analysis of Sample Surveys for Small Area Estimation

**Stephen Haslett**
*Institute of Fundamental Sciences – Statistics, Massey University, Palmerston North, New Zealand*
Received 30 June 2011; Revised 12 September 2011; Accepted 13 September 2011

**SUMMARY**

This paper provides practical guidelines for the design and analysis of sample surveys that are to be used for small area estimation using regression type methods. It is based on the author's experience using small area estimation in a range of studies including small area modeling of employment and unemployment, small area estimation of poverty and small domain estimation of ethnicity, in a range of countries including USA, UK, Bangladesh, Philippines, Nepal, Cambodia, and New Zealand, and for feasibility studies in Bhutan and Timor-Leste. The importance of recognising at design stage that one of the uses of the survey data will or may be small area estimation, and identifying all the parameters that will require estimation (including variance components) are discussed, as are issues of clarity of aim, data availability, and model choice at the analysis stage.

*Keywords* : Data cleaning, Feasibility assessment, Modelling, Multiple goals, Poverty mapping, Variance components.

## 1. INTRODUCTION

Small area estimation uses statistical models that "borrow strength" from data that are in some sense related (e.g. by similarity of type or proximity). Such methods improve direct estimates that by definition rely only on data collected within each small area and which may not be sufficiently accurate for reliable use. Small area estimation is a technique, or rather a range of techniques, covering a wide variety of statistical models with an extensive range of data requirements. Small area estimation also has a very wide range of applications, not only geographically, but also in the domains and measures of interest.

This breadth makes both generalisations and guidelines for design of surveys that are to be used for small area estimation more difficult, which is why the limited literature on the topic has tended to focus on particular projects, for example those undertaken by or for government statistical agencies. See for example, Singh *et al.* (1994) which discusses issues and strategies in the context of Statistics Canada's programmes. Marker (2001) considers a slightly wider range of countries of origin (for United States: Current Population Survey, National Health Interview Survey, Survey of Income and Program Participation, National Employer Health Insurance Survey; Europe: Community Household Panel Survey), but again the focus is on major government statistical agencies.

The present paper covers or mentions by implication a wider range of countries: including United States, United Kingdom, Bangladesh, Philippines, Nepal, Cambodia, and New Zealand, and small area estimation feasibility studies in Bhutan and Timor-Leste. However like earlier papers, it is not comprehensive, since in a number of these countries it covers only small area estimation of poverty.

Nevertheless, a broad view is taken of what constitutes small area estimation, with a focus on small area estimation techniques that make use of sample survey data, rather than the methods developed, for

*E-mail address* : s.j.haslett@massey.ac.nz

example in demography, that use only administrative or census data and tend to emphasise population counts. The suggestions and recommendations given in the following sections are primarily for design, implementation and analysis of sample surveys when it is intended (or even possible) that survey data will be used later as part of a small area estimation exercise. While comments are made in later sections on methodological and theoretical matters, much of the material focuses on implementation of small area estimation projects and related non-sampling issues.

The standard statistical reference on small area estimation is Rao (2003), but despite its encyclopaedic quality, there are small area techniques which fall outside of its orbit. One such set with an economic poverty emphasis (e.g. Elbers *et al.* 2003) has more recently been extended to malnutrition measures, such as underweight, stunting and wasting in children (see for example, Jones *et al.* 2006, and the proposal that such methods be examined in the earlier research in Vietnam of Minot 2000 and more explicitly Minot *et al.* 2003).

There is also research in related disciplines beyond econometrics and demography that has lead to other developments. An example is spatial microsimulation, used extensively by geographers and policy planners, and until recently (for details see Haslett *et al.* 2010) not recognised as a small area estimation technique at all. Essentially, spatial microsimulation is a small area estimation technique that combines survey and census data by fitting an implicit model (c.f. Elbers *et al.* 2003 where the model is explicit) and then uses the model to make census based predictions. The model can also be projected in time, to provide long term projections into the future for policy purposes under a range of pre-specified scenarios. Putting aside the complications of predicting into the future, usually from a single time point, a major statistical issue with spatial microsimulation is that, instead of selecting and checking a small area model based on statistical principles, the variables included in the model are chosen on *a priori* grounds (e.g. based on what cross-tabulated census data is readily available). Spatial microsimulation models often involve population counts and estimates of population counts via low dimensional contingency tables, and although more than one pseudo-census could be created to allow standard error estimation usually only one such pseudo-census is created. See Haslett *et al.* (2010) for a more complete analysis. Earlier references on microsimulation techniques include: Orcutt *et al.* (1961), Orcutt *et al.* (1986), Bramley (1992), Hancock and Sutherland (1992), Smart (1996), Bramley and Lancaster (1998), Caldwell *et al.* (1998), Clarke and Keister (1998), Mitton *et al.* (2000), and Ballas *et al.* (2007).

