1

*INTRODUCTION

1.1 WHAT IS A SMALL AREA?

Sample surveys have long been recognized as cost-effective means of obtaining information on wide-ranging topics of interest at frequent intervals over time. They are widely used in practice to provide estimates not only for the total population of interest but also for a variety of subpopulations (domains). Domains may be defined by geographic areas or socio-demographic groups or other subpopulations. Examples of a geographic domain (area) include a state/province, county, municipality, school district, unemployment insurance (UI) region, metropolitan area, and health service area. On the other hand, a socio-demographic domain may refer to a specific age-sex-race group within a large geographic area. An example of "other domains" is the set of business firms belonging to a census division by industry group.

In the context of sample surveys, we refer to a domain estimator as "direct" if it is based only on the domain-specific sample data. A direct estimator may also use the known auxiliary information, such as the total of an auxiliary variable, x, related to the variable of interest, y. A direct estimator is typically "design based," but it can also be motivated by and justified under models (see Section 2.1). Design-based estimators make use of survey weights, and the associated inferences are based on the probability distribution induced by the sampling design with the population values held fixed (see Chapter 2). "Model-assisted" direct estimators that make use of "working" models are also design based, aiming at making the inferences "robust" to possible model misspecification (see Chapter 2).

Small Area Estimation, Second Edition. J. N. K. Rao and Isabel Molina.

A domain (area) is regarded as large (or major) if the domain-specific sample is large enough to yield "direct estimates" of adequate precision. A domain is regarded as "small" if the domain-specific sample is not large enough to support direct estimates of adequate precision. Some other terms used to denote a domain with small sample size include "local area," "subdomain," "small subgroup," "subprovince," and "minor domain." In some applications, many domains of interest (such as counties) may have zero sample size.

In this text, we generally use the term "small area" to denote any domain for which direct estimates of adequate precision cannot be produced. Typically, domain sample size tends to increase with the population size of the domain, but this is not always the case. Sometimes, the sampling fraction is made larger than the average fraction in small domains in order to increase the domain sample sizes and thereby increase the precision of domain estimates. Such oversampling was, for example, used in the US Third Health and Nutrition Examination Survey (NHANES III) for certain domains in the cross-classification of sex, race/ethnicity, and age, in order that direct estimates of acceptable precision could be produced for those domains. This oversampling led to a greater concentration of the sample in certain states (e.g., California and Texas) than normal, and thereby exacerbated the common problem in national surveys that sample sizes in many states are small (or even zero). Thus, while direct estimates may be used to estimate characteristics of demographic domains with NHANES III, they cannot be used to estimate characteristics of many states. States may therefore be regarded as small areas in this survey. Even when a survey has large enough state sample sizes to support the production of direct estimates for the total state populations, these sample sizes may well not be large enough to support direct estimates for subgroups of the state populations, such as school-age children or persons in poverty. Due to cost considerations, it is often not possible to have a large enough overall sample size to support reliable direct estimates for all domains. Furthermore, in practice, it is not possible to anticipate all uses of the survey data, and "the client will always require more than is specified at the design stage" (Fuller 1999, p. 344).

In making estimates for small areas with adequate level of precision, it is often necessary to use "indirect" estimators that "borrow strength" by using values of the variable of interest, y, from related areas and/or time periods and thus increase the "effective" sample size. These values are brought into the estimation process through a model (either implicit or explicit) that provides a link to related areas and/or time periods through the use of supplementary information related to y, such as recent census counts and current administrative records. Three types of indirect estimators can be identified (Schaible 1996, Chapter 1): "domain indirect," "time indirect," and "domain and time indirect." A domain indirect estimator makes use of y-values from another domain but not from another time period. A time indirect estimator uses y-values from another time period for the domain of interest but not from another domain. On the other hand, a domain and time indirect estimator uses y-values from another domain as well as from another time period. Some other terms used to denote an indirect estimator include "non-traditional," "small area," "model dependent," and "synthetic."

Availability of good auxiliary data and determination of suitable linking models are crucial to the formation of indirect estimators. As noted by Schaible (1996, Chapter 10), expanded access to auxiliary information through coordination and cooperation among different agencies is needed.

1.2 DEMAND FOR SMALL AREA STATISTICS

Historically, small area statistics have long been used. For example, such statistics existed in eleventh-century England and seventeenth-century Canada based on either census or administrative records (Brackstone 1987). Demographers have long been using a variety of indirect methods for small area estimation (SAE) of population and other characteristics of interest in postcensal years. Typically, sampling is not involved in the traditional demographic methods (see Chapter 3 of Rao 2003a).

