Editorial – Recent Developments in Capture-Recapture Methods and Their Applications

Dankmar Böhning*

Quantitative Biology and Applied Statistics, School of Biological Sciences, Harry Pitt Building, Whiteknights, Reading, RG6 6FN, UK

Key words: Biodiversity; Dual-system estimation; Heterogeneity; Horvitz-Thompson estimator; Lincoln-Petersen estimator; Log-linear models; Mark-recapture; Mixture models; Population size estimation; Species estimation.

In recent years capture-recapture methods have experienced important theoretical developments. New application areas have been added to their spectrum, in turn supporting new developments on the methodological side. Hence it appears appropriate to take a closer look at some of these developments. This is the objective of this special theme in the current issue of the Biometrical Journal. The contributions to this special theme stem from a recent Conference of Recent Developments in Capture-Recapture Methods and their Applications (held on July 12–13, 2007, School of Biological Sciences, University of Reading, UK).

Capture-Recapture has its origin in the Biological/Ecological Sciences with the work of Lincoln and Petersen. More than one hundred years ago Petersen (1896) published his landmark paper suggesting what later became known as the Lincoln–Petersen estimator, since it was also independently developed by Lincoln (1930). This estimator and its stabilized version developed by Chapman (1951) are still in use by numerous practitioners. In a way, this issue could be considered a celebration of the centenary of the Lincoln–Petersen estimator.

Capture-Recapture methods can be seen to be applied in three major sciences, all represented in various contributions of this special topic issue:

- The biological sciences, where the size of animal populations and their diversity are of importance.
- The social sciences, where we are interested in the amount of illegal activities, such as illegal immigration.
- The life and medical sciences, where we often want to know the size of the hidden disease burden in a population, such as depression or drug use.

In the biological sciences the classical book by Seber (1982) characterizes the various biological application fields. In addition, the recent monograph by Borchers, Buckland and Zucchini (2002) supplements a view with biological applications using modern inference tools. More recently, capture-recapture methods have been applied to other areas. The landmark paper by Sekar and Deming (1949) applied capture-recapture methods to estimate the number of births and deaths in an area near Calcutta. This paper started a development of the capture-recapture methodology in demography and social science in general. This development was further supplemented and supported by the appearance of the book by Bishop, Fienberg, and Holland (1975) on multivariate discrete analysis in which the entire chapter 6 was devoted to capture-recapture methods. Yet a further milestone was the establishment of the methodology in medicine and public health by the landmark papers by Hook and Regal (1995) as well as the two papers by the International Working Group for Disease Monitoring and Forecasting (IWGDMF 1995a, b). In the early phases the method was used to develop a more com-

* Corresponding author: e-mail: d.a.w.bohning@reading.ac.uk, Phone: +44 (0) 11 8378 8032, Fax: 44 (0) 11 8975 3169

plete picture of disease occurrence. Nowadays, the methodology is used in screening studies in preventive studies as well as it is establishing its place in clinical settings.

This special theme on capture-recapture is opened by two papers of partly review character. The first one is by *Chao, Pan, and Chiang* reviewing the Lincoln–Petersen estimator including the effects of local list-dependence and heterogeneity in the capture probabilities. The Lincoln-Petersen estimator is one of the most popular approaches in capture-recapture. However, the estimator is not valid when the capture sample and recapture sample are not independent. An intuitive interpretation for "independence" between samples based on 2×2 categorical data formed by capture/non-capture in each of the two samples is provided. The estimator is also extended for the case of two shared populations. This new estimator of the size of the shared population is investigated and its variance is derived. The proposed method is applied to a study of the relapse rate of illicit drug use in Taiwan.

The second introductory paper is by *Bunge and Barger* who consider parametric distributions intended to model heterogeneity in population size estimation with emphasis on parametric stochastic abundance models for species richness estimation. The paper summarizes the results of fitting seven candidate models to frequency-count data, from a database of more than 40 000 such instances, mostly arising from microbial ecology. It is found that finite mixtures of a small number of components (point masses or simple diffuse distributions) represent a promising direction. Finally, the connections between parametric models for abundance and incidence data are explored, again noting the usefulness of finite mixture models.

A first and larger group of papers deals with *mixture models* in their connection with capture-recapture methods. *Mao* investigates nonparametric maximum likelihood estimation for two classes of mixture models: the mixture of binomial and the mixture of geometric distributions. Whereas the first mixture model mimics the situation of a Schnabel census in which a number of fixed "trapping" occasions is used, the second model is designed for removal situations. The paper characterizes theoretical results as well as suggests a fast algorithm to compute the nonparametric maximum likelihood estimator. *Kuhnert, Del Rio Vilas, Gallagher, and Böhning* also consider nonparametric maximum likelihood estimation for mixture models. Here, the emphasis is on the so-called boundary problem describing the fact that the NPMLE of the population size necessarily overestimates. A strategy based upon bagging is suggested to diminish the overestimation effect. *Viwatwongkasem, Kuhnert, and Satitvipawee* also investigate mixture models. They compare the finite mixture model estimator with a list of other popular estimators in a simulation study. The results are favourable for the mixture models although the price to pay is a high computational effort. *Pledger and Phillpot* provide a very flexible class of mixture models allowing for time-, recapture-, and unit-heterogeneity effects in the capture probabilities. A skink data example illustrates the usage of this class of models as well a discussion on the extension to open population models is provided. *Cruyff and van der Heijden* are interested in incorporating covariates into the estimation of population size and present the zero-truncated negative binomial regression model to estimate the population size in the presence of a single registration file. The model is a more general alternative to the zero-truncated Poisson regression model and it may be useful if the data are overdispersed due to unobserved heterogeneity. As an application, the size of the population of opiate users in the city of Rotterdam is estimated with this new model.