Rao (2003) includes extensive material on linear mixed models and generalized linear mixed models for small area estimation, noting that both can be fitted to sample survey data (depending on the data available) at area or at sampling unit level. Rao also considers models with time series and spatial dependence. There is a strong emphasis on best linear unbiased prediction (BLUP), empirical best linear unbiased prediction (EBLUP), and Bayesian techniques: empirical Bayes (EB) and hierarchical Bayes (HB). Necessarily, the book also contains material on estimating mean square error (MSE) of small area estimates conditional on a given model, and hence on the estimation of the variance components required to estimate the mean square error of small area estimates. Other more recent references also tend to focus on such technical issues, supplemented by additional material, for example on using the maximum likelihood, restricted maximum likelihood and the bootstrap. Recent literature has also extended and further developed the use of generalized linear models in small area estimation.

The literature on survey and questionnaire design and implementation of sample surveys for small area estimation, *per se*, is however limited. Essentially, there are the seminal paper of Singh *et al.* (1994) which has a major component on estimation techniques, and that of Marker (2001). What is notable, in both Singh *et al.* and in Marker's paper, is that the main focus is on extending direct estimators in government surveys a little, to cover situations where the survey cannot be expected to produce sufficiently accurate results, rather than considering small area estimation to be a principal aim when designing surveys. The emphasis in both papers is borrowing strength through the use of estimators based on the sample alone, rather than on estimators (such as those used in poverty estimation and by geographers) which include or even focus on incorporating census and administrative data in the form of predictions for census data based on a model fitted

to a survey (and which require the predictor variables used to be statistically equivalent and available in both census and survey).

There are, consequent on choice of technique, differing data requirements for small area estimation using survey data. The minimum requirement is data from a single survey, but this may be supplemented by a requirement for data from other related surveys, data from administrative or census sources, and/or by repetitions of the survey over time or in other administrative regions. Even the "minimum source" may be unit level data or data aggregated to small area level. Many of the small area techniques make use of administrative or census data in the form of area averages (i.e. contextual effects), possibly at a less aggregated level than small area, and used as explanatory variables in a regression-type model fitting process in which such averages are used at the finest level at which models are structured.

An alternative to using survey data alone is to fit models at (sampling) unit level to survey data using only explanatory variables that are also available in a census (or from an administrative source). Examples include Elbers *et al.* (2003) type techniques, and spatial microsimulation. Given the model fitted to the survey data is sound, that survey and census are contemporaneous, and that the explanatory measures from them are statistically equivalent (i.e. match sufficiently well statistically), small area estimates for variables collected in the survey but not in the census can then be formed by aggregating survey model-based predictions from census or administrative sources at sampling unit level. While estimating standard errors (even if the model is correct) can remain an issue, such methods have the potential to produce small area estimates at a very much finer level than those using survey data alone, so that the comments of Singh *et al.* and of Marker on requirements for small area estimation need some extension. The distinction between these types of small area estimates is explained more fully and discussed in the following sections.

Although he is referring only to their use in regression-type models as contextual variables, in both this situation or in the types of small area models based on census predictions, Kalton's (2003) comments in the preface to Rao (1973, p xviii) apply, "The essence of all small area estimation methods is the use of auxiliary data at the small area level, such as administrative data or data from the last census".

Although the extent of available small area estimation techniques, at first glance, provides a bewildering range of possibilities and complications, for design of and analysis of surveys within this structure there are nevertheless guiding principles.

Early sample survey theory (for example, Hansen *et al.* 1953, Kish 1965, Cochran 1977, Raj 1972) focused on optimal or near optimal survey design that minimises design mean square error for a given cost. The emphasis was usually on estimates for the entire sample, although subpopulation estimates were also considered, albeit for subpopulations with sufficiently large sample sizes that direct estimation is sufficiently accurate. The early focus was on design issues such as stratification, number of strata, allocation of sample to strata, clustering and cluster sizes, and sample size within clusters.

