Bibliography on Capture-Recapture Modelling With Application to Census Undercount Adjustment

STEPHEN E. Fienberg

ABSTRACT

This article presents a selected annotated bibliography of the literature on capture-recapture (dual system) estimation of population size, on extensions to the basic methodology, and the application of these techniques in the context of census undercount estimation.

KEY WORDS: Capture-recapture; Census undercount; Dual system estimation; Loglinear models.

1. INTRODUCTION

The method of capture-recapture for estimating the size of a closed population has been in use since at least the nineteenth century, when Peterson (1896) developed the standard estimator that bears his name for the use with fish populations. Subsequent application to other types of populations include Geiger and Werner (1924) – physics; Lincoln (1930) – wildlife; Chandrasekar and Deming (1948) – vital statistics for human populations; Wittes and Sidel (1968), Wittes, Colton and Sidel (1974) – epidemiology; Sanathanan (1972b) – particle scanning in physics; Blumenthal and Marcus (1975) – life testing; Green and Stollmack (1981), Rossmo and Routledge (1990) – crimes and criminals. In the context of the study of human populations and demography the method is often referred to as dual system estimation. We have included virtually no references to the related problem of counting the number of species, which goes back to the work of R.A. Fisher in the 1940s and had an elegant formulation in Efron and Thisted’s (1976) Biometrika paper on “How many words did Shakespeare know?”.

The basic capture-recapture approach rests on a number of assumptions, e.g.: (1) the population under study is closed; (2) individuals (units) can be perfectly matched from capture to recapture; (3) capture probabilities are constant across the individuals (units) in the population; (4) the probability of inclusion of an individual (unit) in recapture sample is independent of inclusion in original census or sample. Beginning in the late 1930s various investigators began to explore extensions that allowed for departures from the assumptions. These methods typically require additional data such as a second recapture (or even a third) and the full capture-recapture history of each individual.

For human populations and the study of vital statistics the methodology has long been linked to census data, e.g., see Tracy (1941) and Shapiro (1949, 1954). In connection with the 1950 decennial census of population, the U.S. Bureau of the Census introduced the use of a sample matched to the census records for coverage evaluation. This approach has evolved into what is currently known as the Post Enumeration Survey approach to undercount and overcount estimation, and it has been the focal point of the recent and ongoing controversy of the possible adjustment of the 1980 and 1990 censuses, e.g., see Eriksen and Kadane (1985); Freedman and Navidi (1986, 1992); Freedman (1991); Wolter (1991).

1 Stephen E. Fienberg, Office of Vice President (Academic Affairs), York University, North York, Ontario M3J 1P3, Canada.
This selected annotated bibliography presents an overview of published literature on capture-recapture estimation of population totals. It includes historical references, articles that explore departures from assumptions and extensions of the basic methodology, and is most complete in connection with papers that describe the dual and multiple system approaches in the context of census undercount estimation. In this regard, however, we have not included references to any of the unpublished memoranda and papers from the U.S. Bureau of the Census (primarily because most of these have been replicated in some form in the published literature). We have tended to exclude articles published in un refereed proceedings for related reasons. Because the literature on specialized applications of capture-recapture techniques to wildlife populations is so extensive, and only some of it is of relevance for human populations, we have provided primarily references to reviews of this literature, e.g., see Brownie et al. (1977); Otis et al. (1978); Seber (1973, 1982). Similarly we have included only a small number of references to the more specialized methods in use for life testing, e.g., see Dahiya and Blumenthal (1986), as well as those in use for software reliability applications, e.g. Jelinski and Moranda (1972), and Duran and Workowski (1981). The methods in this latter literature diverge in significant ways from those used in the basic capture-recapture and dual system approaches.

2. SELECTED BIBLIOGRAPHY

• Extends usual dual systems approach to allow for multiplicative stratification effects.

• Links cross-classification of covariates to the capture and recapture via loglinear models and then uses EM algorithm to estimate population size.

• Develops models for evaluating impact of matching error on census coverage.

• Monograph on loglinear models which includes a chapter on the relationship to capture-recapture models.

• Uses exponential distribution to estimate population size based on a subset of observations obtained by truncated sampling.

• Gives succinct summary of several basic variants on capture-recapture models and their estimation.

• Describes a comprehensive range of capture-recapture models and appropriate goodness-of-fit tests, with emphasis on banding experiments.

- Describes the survey-based accounting approach of the reverse record check for undercount estimation. Does not deal with issue of exclusion of individuals from census and other lists.


- Combines methodology of Brownie et al. for band recovery with survival estimation under Jolly-Seber mark-recapture models.


- Develops a capture-recapture model with heterogeneity for animals but constant probabilities of capture across samples. Model induces dependencies amongst captures.


- Develops a Bayesian approach using beta priors for traditional independence-based Schnabel census model for multiple recapture data.


- Extends Chandrasekar-Deming approach to three or more sources.