In recent years, the demand for small area statistics has greatly increased worldwide. This is due, among other things, to their growing use in formulating policies and programs, in the allocation of government funds and in regional planning. Legislative acts by national governments have increasingly created a need for small area statistics, and this trend has accelerated in recent years. Demand from the private sector has also increased significantly because business decisions, particularly those related to small businesses, rely heavily on the local socio-economic, environmental, and other conditions. Schaible (1996) provides an excellent account of the use of traditional and model-based indirect estimators in US Federal Statistical Programs.

SAE is of particular interest for the economies in transition in central and eastern European countries and the former Soviet Union countries. In the 1990s, these countries have moved away from centralized decision making. As a result, sample surveys are now used to produce estimates for large areas as well as small areas. Prompted by the demand for small area statistics, an International Scientific Conference on Small Area Statistics and Survey Designs was held in Warsaw, Poland, in 1992 and an International Satellite Conference on SAE was held in Riga, Latvia, in 1999 to disseminate knowledge on SAE (see Kalton, Kordos, and Platek (1993) and IASS Satellite Conference (1999) for the published conference proceedings).

Some other proceedings of conferences on SAE include National Institute on Drug Abuse (1979), Platek and Singh (1986), and Platek, Rao, Särndal, and Singh (1987). Rapid growth in SAE research in recent years, both theoretical and applied, led to a series of international conferences starting in 2005: Jyvaskyla (Finland, 2005), Pisa (Italy, 2007), Elche (Spain, 2009), Trier (Germany, 2011), Bangkok (Thailand, 2013), and Poznan (Poland, 2014). Three European projects dealing with SAE, namely EURAREA, SAMPLE and AMELI, have been funded by the European Commission. Many research institutions and National Statistical Offices spread across Europe have participated in these projects. Centers for SAE research have been established in the Statistical Office in Poznan (Poland) and in the Statistical Research Division of the US Census Bureau.

Review papers on SAE include Rao (1986, 1999, 2001b, 2003b, 2005, 2008), Chaudhuri (1994), Ghosh and Rao (1994), Marker (1999), Pfeffermann (2002,

2013), Jiang and Lahiri (2006), Datta (2009), and Lehtonen and Veijanen (2009). Text books on SAE have also appeared (Mukhopadhyay 1998, Rao 2003a, Longford 2005, Chaudhuri 2012). Good accounts of SAE theory are also given in the books by Fuller (2009) and Chambers and Clark (2012).

1.3 TRADITIONAL INDIRECT ESTIMATORS

Traditional indirect estimators, based on implicit linking models, include synthetic and composite estimators (Chapter 3). These estimators are generally design based, and their design variances (i.e., variances with respect to the probability distribution induced by the sampling design) are usually small relative to the design variances of direct estimators. However, the indirect estimators will be generally design biased, and the design bias will not decrease as the overall sample size increases. If the implicit linking model is approximately true, then the design bias is likely to be small, leading to significantly smaller design mean-squared error (MSE) compared to the MSE of a direct estimator. Reduction in MSE is the main reason for using indirect estimators.

1.4 SMALL AREA MODELS

Explicit linking models with random area-specific effects accounting for the between-area variation that is not explained by auxiliary variables will be called "small area models" (Chapter 4). Indirect estimators based on small area models will be called "model-based estimators." We classify small area models into two broad types. (i) Aggregate (or area) level models are the models that relate small area direct estimators to area-specific covariates. Such models are necessary if unit (or element) level data are not available. (ii) Unit level models are the models that relate the unit values of a study variable to unit-specific covariates. A basic area level model and a basic unit level model are introduced in Sections 4.2 and 4.3, respectively. Various extensions of the basic area level and unit level models are outlined in Sections 4.4 and 4.5, respectively. Sections 4.2–4.5 are relevant for continuous responses y and may be regarded as special cases of a general linear mixed model (Section 5.2). However, for binary or count variables y, generalized linear mixed models (GLMMs) are often used (Section 4.6): in particular, logistic linear mixed models for the binary case and loglinear mixed models for the count case.

A critical assumption for the unit level models is that the sample values within an area obey the assumed population model, that is, sample selection bias is absent (see Section 4.3). For area level models, we assume the absence of informative sampling of the areas in situations where only some of the areas are selected to the sample, that is, the sample area values (the direct estimates) obey the assumed population model.