Rao (2003, Section 2.6) discusses the extension of these survey based issues to small area estimation, noting particularly the earlier work of Singh *et al.* (1994) and Marker (2001) already mentioned above. Singh *et al.* (1994) comment on the need to develop an overall strategy and specify salient features of the survey design that have an impact on small area estimation. Marker (2001) focuses on stratification, dual frames and oversampling. Rao (2003) discusses stratification, sample allocation, integration of surveys, dual frame surveys, and repeated surveys. He gives a number of examples for dual frames, integration of information from different surveys that have the same questions, and repetition of surveys, which are all methods that emphasise combining data from different sources at the model fitting stage. He does not however discuss use of census or administrative data sources for small area estimation using predictions from census data based on a survey model (such as used by econometricians and geographers).

More recently however, Molina and Rao (2010) outline such a technique by considering best prediction at small area level which conditions estimation on the sample. Although their technique may seem very different, when there is no sample in a small area (as often occurs in poverty mapping in developing countries) Molina and Rao's estimator essentially reduces to a synthetic estimator which can be

reformulated as a variant of Elbers *et al.* (2003). Details of this required reformulation have been discussed in Haslett *et al.* (2010) and Haslett and Jones (2010) and have been applied to poverty estimation in Bangladesh, Philippines, Nepal and the Cambodia over the period since 2003. The essential differences between Molina and Rao (2010) and Elbers *et al.* (2003) are that in Molina and Rao the predictor of the small area error component in the model is included in the small area estimate (although this is not possible if the small area is not sampled), that the regression part of the model is re-estimated at each iteration to better incorporate the survey design (since Elbers *et al.* and their PovMap software use ordinary least squares, so that their estimated variance of the parameter estimates is downwardly biased), and that Molina and Rao use the known survey values for the (usually very small) percentage of the population that have been sampled (which usually has little effect, and is not possible if survey and census respondents do not have linked identification numbers).

The links between Molina and Rao (2010) and the improved variants of Elbers *et al.* (2003) remain an active area of research, but the following comments may be a useful guide.

Molina and Rao (2010) apply their technique to Spanish data where there is sample in every small area, and where the small areas are much larger in terms of population size (and poverty levels rather lower) than is common for small area estimation in third world countries. In these circumstances, with the possible exception of various distributional assumptions they make, their methods are clearly superior to those of Elbers *et al.* and their variants, because they incorporate the small area error component, and in addition superior to the original Elbers *et al.* technique because the additional variation from the complex design (which is the larger part of the mean square error of the small area estimates at such higher levels of aggregation) is unbiasedly estimated. Interestingly, in third world applications the relative contribution of this error from the regression fit is usually small (as it will often be when a good percentage of small areas contain no sample), so that for estimation of mean square error it is estimation of variance components that is the principal issue. In such circumstances, Molina and Rao (2010) and the variants of Elbers *et al.* (2003) discussed in Haslett *et al.* (2010) and Haslett and Jones (2011) should give very similar estimates and estimates of mean square error. In summary, at the level of aggregation used in many small area poverty studies in developing countries, Molina and Rao (2011) is almost identical to the methods outlined in Haslett *et al.* (2010) and Haslett and Jones (2010).

The comments in following two sections extend these ideas to situations where there is supplementary data used for prediction, as well as discussing a range of issues that apply even when only data from a single survey is available.

## 2. MODEL SPECIFICATION – GENERAL CONTEXT

Although fully specifying all the types of linear and generalized linear models that are used in small area estimation goes beyond the intent of this paper, it is nevertheless useful to characterise small area models used with survey data in a general sense. The intention is not to be definitive but to characterise sufficiently those aspects of small area models which have important implications for survey design and analysis.

Small area estimation using survey data only, and small area estimation using census or administrative data for unit level predictions, have a number of aspects in common. This broad classification may include small area estimation techniques which incorporate census or administrative data into the modelling stage as contextual effects only within the first category, since the estimation phase does not use census data at unit record level. Even though the final stages of estimation differ between these two categories of technique, they do however have major underlying features in common. The principal similarities are that at the first stage both types require statistical models to be fitted to survey data, that the models are often regression-type models, and that both types may or may not involve contextual effects.