A second group contains two papers using Bayesian approaches as well as an application study on capture-recapture procedures to estimate cancer registry completeness. *Fienberg and Manrique-Vallier* revisit the heterogeneous closed population multiple recapture problem and model individual-level heterogeneity with the Grade of Membership model. This model is an extended and further developed version of the *latent class model* (Lazarsfeld and Henry, 1968) which allows to postulate the existence of homogeneous latent "ideal or "pure classes within the population, and construct a soft clustering of the individuals, where each one is allowed partial or mixed membership in all of these classes. A full hierarchical Bayes specification and a MCMC algorithm is proposed to obtain samples from the posterior distribution. The method is applied simulated data and to three real life examples dealing with diabetes patients in Casale Monferrato (Italy), children suffering a specific congenital anomaly in Massachusetts (USA), and killings and disappearances due to political violence in the District of

Chungui (Peru). *Barger and Bunge* focus again in estimating the number of species in a closed population. In order to conduct a noninformative Bayesian inference when modeling this data, they derive Jeffreys and reference priors from the full likelihood. Two specific cases are considered which assume that the mean abundances are constant or exponentially distributed. Then, the Jeffreys and reference priors are functions of the Fisher information for the model parameters and the information is calculated in part using the linear difference score for integer parameter models. It is shown that the Jeffreys and reference priors perform similarly in a data example consisting of a sample of microbial organisms collected from the Framvaren Fjord in Norway. *Schmidtmann* evaluates various methods for determining the completeness of a disease registry. Completeness of registration is one of the quality indicators usually reported by cancer registries. Several methods have been suggested to estimate completeness. In this paper a multi-state model for the process of cancer diagnosis and treatment is presented. Data were simulated according this model and several capture-recapture methods have been applied to the simulated data. In the scenarios investigated here, all capture-recapture estimators tended to underestimate completeness whereas a modified DCN method and one type of log-linear model yielded reasonable estimates.

Acknowledgement *The author is very grateful to the Editors of the Biometrical Journal for the opportunity to publish a number of selected papers in this special topic issue. In addition, the author would like to thank the following colleagues for their service as Associate Editors for this special issue: Kathryn Barger, David Borchers, John Bunge, Steven Fienberg, Changxuan Mao, Peter van der Heijden, Shirley Pledger, and Irene Schmidtmann.*

References

Bishop, Y. M. M., Fienberg, S. E., and Holland, P. W. (1975). *Discrete Multivariate Analysis: Theory and Practice*. MIT Press, Cambridge.

Borchers, D. L., Buckland, S. T., and Zucchini, W. (2002). *Estimating Animal Abundance: Closed Populations*. Heidelberg: Springer.

Chapman, D. G. (1951). Some Properties of the Hypergeometric Distribution With Applications to Zoological Censuses. *University of California Publications in Statistics* **1**, 131–160.

Hook, E. B. and Regal, R. (1995). Capture-Recapture Methods in Epidemiology: Methods and Limitations. *Epidemiologic Reviews* **17**, 243–264.

International Working Group for Disease Monitoring and Forecasting (IWGDMF). (1995a). Capture-recapture and multiple record systems estimation I. History and theoretical development. *American Journal of Epidemiology* **142**, 1047–1058.

International Working Group for Disease Monitoring and Forecasting (IWGDMF). (1995b). Capture-recapture and multiple record systems estimation II. Application in human diseases. *American Journal of Epidemiology* **142**, 1059–1068.

Lazarsfeld, P. F. and Henry, N. W. (1968). *Latent Structure Analysis.* Boston: Houghton Mifflin.

Lincoln, F. C. (1930). Calculating Waterfowl Abundance on the Basis of Banding Returns. *United States Department of Agriculture Circular* **118**, 1–4.

Petersen, C. G. J. (1896). The Yearly Immigration of Young Plaice Into the Limfjord From the German Sea. *Report of the Danish Biological Station (1895)* **6**, 5–84.

Seber, G. A. F. (1982). *The Estimation of Animal Abundance*, 2nd edition. London: Griffin.

Sekar, C. and Deming, W. E. (1949). On a method of estimating birth and death rates and the extent of registration. *Journal of the American Statistical Association* **44**, 101–115.