Inferences from model-based estimators refer to the distribution implied by the assumed model. Model selection and validation, therefore, play a vital role in model-based estimation. If the assumed models do not provide a good fit to the

data, the model-based estimators will be model biased which, in turn, can lead to erroneous inferences. Several methods of model selection and validation are presented throughout the book. It is also useful to conduct external evaluations by comparing indirect estimates (both traditional and model-based) to more reliable estimates or census values based on past data (see Examples 6.1.1 and 6.1.2 for both internal and external evaluations).

1.5 MODEL-BASED ESTIMATION

It is now generally accepted that, when indirect estimators are to be used, they should be based on explicit small area models. Such models define the way that the related data are incorporated in the estimation process. The model-based approach to SAE offers several advantages: (i) "Optimal" estimators can be derived under the assumed model. (ii) Area-specific measures of variability can be associated with each estimator unlike global measures (averaged over small areas) often used with traditional indirect estimators. (iii) Models can be validated from the sample data. (iv) A variety of models can be entertained depending on the nature of the response variables and the complexity of data structures (such as spatial dependence and time series structures).

In this text, we focus on empirical best linear unbiased prediction (EBLUP) (Chapters 5–8), parametric empirical Bayes (EB) (Chapter 9), and parametric hierarchical Bayes (HB) estimators (Chapter 10) derived from small area models. For the HB method, a further assumption on the prior distribution of model parameters is also needed. EBLUP is designed for estimating linear small area characteristics under linear mixed models, whereas EB and HB are more generally applicable.

The EBLUP method for general linear mixed models has been extensively used in animal breeding and other applications to estimate realized values of linear combinations of fixed and random effects. An EBLUP estimator is obtained in two steps: (i) The best linear unbiased predictor (BLUP), which minimizes the model MSE in the class of linear model unbiased estimators of the quantity of interest is first obtained. It depends on the variances (and covariances) of random effects in the model. (ii) An EBLUP estimator is obtained from the BLUP by substituting suitable estimators of the variance and covariance parameters. Chapter 5 presents some unified theory of the EBLUP method for the general linear mixed model, which covers many specific small area models considered in the literature (Chapters 6 and 8). Estimation of model MSE of EBLUP estimators is studied in detail in Chapters 6–8. Illustration of methods using specific R software for SAE is also provided.

Under squared error loss, the best predictor (BP) of a (random) small area quantity of interest such as mean, proportion, or more complex parameter is the conditional expectation of the quantity given the data and the model parameters. Distributional assumptions are needed for calculating the BP. The empirical BP (or EB) estimator is obtained from BP by substituting suitable estimators of model parameters (Chapter 9). On the other hand, the HB estimator under squared error loss is obtained by integrating the BP with respect to the (Bayes) posterior distribution

derived from an assumed prior distribution of model parameters. The HB estimator is equal to the posterior mean of the estimand, where the expectation is with respect to the posterior distribution of the quantity of interest given the data. The HB method uses the posterior variance as a measure of uncertainty associated with the HB estimator. Posterior (or credible) intervals for the quantity of interest can also be constructed from the posterior distribution of the quantity of interest. The HB method is being extensively used for SAE because it is straightforward, inferences are "exact," and complex problems can be handled using Markov chain Monte Carlo (MCMC) methods. Software for implementing the HB method is also available (Section 10.2.4). Chapter 10 gives a self-contained account of the HB method and its applications to SAE.

"Optimal" model-based estimates of small area totals or means may not be suitable if the objective is to produce an ensemble of estimates whose distribution is in some sense close enough to the distribution of the corresponding estimands. We are also often interested in the ranks (e.g., ranks of schools, hospitals, or geographical areas) or in identifying domains (areas) with extreme values. Ideally, it is desirable to construct a set of "triple-goal" estimates that can produce good ranks, a good histogram, and good area-specific estimates. However, simultaneous optimization is not feasible, and it is necessary to seek a compromise set that can strike an effective balance between the three goals. Triple-goal EB estimation and constrained EB estimation that preserves the ensemble variance are studied in Section 9.8.

1.6 SOME EXAMPLES

We conclude the introduction by presenting some important applications of SAE as motivating examples. Details of some of these applications, including auxiliary information used, are given in Chapters 6–10.

1.6.1 Health

SAE of health-related characteristics has attracted a lot of attention in the United States because of a continuing need to assess health status, health practices, and health resources at both the national and subnational levels. Reliable estimates of health-related characteristics help in evaluating the demand for health care and the access that individuals have to it. Healthcare planning often takes place at the state and substate levels because health characteristics are known to vary geographically. Health System Agencies in the United States, mandated by the National Health Planning Resource Development Act of 1994, are required to collect and analyze data related to the health status of the residents and to the health delivery systems in their health service areas (Nandram, Sedransk, and Pickle 1999).