The papers of Singh *et al.* (1994) and Marker (2001), and the book of Rao (2003), all focus on (and are limited to) survey design and other requirements for this modelling aspect of small area estimation when providing their design and analytic guidelines. Their focus is proper, because the econometric and geographic small area techniques that use unit level census data are outside their orbit.

It is useful at this stage to consider aims, survey and questionnaire design, data issues, and statistical modelling in sequence.

## 3. ASPECTS OF SURVEY DESIGN, IMPLEMENTATION AND ANALYSIS RELEVANT TO SMALL AREA ESTIMATION

### 3.1 Aims and Objectives

For small area estimation, as with other types of research, it is necessary to be clear what the research objectives are in detail, and the range and quality of the data sources available. More complicated and detailed aims are almost inevitably more difficult to achieve than simpler ones, so there is merit in simplicity. In a technical sense, aims for small area estimation may be different or need to be modified depending on whether the relevant survey or surveys have already been undertaken, or whether instead the survey design can be amended to better incorporate or optimise small area estimation requirements. The survey design structure and the questionnaire content also impact on what is achievable (for further details see below) so that the advantages of designing a survey with small area estimation in mind, rather than having small area estimation as an afterthought, can be considerable.

### 3.2 Survey and Questionnaire Design

Questionnaire design can have considerable influence on small area estimation models. Problems can be as simple as those imposed by restricted questionnaire content, because even if small area estimation has been noted as a possibility at design stage, the range of questions may be too limited to allow development of sound small area models.

This issue appears in another guise when small area models using unit level census data as predictors are used (as in the aid industry standard method for small area estimation of poverty) since then (excepting any contextual variables) the explanatory variables (i.e. those used to predict the variable of interest) need to be statistically equivalent in survey and census. This is certainly not possible, *a priori*, if candidate explanatory variables are not in both survey and census. Complications ensue when the survey is nutrition or health based and the census (as is usually the case internationally), in the main, collects economic information. For example, the Cambodia Anthropometric Survey 2008 (CAS2008) asks detailed questions about mothers and children under the age of five years, but there are no questions about other family members (including household type) or household size. On the other hand there are indicators of child malnutrition, such as those about child diarrhoea, that are asked in CAS2008 but not in the General Population Census of Cambodia 2008.

There may also be benefit in asking a more detailed question in the survey than is necessary for the survey itself, as additional categories or further detail may greatly facilitate modelling if the variables derivative from such questions are or may be important predictors.

When fitting small area models using unit level census data as predictors, it is also very important if data collection from survey and census have not been completed, to ask the same questions using the same categories in survey and census. Too often, even within a single agency, survey and census questionnaires are developed by different personnel without adequate consultation, so that even simple questions such as type of roof or walls, or more complicated ones about educational attainment are asked in different ways and/or using different categories.

There are also important survey design considerations for small area estimation, and these may or may not fit with optimality of design for direct estimators from the survey itself. This conflict can occur however even if only direct estimators for subpopulations are required in a standard design, because then the estimator that is optimal overall may give poor subpopulation estimates. An example from New Zealand is surveys which are required to provide estimates for Maori and/or other ethnic groups and where, since Maori are around 15% of the population, various forms of oversampling or other design adjustments are commonly used to improve subpopulation estimates. When direct area-based (rather than domain-based) estimates are required, this focus on subpopulation estimates leads to samples that are similar in size for different areas or groupings, even if the areas or groupings are themselves of different sizes.

In contrast, samples for national estimates only (especially in a business context where techniques such as probability proportional to size are used) do not

necessarily provide good regional estimates (particularly where size of business varies by location and/or the same type of business tends to be co-located). Focusing, for the purpose of this discussion, on small area (rather than small domain) estimation, a similar situation exists. Even for social surveys, where the geographic distribution of the sample tends to be more even, design issues can be important. Examples include choice of clusters. Fewer, larger clusters may be quite adequate for national and regional estimates, but may lead to complications for small area estimation. A smaller number of clusters can limit geographical spread and this leads to stronger reliance on the small area model especially in small areas where no sample is taken.

