Estimating transition probabilities in unmarked populations – entropy revisited

EVAN G. COOCH[1]* *and* WILLIAM A. LINK[2] [1]*Department of Biological Sciences, Simon Fraser University, Burnaby, British Columbia, Canada and* [2]*US National Biological Service, Patuxent Wildlife Research Center, Laurel, Maryland, USA*

The probability of surviving and moving between 'states' is of great interest to biologists. Robust estimation of these transitions using multiple observations of individually identifiable marked individuals has received considerable attention in recent years. However, in some situations, individuals are not identifiable (or have a very low recapture rate), although all individuals in a sample can be assigned to a particular state (e.g. breeding or non-breeding) without error. In such cases, only aggregate data (number of individuals in a given state at each occasion) are available. If the underlying matrix of transition probabilities does not vary through time and aggregate data are available for several time periods, then it is possible to estimate these parameters using least-squares methods. Even when such data are available, this assumption of stationarity will usually be deemed overly restrictive and, frequently, data will only be available for two time periods. In these cases, the problem reduces to estimating the most likely matrix (or matrices) leading to the observed frequency distribution of individuals in each state. An entropy maximization approach has been previously suggested. In this paper, we show that the entropy approach rests on a particular limiting assumption, and does not provide estimates of latent population parameters (the transition probabilities), but rather predictions of realized rates.

The probability of successfully making the transition from one state to another (where 'states' could be developmental stages, or a physiological condition, for example) is of fundamental importance to the analysis of life histories. Consider the general question of breeding propensity – the probability of breeding in a given year. Clearly, selection will favour the optimal scheduling of breeding versus non-breeding over a lifetime, subject to the potential costs (i.e. mortality) and benefits (i.e. offspring production) of being in a particular state at any time *t*.[1–3]

Although there is considerable theory on the optimal schedule of life-history transitions under a specified set of constraints, comparatively little effort has been focused on the general problem of estimation of transition rates in wild populations. Recent work, however, has shown that mark–recapture analysis can provide robust estimates of both the transition rates themselves, as well as the covariances amongst these transitions.[4–10]

Mark–recapture approaches generally assume a sample of uniquely marked individuals which can be followed through time. What, if anything, can be estimated if few or no individuals are individually marked? In such cases, only the total numbers in particular states at each time are known. Is it possible to estimate the transition probabilities from aggregate data alone? There is evidence suggesting that estimates can be derived under time-invariant conditions using least-squares methods, and that these estimates are generally robust with a sufficient number of sampling occasions,[11–13] with increasing precision when supplemented with even low amounts of additional information

*Correspondence author. Present address: US Fish & Wildlife Service, Patuxent Wildlife Research Center, Laurel, MA, USA.
Email: Evan_Cooch@mail.fws.gov

(e.g. incidental recaptures of identifiable individuals).[14] However, the assumption of a stationary probability vector for the individual transitions is unlikely to be met in most biological situations. One approach which makes no *a priori* assumptions about time-invariance of the joint distribution involves an entropy maximiztion procedure.[15] Although entropy maximization is relatively straightforward to apply (and has in fact been previously applied to multistate problems in human demography[16,17]), we show that the approach rests on several significant, and generally limiting, assumptions. However, while entropy maximization may yield little information from aggregate data, we show that analysis of such data does yield some information, which may be of utility in supplementing analysis of sparse mark–recapture data sets.

ENTROPY MAXIMIZATION: BACKGROUND

Consider a system where individuals must be in one of two states (say, breeding B and non-breeding N). The only information available is the total number of individuals in either state observed at time i and $i + 1$. For simplic-ity, we assume that the population is effectively closed between these two occasions, and is completely enumerated. Arguably, these assumptions are no more or less realistic than assuming the transition matrix is stationary and Markovian (which would allow for least-squares estimation[11-14]). However, our intent at this stage is to reduce the initial problem to its simplest form. Some possible consequences of relaxing these assumptions are briefly noted in a subsequent section.