- Develops dual-system technique and suggests the use of stratification for eliminating heterogeneity. Applies approach to estimation of number of births and deaths in several Indian villages.


- Explores heterogeneous catchability model of Burnham and Overton using a moment inequality to get a lower bound on population size.


- Explores adequacy of estimator resulting from moment inequality for heterogeneous catchability model in settings involving sparse data.


- Develops the hypergeometric sampling model for estimating the population size in capture-recapture studies.


- Describes the use of dual system estimation and a post enumeration survey to adjust the results of the Australian census. Also applies Wolter sex-ratio technique to check on sensitivity of dual system estimator.


- Report of a panel of the Committee on National Statistics on census methodology including an examination of the dual systems approach to undercount correction.


- Proposes the use of two lists covering the same sample from a population.


- Examines effect of bias and variability on accuracy of adjusted and unadjusted census counts and the impact on the reapportionment of the U.S. House of Representatives.


- Introduces Poisson model for capture-recapture and uses it with loglinear models to extend standard approach to allow for birth, death, and trap dependency.


- Uses Poisson model and loglinear representation for inclusion of birth, death, and trap dependency into standard capture-recapture approach.


- Compares MLEs of parameters under the two models and presents relationship between the corresponding asymptotic variances and covariances.


- Extends dual systems approach to situation involving clustered observations as in the U.S. census coverage improvement program.


- Proposes a PES-based model for undercount adjustment utilizing an empirical Bayes estimation scheme and a family of loss functions.


- Develops and applies empirical Bayes smoothing methods for census adjustment factors produced from dual systems approach for geographic by demographic stratification. Applies approach to state data from 1980 U.S. census.


- Shows that synthetic estimation approach used by Isaki *et al.* is special case of empirical Bayes.

• Reviews theory underlying population size estimation from truncated sampling for discrete distributions and provides references to domains of application.

• Describes the maximum likelihood approach to the multiple recapture problem under complete independence.

• Extends the maximum likelihood approach under independence to open populations with either immigration or death.

• Extends the maximum likelihood approach with independence to the situation where the original captured individuals are stratified into s groups and the individuals in the recapture sample are stratified, but according to t (possibly different) strata.

• Extends triple system estimation to allow for individual heterogeneity and selected forms of dependence. Applies estimators to triple system data from census dress rehearsal in St. Louis.

• Presents an alternative estimator for capture-recapture problems with interesting asymptotic properties.

• Extends Chandrasekar-Deming approach to three or more sources.


• Extends of Chandrasekar-Deming approach to three sources.

• Describes implementation of post-enumeration survey approach to dual system estimation in a test census.

• Develops a probabilistic matching model for use with dual and multiple system estimation, and considers a Bayesian approach for estimating the population size. Illustrates techniques using data from test census results from Los Angeles.

- Develops a probabilistic matching model for use with dual system estimation and illustrates its application to data from test census results from Los Angeles.


- Suggests a parametric model due to Fisher and a nonparametric model for the classical species problem using empirical Bayes methods. Applies approach to the vocabulary of Shakespeare.


- Review of literature and methods for dual- and multiple systems estimation. Includes sections comparing use of techniques and departures from assumptions in wildlife and human populations.


- Applies dual system approach to 1980 census data, including the regression-based smoothing of undercount estimates and the estimation of adjusted odds ratios using demographic estimates.


- Presents revisions and extensions to the Ericksen and Kadane methodology and a critique of Freedman and Navidi.


- Attempts to test hypotheses regarding the causes of census undercount for a hard-to-enumerate Hispanic urban population.


- Introduces a method for estimating dependencies among multiple lists using loglinear models and develops a general approach for estimation using results on $2^k$ incomplete contingency tables and conditional estimation.


- Presents historical background on the differential undercount of the U.S. population and brief descriptions of demographic analysis and the dual system estimation approaches.


- Critique of dual systems approach to adjustment of the 1990 census.
• Critique of Ericksen and Kadane dual systems methodology as applied to 1980 census data.

• Continues critique of the use of dual system estimation and synthetic adjustment as applied to 1980 census.

• Uses a recursive relationship to generate point and interval estimate for multiple-recapture census under independence.

• Applies a capture-recapture method to radium ion particle detection estimation.

• Applies capture-recapture models to problems in medical screening.


• Introduces a non-zero value for the response correlation, by taking the mid-point of the range of permissible correlation values, and consequently derives a value for missing cell. Applies approach to census data from Malawi.

• Applies bounds on correlation in a $2 \times 2$ table to dual system estimation in the presence of event correlation induced by heterogeneity.

• Uses approximation for upper bound for response correlation to derive an upper bound for missing cell.

• Reports on Los Angeles Test of Adjustment Related Operations and estimates of sources of bias in post enumeration survey and census-based dual systems estimates.