There is a balancing consideration however. Many small area models contain random effects at small area, cluster (i.e. primary sampling unit - psu), household, or even individual level (e.g. children within household for stunting, underweight and wasting). To estimate standard errors for small area estimates requires estimates of these variance components and, to estimate the variance components sufficiently accurately, the sample needs to contain at least some small areas with sufficient sampled clusters, as well as sampled clusters with sufficient sampled households, etc. Estimation of random effects at the higher levels (e.g. at small area level from clusters within small areas) is most important (since it usually has the largest effect on estimated standard errors of the small area estimates), and this requirement should be considered carefully at design stage. What should be avoided, if possible (and it not always is), is having most small areas containing only one sampled cluster, since then there is limited information that can be derived from the sample about the corresponding variance component. This complication can be relatively common in small area models where census data is used at unit level, small areas are consequently small, and the survey (while sound for direct estimates at a more aggregated level) has not been designed with small area estimation in mind. For example, for the Nepal Living Standards Survey (NLSS), 2003/04 the coverage is adequate at regional level, but three of the 75 districts are not sampled and at least one of the others has only one sampled primary sampling unit. At the finer ilaka level of which there are approximately 1000 in Nepal, very few contain more than one sampled psu, and the majority contain no sampled psu at all. Of course, in this situation of few small areas with more than one sampled cluster, if the cluster level effect is dropped from the model and the area level effect retained, an upper bound for the variance can be obtained since cluster level variation is then included in the area level effect estimates, although even then care is still required because estimation of the area level variance component can still be difficult.

### 3.3 Data Requirements

Data availability can be limited for surveys that have been designed and undertaken before small area estimation is considered. In the extreme, when survey data release requires permission, no data beyond questionnaire structure may be available and (as was the case for the small area estimation of poverty feasibility study for the World Food Programme in Timor-Leste, based on the Timor-Leste Survey of Living Standards 2007) nothing is possible at feasibility stage beyond assessing the range of variables collected in the survey and preliminary matching of survey and census variables based on the questionnaires. An even more extreme example is the Global Entrepreneurship Monitor survey in Tonga, outlined in Frederick *et al.* (2011), where an additional journey from New Zealand to Tonga was required to get design information for the national survey, because the necessary formal letters had not been sent in time to the Tongan Statistics Department and to other government officials.

Where data is available it needs to be assessed for quality. Non-response can be an issue, particularly where (even if the overall response rate is good) there are pockets of non-response for subpopulations and/or areas, and the relevant subpopulations are ghettoised. This occurs for example in a range of national surveys in New Zealand, particularly for with young Maori and Polynesian men. Interestingly, non-response is generally not a problem in poverty surveys in developing countries with a reasonably sound statistical system since response rates in excess of 98% are the norm (e.g. Bangladesh, Bhutan, Cambodia, Nepal, Philippines). Even the veracity of responses may be better, although not necessarily for sensitive questions. For example, in both Nepal National Population Census 2001 and the Nepal Demographic and Health Survey 2001 there is a question about the number of deaths of children in sampled families. Although the questions are very

similar, the reported death rates are very much lower in the census where the questionnaire is much shorter and the question asked in the context of economic information rather than as part of an extended set of questions on family health.

Data preparation and cleaning for small area estimation is vitally important, even when survey data has been cleaned and official estimates produced at national at regional level. The additional care is required because the survey is being considered at a finer level, and data related issues that may not be important for the accuracy descriptive statistics (e.g. means, totals, tables of counts) at a more aggregated level can nevertheless become very important when making use of the survey for analytic purposes at small area level. There are parameters to be estimated for models, the data are being considered in a multivariate context, and outliers (especially in unit record data) which have little influence on overall means, can have considerable influence on regression coefficient and variance component estimates. Robust methods have been developed to deal with such problems, but considerable care is required in their use, especially when estimating variance components, because observations that are down-weighted or ignored when robust estimates are used, rather than needing to be down-weighted, may be better used as important indicators or diagnostics for model fit. Robust methods can also severely bias variance component estimates, resulting in underestimates.

Data coding particularly needs watching, because misclassification can have marked effect on model fits. If the small area estimation is being carried out by non-locals, there are an additional range of issues that need care. Questionnaires in local languages may or may not be fully reflected in their translations into English or into any other foreign language. A number of the differences may simply reflect local culture. For example, words that are the same in English may be different in local language, and vice versa. This is particularly an issue when survey and census questions, or questionnaires from more than one survey, need to be compared. Even apparently objective questions which seem not to involve opinion (e.g. type of roof or wall) warrant clarification. Fieldwork procedures also justify discussion (e.g. what was the classification of wall type when some walls in a sampled household were wood and others concrete?). Discussing interpretation of results from regression modelling with of local statisticians and official statistics staff are also highly recommended, as well as providing a training opportunity - for all participants.