At each occasion, we can determine (with perfect precision) the number of individuals in each of the two states. Following Willekens,[16,17] let m_{xy} equal the absolute number of individuals moving from state x at time i to state y at time $i + 1$. Thus, m_{BN} is the absolute number of individuals in breeding condition B at time i moving to non-breeding status N at time $i + 1$. There are, for a two-state model, four possible transitions: m_{BB}, m_{NB}, m_{BN}, and m_{NN}.

Let m_{B*} represent the total number of individuals in breeding state at time i. Reading the subscript from left to right, the B indicates that the individuals were originally in breeding condition, and then made a transition during the interval from i to $i + 1$. The asterisk reflects the fact that the transition could have been (a) from breeding to non-breeding (i.e. m_{BN}), or (b) from breeding to breeding – remaining in breeding state (i.e. m_{BB}). Thus, m_{N*} is the number of non-breeders at time i, m_{*B} is the number of breeders at time $i + 1$, and m_{*N} is the number of non-breeders at time $i + 1$. We can tabulate these values in an origin–destination (O–D) table (Table 1) where 'origin' is the state of an individual at time i, and 'destination' is the state of the same individual at time $i + 1$. We observe only the row and column totals; we are then faced with the problem of 'predicting' what m_{ij} (and consequently, the probability ϕ_{ij} of making the transition from state i to state j values must have been to 'realize' these totals.

Table 1. Origin-destination table.

	i (origin)		
i + 1 (destination)	B	N	'Arrivals'
B	m_{BB}	m_{NB}	m_{*B}
N	m_{BN}	m_{NN}	m_{*N}
'Departures'	m_{B*}	m_{N*}	m_{**}

For example, suppose that $m^{**} = 20$ individuals (i.e. 20 individuals in the population on both occasions). On occasion i, there were ten individuals in breeding condition (B; $m_{B*} = 10$), and ten individuals in non-breeding condition (N; $m_{N*} = 10$). On occasion $i + 1$, there were 14 individuals in breeding condition ($m_{*B} = 14$), and six individuals in non-breeding condition ($m_{*N} = 6$).

In this example, there are seven possible combinations (**M**, 'macrostates') of m_{ij} entries of the matrix which lead to the observed numbers of individuals in each state at each occasion in this example (i.e. seven macrostates which satisfy the observed column and row totals in this example).

$$\mathbf{M}_a = \begin{bmatrix} 10 & 4 \\ 0 & 6 \end{bmatrix} \quad \mathbf{M}_b = \begin{bmatrix} 9 & 5 \\ 1 & 5 \end{bmatrix}$$

$$\mathbf{M}_c = \begin{bmatrix} 8 & 6 \\ 2 & 4 \end{bmatrix} \quad \mathbf{M}_d = \begin{bmatrix} 7 & 7 \\ 3 & 3 \end{bmatrix}$$

$$\mathbf{M}_e = \begin{bmatrix} 6 & 8 \\ 4 & 2 \end{bmatrix} \quad \mathbf{M}_f = \begin{bmatrix} 5 & 9 \\ 5 & 1 \end{bmatrix} \quad \mathbf{M}_g = \begin{bmatrix} 4 & 10 \\ 6 & 0 \end{bmatrix}$$

where

$$\mathbf{M}_i = \begin{bmatrix} m_{\mathrm{BB}} & m_{\mathrm{NB}} \\ m_{\mathrm{BN}} & m_{\mathrm{NN}} \end{bmatrix}$$

Each macrostate is consistent with the observed row and column totals (i.e. the column totals $m_{\mathrm{B}*}$ and $m_{\mathrm{N}*}$ and row totals $m_{*\mathrm{B}}$ and $m_{*\mathrm{N}}$, and the total number of individuals m_{**} are the same for all seven macrostates $\mathbf{M}_i$). However, which macrostate ($\mathbf{M}_a$, ..., $\mathbf{M}_g$) is 'most likely'?