• Applies results on successive sampling to derive asymptotic distribution of usual Peterson estimator when there are heterogeneous capture probabilities or the effects of matching.
- Applies different loglinear related methods used to study the number of infants born with Downs syndrome.

- Uses linear logistic models for capture probabilities for individuals and capture occasions.

- Uses linear logistic models for capture probabilities and exploits temporal order of captures to introduce dependence amongst captures and on measurable covariates for those captured at least once.

- Exploits upper bound on correlation bias to reduce the bias of the dual system estimator.

- Uses demographic analysis data to get revised dual system estimates for 1980 census using different models for correlation bias.

- Develops a simple matching error model in the presence of correlation bias to compare three dual system estimators.

- Develops simulation populations based on 1980 census and coverage evaluation results, evaluates regression-based synthetic undercount estimation methods, and shows superiority of synthetic approaches to raw census counts.

- Shows that positive correlation bias produces a downward bias in estimate of total population size.

- Describes census methodology for matching census and post-enumeration survey records, with the results from their application to 1985 test census.

- Proposes a model with exponentially distributed failures to estimate total number of program faults based on times of occurrence of failures in fixed time period.


- Estimation from multiple-recapture data for open populations.
• Demonstrates the impact of correlation bias resulting from collapsing over heterogeneous strata with different catchability probabilities in each strata, subject to a monotonicity constraint.

• Reviews the use of dual system estimation for vital records in various countries. Includes technical details on the use of complex samples and elaborations on basic techniques.

• Estimates population size from the last of $k$ lists.

• Applies dual system estimation to surveillance of infectious diseases.

• Applies capture-recapture method to estimating size of waterfowl populations.

• Shows how heterogeneity induces correlation bias (event correlation) in the estimation of disease prevalence.

• Comprehensive review of dual-systems estimation, assumptions, background, design, and problems. Contains claim that the basic method has been used for more than three centuries for estimating size of animal populations.


• Develops a total error model for dual systems approach applied to Los Angeles Test of Adjustment Related Operations.

• Extends earlier Mulry-Spencer development of total error model for dual systems approach and applies it to 1988 dress rehearsal census in St. Louis and east-central Missouri.


• Reviews capture-recapture and related methods for wildlife populations.

- Classic development of method of capture-recapture and its application to the estimation of the size of fish populations.


- Develops a formula for bias in dual system estimator under a general model of response errors and explores use of double sampling to correct bias.


- Presents a methodology matching for undercount estimation which utilizes an imputation approach rooted in loglinear models in the presence of missing data.

- Demonstrates asymptotic equivalence of conditional and unconditional estimators for the population size.

- Develops latent model to estimate the number of particles in scanning records which allows for differential detectability and induces dependencies amongst detectors.

- Applies traditional capture recapture model and latent models to data from actual visual scanning experiments.

- Shows relationship between the asymptotic variances of the population size under general capture-recapture model for the two alternate sampling schemes.

- Examines effect of missing data on dual system estimate applied to test census data.

- Applies synthetic estimation approaches to 1980 census data to evaluate the impact of undercount estimation.

- Extends the basic capture-recapture approach to multiple recaptures, with information at each recapture on whether individuals were captured previously.


- Estimation from multiple-recapture data for open populations with time-specific parameters.


- Contains an up-to-date review of capture-recapture techniques and their extensions for animal populations, with emphasis on applications.


- Reviews capture-recapture approach for both closed and open populations and provides guide to wildlife and fisheries applications.


- Describes use of 1940 U.S. census data to check on completeness of birth registration using Chandrasekar-Deming estimator.


- Describes use of U.S. census data from 1940 and 1950 to check on completeness of birth registration using Chandrasekar-Deming estimator.


- Adapts author's approach of multiplicity surveys for rare events to dual systems problem.


- Uses Poisson approximation and Gamma prior for a Bayesian approach to estimation under independence in multiple recapture model.


- Develops exact Bayesian posterior distribution for multiple recapture census under independence of recaptures.


- An early application of dual systems approach to census data.
- Examines methods of adjusting linkage rules for matching dual-system records when independence cannot be assumed.

- Applies multiple systems and independence to estimate the size of a population of children with a congenital anomaly and other problems.

- Applies multiple systems and independence to estimate the size of a population of children with a congenital anomaly.

- Uses capture-recapture approach to estimate number of hospital patients using methicillin.

- Describes basic version of dual systems approach as used for the 1980 census, including the elimination of erroneous enumerations.

- Uses sex ratio to get a modified capture-recapture estimator for data with stratification by sex, assuming either a common odds ratio, or independence in one stratum. Applies approach to animal data and describes application to census data.

- Describes 1990 census adjustment procedures and why they are statistically defensible.

- Examines through simulations the improvements of synthetic adjustment over the census counts for small areas.

- Applies various loglinear and related models of dependence to analyze triple system data from census dress rehearsal in St. Louis.