One final point: where both census and survey data are being used for small area modelling, (even if census data only provides aggregated contextual variables) understanding in detail the connections between area codes used in the survey design and those used in the census is imperative. This linking can be a very time consuming job indeed where a previous census has been used as the frame for the survey, and data from a new census is later available. Again, local knowledge is imperative if this linking of area codes is to be done well, particularly if it involves use of local names that after translation contain spelling variants.

### 3.4 Modelling

There are a number of caveats when modelling. Some are obvious since they connect back to data quality, such as checking model residuals - especially where they are used to estimate variance components. There can be a pressure to look for models with high percentage of variance explained ($R^2$, or adjusted $R^2$) but if this is achieved by dividing survey data into pieces (e.g. strata), gains can be illusionary as they tend to be highly sample dependent and not to reflect population characteristics. Such overfitting also ignores the fact that $R^2$ may not be the most important diagnostic for a small area model. Where variance components are necessary, sound estimation of these (which is of course more difficult with subsetted data) is often more important. Models with a greater proportion of overall variation at lower level (e.g. individual or household level rather than psu or area level) are preferable as the variation at lower levels has much less effect of small area estimate standard errors. It is also wise to be aware that standard errors estimated from small area estimation models are almost invariably conditional on the model being correct, so that extensive model testing is not only warranted but necessary. For further details and discussion of these issues in the context of small area estimation of poverty, see Haslett and Jones (2010) and Haslett *et al.* (2010).

For social surveys, fitted statistical models will test, and likely incorporate, stratum and cluster effects, as well as exogenous variables such as household type. Since interactions between stratum and exogenous variables are reasonably common in models, having a

sufficient sample size within each stratum to properly do the fitting is important. Unless designs are of the two units per stratum type, this is not usually an issue (except if data are subset before model fitting, which is the non-recommended procedure discussed in the previous paragraph). What tends to be crucial is the selection method that has been used for clusters and number of clusters sampled per stratum. Clusters are often selected with probability proportional to size (pps). While then sampling a fixed number of households per sampled cluster then gives a self-weighting sample of households, at least within strata, different cluster characteristics (e.g. their size) can cause problems when models need to incorporate and estimate cluster-level variances. As noted above, the estimation of these variances is further complicated if the sampling scheme only samples a few clusters within strata, especially if the differences in cluster variance between strata need testing as part of the model and/or there are underlying differences related to size between clusters. For further discussion at a more technical level, see Haslett *et al.* (2010). Robust estimates of means square error are possible (with the proviso mentioned earlier that down-weighted observations should first be checked as indicators of model unsuitability, and that variance components may be downwardly biased if robust methods are used uncritically).

In business surveys the situation is different, although for area based survey designs for business the conclusions are rather similar. However, for samples selected from a list, for example by probability proportional to size, because differences in size of businesses can be marked, the sample needs to be strongly weighted (inversely proportional to size) to produce unbiased estimates. The strong weighting can complicate model fitting to the survey data for small area estimation. It can also mean that subgroups of the population that are important to the modelling but not to the overall national figures can be under-represented if the primary consideration at design stage was not small area estimation. An alternative, if they are known, is to include the size measures explicitly as an explanatory variable in the model.

The general conclusion is that for modelling, designs which spread their choice of sample across the population are preferable. The situation has parallels to the trade-off necessary when direct subpopulation estimates as well as national totals or averages are required from a sample. Designing for sound direct subpopulation estimates and estimates based on statistical models have similar requirements and, because it is a trade-off, meeting these requirements will always lower the accuracy of national-level estimates to some extent relative those possible for a more complicated design.

It is consequently important to decide the design of a survey based on the entire range of uses to which it will be put, and not to expect good statistical modelling for small area estimation will be possible simply as an add-on to a design constructed to meet other optimality criteria.