The entropy approach makes two key logical assumptions:

- that the probability that a given macrostate is the true ('most likely') macrostate is proportional to the number of ways in which it could be achieved (number of 'microstates', $\mathbf{W}_i$);
- that each of the ways in which it could be achieved is equally likely.[15]

The number of microstates is clearly related to the number of ways to select X individuals from n total individuals without replacement, which is given by the binomial:

$$\binom{n}{X} = \frac{n!}{(n-X)!X!}$$

In general, the number of microstates $\mathbf{W}$ for a given macrostate $\mathbf{M}_i$ is:

$$\mathbf{W} = \frac{m_{**}!}{\prod_{i,j} m_{ij}!} \qquad (1)$$

For the macrostates $\mathbf{M}_a$,..., $\mathbf{M}_g$ the number of microstates ($\mathbf{W}$) is maximal for macrostate $\mathbf{M}_d$ (Fig. 1). The realized transition rates (ϕ_{ij}) are determined by dividing each m_{ij} value by the column totals of the O–D table (since the column totals $m_{\mathrm{B}*}$ and $m_{\mathrm{N}*}$ are the total number of individuals that have moved from either state):

$$\mathbf{M}_d = \begin{bmatrix} 7 & 7 \\ 3 & 3 \end{bmatrix}$$

$$\mathbf{A} = \begin{bmatrix} \phi_{\mathrm{BB}} & \phi_{\mathrm{NB}} \\ \phi_{\mathrm{BN}} & \phi_{\mathrm{NN}} \end{bmatrix} = \begin{bmatrix} 0.7 & 0.7 \\ 0.3 & 0.3 \end{bmatrix}$$

These realized rates are to be distinguished from the latent rates (ψ_{ij}), which are demographic parameters governing the system; we highlight this distinction in our subsequent discussion.

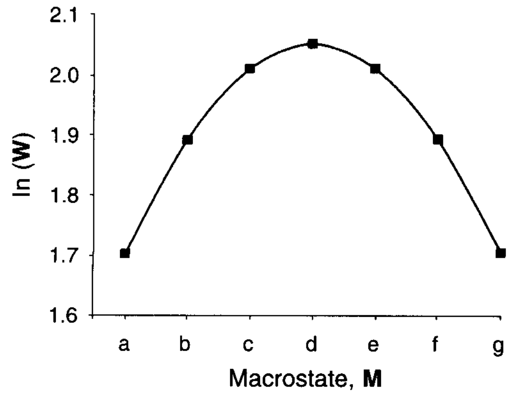

Figure 1. Plot of **W** (Eqn. 1) for seven possible macrostates in numerical example in text. For ease of computation, **W** is calculated using natural logarithms of factorials.

In fact, since the margins of the O–D table are fixed, the m_{ij} elements are completely determined by any one of the four entries, and thus the assumptions of the entropy method lead to the same probability structure as treating any one entry as a hypergeometric random variable. Consider m_{BB}. Since $m^{**} = m_{\mathrm{B}*} + m_{\mathrm{N}*}$, we can write:

$$\Pr(m_{\mathrm{BB}}) = \frac{\binom{m_{\mathrm{B}*}}{m_{\mathrm{BB}}}\binom{m_{\mathrm{N}*}}{m_{\mathrm{NB}}}}{\binom{m_{**}}{m_{*\mathrm{B}}}}$$

which can be rewritten as:

$$\Pr(m_{\mathrm{BB}}) = \frac{\left(\dfrac{m_{**}!}{m_{\mathrm{BB}}!\,m_{\mathrm{BN}}!\,m_{\mathrm{NB}}!\,m_{\mathrm{NN}}!}\right)}{\binom{m_{**}}{m_{*\mathrm{B}}}\binom{m_{**}}{m_{\mathrm{B}*}}}$$

$$= \frac{\left(\dfrac{m_{**}!}{\prod_{i,j} m_{ij}!}\right)}{\binom{m_{**}}{m_{*\mathrm{B}}}\binom{m_{**}}{m_{\mathrm{B}*}}} \qquad (2)$$