(i) The US National Center for Health Statistics (NCHS) pioneered the use of synthetic estimation based on implicit linking models. NCHS produced state synthetic estimates of disability and other health characteristics for different

groups from the National Health Interview Survey (NHIS). Examples 3.2.2 and 10.13.3 give health applications from national surveys. Malec, Davis and Cao (1999) studied HB estimation of overweight prevalence for adults by states, using data from NHANES III. Folsom, Shah, and Vaish (1999) produced survey-weighted HB estimates of small area prevalence rates for states and age groups, for up to 20 binary variables related to drug use, using data from pooled National Household Surveys on Drug Abuse. Chattopadhyay et al. (1999) studied EB estimates of state-wide prevalences of the use of alcohol and drugs (e.g., marijuana) among civilian non-institutionalized adults and adolescents in the United States. These estimates are used for planning and resource allocation and to project the treatment needs of dependent users.

(ii) Mapping of small area mortality (or incidence) rates of diseases, such as cancer, is a widely used tool in public health research. Such maps permit the analysis of geographical variation that may be useful for formulating and assessing etiological hypotheses, resource allocation, and the identification of areas of unusually high-risk warranting intervention (see Section 9.6). Direct (or crude) estimates of rates called standardized mortality ratios (SMRs) can be very unreliable, and a map of crude rates can badly distort the geographical distribution of disease incidence or mortality because the map tends to be dominated by areas of low population. Disease mapping, using model-based estimators, has received considerable attention. We give several examples of disease mapping in this text (see Examples 9.6.1, 9.9.1, 10.11.1, and 10.11.3). Typically, sampling is not involved in disease mapping applications.

1.6.2 Agriculture

The US National Agricultural Statistics Service (NASS) publishes model-based county estimates of crop acreage using remote sensing satellite data as auxiliary information (see Example 7.3.1 for an application). County estimates assist the agricultural authorities in local agricultural decision making. Also, county crop yield estimates are used to administer federal programs involving payments to farmers if crop yields fall below certain levels. Another application, similar to Example 7.3.1, to estimate crop acreage in small areas using ground survey and remote sensing data, is reported in Ambrosio Flores and Iglesias Martínez (2000). Remote sensing satellite data and crop surveys are used in India to produce direct estimates of crop yield at the district level (Singh and Goel 2000). SAE methods are also used in India to obtain estimates of crop production at lower administrative units such as "tehsil" or block, using remote sensing satellite data as auxiliary information (Singh et al. 2002).

An application of synthetic estimation to produce county estimates of wheat production in the state of Kansas based on a non-probability sample of farms is presented in Example 3.2.4. Chapters 6 and 7 of Schaible (1996) provide details of traditional and model-based indirect estimation methods used by NASS for county crop acreage and production.

1.6.3 Income for Small Places

Example 6.1.1 gives details of an application of the EB (EBLUP) method of estimation of small area incomes, based on a basic area level linking model (see Section 6.1). The US Census Bureau adopted this method, proposed originally by Fay and Herriot (1979), to form updated per capita income (PCI) for small places. This was the largest application (prior to 1990) of model-based estimators in a US Federal Statistical Program. The PCI estimates are used to determine fund allocations to local government units (places) under the General Revenue Sharing Program.

1.6.4 Poverty Counts

The Fay–Herriot (FH) method is also used to produce model-based county estimates of poor school-age children in the United States (National Research Council 2000). Using these estimates, the US Department of Education allocates annually several billions of dollars called Title I funds to counties, and then states distribute the funds among school districts. The allocated funds support compensatory education programs to meet the needs of educationally disadvantaged children. In the past, funds were allocated on the basis of updated counts from the previous census, but this allocation system had to be changed since the poverty counts vary significantly over time. EBLUP county estimates in this application are currently obtained from the American Community Survey (ACS) using administrative data as auxiliary information. Example 6.1.2 presents details of this application.

1.6.5 Median Income of Four-Person Families

Estimates of the current median income of four-person families in each of the states of the United States are used to determine the eligibility for a program of energy assistance to low-income families administered by the US Department of Health and Human Services. Current Population Survey (CPS) data and administrative information are used to produce model-based estimates, using extensions of the FH area level model (see Examples 8.1.1 and 8.3.1).

1.6.6 Poverty Mapping

Poverty measures are typically complex non-linear parameters, for example, poverty measures used by the World Bank to produce poverty maps in many countries all over the World. EB and HB methods for estimating poverty measures in Spanish provinces are illustrated in Example 10.7.1.