## 4. CONCLUSION AND GUIDELINES

The conclusion of this paper is provided below in the form of a set of guidelines. These outline and summarise what best to do when considering design, implementation and analysis of sample surveys for small area estimation:

1. Consider carefully what the aims of the small area estimation research are, in detail.
2. Be aware that more complicated and detailed aims will be more difficult to achieve than simpler less detailed ones.
3. Plan, including specification of resources available and a timeline.
4. Ensure necessary permissions have been gained in writing for access to any required survey and questionnaire design information, and existing survey and census data, where relevant.
5. If it is possible or necessary to design the survey around the aims of the small area estimation project (rather than using only pre-existing sample survey and other data collections) then:
   (a) Consider the questionnaire design and, if census data is also to be used in modelling (either at unit record level or aggregated as contextual variables), or if data from additional surveys is to be used, make sure the question structure for candidate variables for the modelling is identical for all data sources, or (where variables are categorical) that they can be collapsed to equivalent categories.
   (b) Consider how the survey design can best incorporate or optimise small area estimation requirements. For example, ensure that primary sampling units are spread across areas so that as many small areas as possible contain at least some sample, and that the

sample design ensures that all required variance components can be accurately estimated.

(c) If census based prediction of survey models is to be used, to the extent possible ensure candidate explanatory (i.e. exogenous) variables are identically defined and in both survey and census.

(d) Ensure the survey design uses the same area codes as other data sources, or that there is a method of linking disparate area codes at all administrative levels.

(e) Be particularly careful with the design of business surveys which have a wide range of sampling unit selection probabilities, to ensure the information required for small area estimation from units with low selection probability can still be estimated from the survey data.

6. Fully assess and the range and quality of the data sources available, including resolving any non-response issues for subpopulations. Be aware of the effect of context on responses for sensitive questions, and of any possible language translation related issues.
7. Make good use of local knowledge.
8. Establish whether multiple data sources are compatible (e.g. same or similar time period, same definitions of variables, level of disaggregation, area coding), and sufficient. Assess what proportion of the data required is available from each of these sources, both in terms of the population under study and the variables available.
9. Consider carefully the structure of any dataset you are trying to create (e.g. if using unit record census data for prediction), in terms of variables and level and numbers of observations, and how it will be aggregated as required to form small area estimates.
10. Reassess what proportion of the data is missing on key variables, and establish the pattern of missingness (e.g. random from a sample survey; missing due to purposive administrative procedures).
11. Consider the range of types of statistical models (e.g. linear models with or without random effects, generalized linear models, generalized linear mixed models) that seem suitable for modelling, and make a preliminary choice.
12. List the candidate variables available to predict the key variables.
13. Fit preliminary small area models and then recheck, and if necessary re-clean, all data to be used.
14. Avoid subsetting survey data and fitting separate models to the subsets – fit models with specified interactions with area (e.g. region or stratum) instead. Assess how much of the variation the preliminary models explain, and whether any required variance components can be accurately estimated.
15. If time series projections are required, or models of future scenarios are needed, consider what models are suitable for projection, whether required variables are available, and whether a statistical time series model can be built from the available data.
16. For census projection models, even for a single time period, consider what prediction errors are likely to be at unit (e.g. individual) level.
17. For census projection models, think about what aggregations of data you will need to use to get sufficient accuracy to be useful.
18. Go back to your aims, and ask whether the project is feasible given the likely accuracy.
19. If the project remains feasible, continue fitting and checking small area models. For census projection models based on a sample survey model, this includes setting up initial simulations for assessing small area standard errors.
20. Run any such simulations many times, for every scenario if relevant. (This is a form of multiple imputation.) Use your multiple simulation results to assess accuracy for key variables (noting that accuracy estimates are conditional on the simulation model being used).
21. Recheck data for errors, and reassess what types of statistical models are most appropriate.
22. Begin a more thorough search for models that fit well, checking them and their component parts for statistical significance. Loop back to step 16, or if necessary step 11, as many times as necessary. Continue this process for as long as the timeline and resources allow.
23. Reassess whether you can meet your aims.
24. Consider carefully at what level of aggregation (e.g. at what area size) you can get small area estimates that are sufficiently accurate for purpose.
25. Produce results at this level or a more aggregated level, and relate the results at this level back to your aims.

## REFERENCES


Ballas, D., Kingston, R., Stillwell, J., and Jin, J. (2007). Building a spatial microsimulation-based planning support system for local policy making. *Environ. Plann.*, **A39(10)**, 2482-2499.

Bramley, G. (1992). Homeownership affordability in England. *Housing Policy Debate*, **3, 3**, 143-182.