The numerator is, in fact, equivalent to $\mathbf{W}$ (Equation 1), and the denominator is a constant not dependent on the structure of the O–D table. Using known properties of the hypergeometric distribution, we can derive a closed-form 'estimate' for the realized value ϕ_{BB}, based on the maximizer of the hypergeometric frequency distribution:[18]

$$\phi_{BB} = \frac{[\theta]}{m_{B*}} \qquad (3)$$

where $[\theta]$ is the maximum integer value less than or equal to:

$$\theta = \frac{(m_{*B} + 1)(m_{B*} + 1)}{(m_{**} + 2)}$$

If θ is an integer, there is a tie for the maximum probability at $[\theta]$ and $[\theta - 1]$. The mean and variance of ϕ_{BB} can also be derived.

ASSUMPTIONS, HIDDEN AND OTHERWISE

Although the hypergeometric entropy approach appears to have attractive properties (notably, the ability to derive 'estimates' of the mean and variance of the realized transition rates, ϕ_{ij}), in fact, there are several real limits to its utility.

First, by conditioning on both fixed margins, there are, in fact, no data – the only 'random variable' admitted by the fixed-margins structure is any one of the m_{ij} elements, which we have not observed. As such, the entropy method does not provide an estimate of the latent population parameter (ψ_{ij}), but predicts an unobserved random variable (ϕ_{ij}) based on the realized column and row totals in the O–D table.

Second, the conditional distribution of m_{BB} given m_{*B} is hypergeometric if, and only if, we assume equivalence of the latent rates (population parameters, ψ_{ij}) for the transition to the breeding state (i.e. $\psi_{BB} = \psi_{NB}$);[19] the most likely macrostate is always the one where the latent rates are equal (as in the preceding example). This conditional equivalency has not previously been noted in this context, and immediately limits the utility of the entropy approach. We also note that under this equivalency restriction, the latent demographic parameter ψ_{ij} would be estimated as m_{*B}/m_{**}, which is a simple binomial proportion.

Third, for simplicity, we assumed complete enumeration and closure of the population. We noted that these assumptions, also made by Willekens[16,17] in his analysis of human demography, were no more realistic than those required by other methods of analysis of aggregate data. Although formal examination of the consequences of relaxing these assumptions remains to be done, some preliminary outcomes seem likely. If the population is not completely enumerated (but is still closed), then the row and column totals are not fixed – each occasion represents a separate sample of the population. In theory we can can proceed, however, if we use the proportions of individuals in either state at each occasion (since the sum of the proportions for both margins must be 1.0). This would assume that individuals in each state are equally observable at both occasions (i.e. encounter rate is not time- or state-specific), which may be reasonable in some cases.

Relaxation of the closure assumption is potentially more problematic, since without marked individuals, we would have to make restrictive assumptions about the state-specificity of individuals entering or leaving the population between occasions (i.e. that departure from i to $i + 1$ was unbiased with respect to state, and that new individuals encountered at occasion $i + 1$ were subject to the latent rates experienced by individuals encountered at occasion i).

CONCLUSIONS – WHAT'S LEFT?

Recapture rates in some studies are extremely low, such that only sparse data are available for classic estimation techniques. For some species, marking of individuals is prohibited for either logistical or biological reasons (e.g. when capture myopathy is significant). In such cases, much of the available data consists of aggregate information on the number of individuals in a given state at a particular sampling occasion.

The initial impression is that there is little information available from the margins of an O–D table. It was previously suggested that entropy maximization might provide a means of estimating the parameters underlying the observed totals.[15,16] However, even under the limiting assumptions of closure and complete enumeration, the entropy method does not provide estimates of population parameters, but rather predictions of the realized structure of the table which are themselves based on a key assumption (i.e. equivalency of the unknown latency rates) which is unlikely to be met in practice.