Bramley, G. and Lancaster, S. (1998). Modelling local and small-area income distributions in Scotland. *Environ. Plann.*, **C16**, 681-706.

Bramley, G. and Smart, G. (1996). Modelling local income distributions in Britain. *Regional Studies*, **30(3)**, 239-255.

Caldwell, S.B., Clarke, G.P. and Keister, L.A. (1998). Modelling regional changes, in US household income and wealth: A research agenda. *Environ. Plann., C: Government and Policy*, **16**, 707-722.

Cochran, W.G. (1977). *Sampling Techniques.* 3rd edition, John Wiley and Sons.

Elbers, C., Lanjouw, J.O., and Peter F., Lanjouw, P.F. (2003). Micro-level estimation of poverty and inequality. *Econometrica,* **71(1)**, 355-364.

Frederick, H., Haslett, S., Wolfgramm, T. and Finau, A. (2011). Measuring entrepreneurial activity in low-teledensity countries. *Int. J. Busi. Global.,* **6, 3/4**, 251-291.

Hancock, R. and Sutherland, H. (eds.) (1992). *Microsimulation Models for Public Policy Analysis: New Frontiers.* Suntory-Toyota International Centre for Economics and Related Disciplines – LSE, London.

Hansen, M.H., Hurwitz, W.N., and Madow, W.G. (1953). *Sample Survey Methods and Theory*. **I & II**, John Wiley and Sons.

Haslett, S., Isidro, M. and Jones, G. (2010). Comparison of survey regression techniques in the context of small area estimation of poverty. *Survey Methodology*, **36(2)**, 157-170.

Haslett, S. and Jones, G. (2010). Small area estimation of poverty: the aid industry standard and its alternatives. *Australian and New Zealand J. Stat.,* **52(4)**, 341-362.

Haslett, S., Jones, G., Noble, A. and Ballas, D. (2010). More for less? Using existing statistical modelling to combine existing data sources to produce sounder, more detailed, and less expensive Official Statistics. *Official Statist., Report Series*, **3**, ISSN 1177-5017, 2010, 75 pages. http://www.statisphere.govt.nz/official-statistics-research/series/2010/page1.aspx

Jones, G., Haslett, S. and Parajuli, D. (2006). *Small area estimation of poverty, caloric intake and malnutrition in Nepal.* Published by the World Food Programme, the World Bank, and the Nepal Central Bureau of Statistics, 218 pages, ISBN 999337018-5.

Kalton, G. (2003). Foreword to Rao, J.N.K., *Small Area Estimation.* xvii - xix, John Wiley and Sons.

Kish, L. (1965). *Survey Sampling*. John Wiley and Sons.

Marker, D.A. (2001). Producing small area estimates from national surveys: Methods for minimizing use of indirect estimators. *Survey Methodology*, **27(2)**, 183-188.

Minot, N.W. (2000). Generating Disaggregated Poverty Maps: An Application to Vietnam. *World Development,* **28(2)**, 319-331.

Minot, N.W., Baulch, B. and Epprecht, M. (2003). *Poverty Mapping and Inequality in Vietnam: Spatial Patterns and Geographical Determinants*. International Food Policy Research Institute. Washington, D.C. http://www.isgmard.org.vn/information%20service/report/General/Poverty%20Mapping%20Final%20Report-e.pdf

Mitton, L., Sutherland, H. and Weeks, M. (eds.) (2000). *Microsimulation Modelling for Policy Analysis: Challenges and Innovations.* Cambridge University Press, Cambridge.

Molina, I. and Rao, J.N.K. (2010). Small area estimation of poverty indicators. *Canad. J. Statist.*, **38(3)**, 369-385.

Orcutt, G.H., Greenberger, M., Korbel, J. and Rivlin, A. (1961). *Microanalysis of Socioeconomic Systems: A Simulation Study*. Harper and Row, New York.

Orcutt, G.H., Mertz, J. and Quinke, H. (eds.) (1986). *Microanalytic Simulation Models to Support Social and Financial Policy.* North-Holland, Amsterdam.

Rao, J.N.K. (2003). *Small Area Estimation.* John Wiley and Sons.

Raj, D. (1972). *The Design of Sample Surveys.* McGraw-Hill.

Singh, M.P., Gambino, J. and Mantel, H.J. (1994). Issues and strategies for small area data. *Survey Methodology*, **20(1)**, 3-22.