So, what is left – is there any information which can be gained from the margins of the O–D table under these assumptions? In fact, the problem with the entropy approach arises because it conditions on both margins. What happens if we condition on only one margin? Any of the column or row totals can be expressed as the sum of two m_{ij} elements. For example, $m_{*B} = m_{NB} + m_{BB}$. Let $X = m_{BB} \sim B(m_{B*}, \psi_{BB})$, a binomial random variable, and let $Y = m_{NB} \sim B(m_{N*}, \psi_{NB})$, also a binomial random variable. The sum $Z = X + Y$ (i.e. $Z = m_{*B}$) is the data, and the system is completely determined by two parameters (ψ_{BB} and ψ_{NB}), which can be estimated by maximum likelihood (previously suggested by Arnason[4]). Applied to our example, let $Z = m_{*B} = 14 = (m_{BB} + m_{NB}) = (X + Y)$. Thus,

$$\Pr(Z=14) = \sum_{t=4}^{10} \Pr(X=t)\Pr(Y=14-t)$$

$$= \sum_{t=4}^{10} \frac{10!\,10!}{t!(10-t)!(14-t)!(t-4)!} \times$$

$$\times\ \psi_{BB}^{t}(1-\psi_{BB})^{10-t}(1-\psi_{NB})^{t-4}\psi_{NB}^{14-t}$$

The plot of the likelihood (Fig. 2a, 2b) indicates that the estimates are highly collinear, with high likelihood along the ridge $\psi_{BB} + \psi_{NB} = 1.4$. The likelihood has two distinct modes, at $(\psi_{BB}, \psi_{NB}) = (1.0, 0.4)$, and at $(\psi_{BB}, \psi_{NB}) = (0.4, 1.0)$; in the first case most individuals remain in their original state, while in the second the maximum possible number of individuals changes state.

Examination of the contour plot (Fig. 2b) of the likelihood shows the shape of the joint confidence regions for the two parameters. For small α, the confidence region consists of two disjoint pieces. As α increases, the overall confidence region has the standard elliptical form of many bivariate confidence regions. While fully specifying the joint 1 – α probability region is tractable with numerical methods, it is considerably more straightforward to use a Bayesian approach to derive credibility regions (treating the ψ_{ij} values as random variables, and deriving the posterior distribution for the pair of parameters integrating the likelihood after scaling to 1).

The plot of the likelihood shows clearly that information is available for the pair of parame-

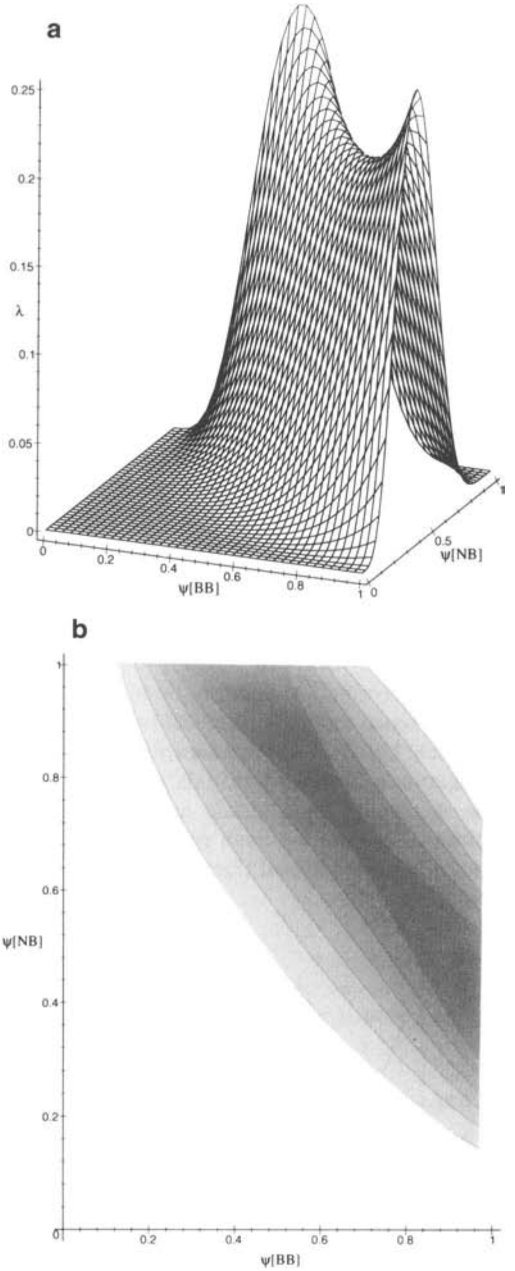

Figure 2. (a) Surface plot of likelihood for parameter pair (ψ_{BB}, ψ_{NB}), using data from numerical example in text. (b) Two-dimensional contour plot of likelihood (from Fig. 2a).

ters (ψ_{BB} and ψ_{NB}), but not for either parameter individually. Thus, although we cannot say whether a particular parameter is large or small, what we can say is that (for this example) if either one is large, then the other is likely to be small. It is also worth noting that the ridge in the likelihood consists of other values for the pair (including the values predicted by the entropy method, which are, in fact, at the minima of the ridge) that would probably not be rejected by a likelihood-ratio test.

If we adopt a Bayesian perspective, we can at least quantify the amount of information available to determine where along the ridge the true parameters lie. Under binomial sampling, the non-informative prior distribution for a proportion has a beta density where $\alpha = \beta = 0.5$ (Jeffreys' prior). The posterior distribution for a proportion uses X successes in n Bernoulli trials to update the prior, by adding X to α and $n - X$ to β. The profile $\psi_{BB} + \psi_{NB} = 1.40$ (i.e. the ridge line) of the likelihood is of a similar shape (Fig. 3) to a beta posterior distribution with parameters $\alpha = \beta = 0.8$. Thus, the margin totals provide slightly less than one Bernoulli trial's worth of information ($X = 0.8 - 0.5 = 0.3$, $n - X = 0.3$, thus $n = 0.6$) as to the location of the pair of parameter values along the ridge. Obviously, the amount of information is determined in part by the magnitude of the difference in proportions of individuals in each state between the two sampling occasions. For example, if all individuals are breeders at time i, and all individuals are non-breeders at time $i + 1$, then we have perfect information concerning the realized transition rates.

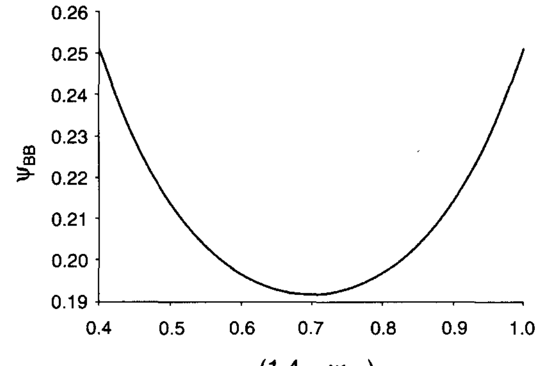

Figure 3. Plot of likelihood profile along ridge of maximum likelihood (see Fig. 2).

Nonetheless, some information about pairs of parameters from the margins of a 2 × 2 O–D table is available when the problem is recast in an estimation context, which might be useful when combined with other information (e.g. sparse recapture or recovery data). Use of additional information is well known to increase the precision of estimates in a variety of contexts.[14,20] However, relaxing the assumptions of closure and complete enumeration may prove ultimately limiting to broader utility of this approach.

ACKNOWLEDGEMENTS

The authors thank the convenors of EURING'97 for inviting this contribution. We also thank Neil Arnason (especially for access to unpublished thesis work), Jean-Dominique Lebreton and Carl Schwarz (for pointing out earlier work on estimation from aggregate data), and Jim Nichols (for broadening our interpretations of the results). Comments from Anne Viallefont and an anonymous reviewer improved the clarity and accuracy of the manuscript.

REFERENCES

1. Stearns, S.C. (1992) *The Evolution of Life Histories.* Oxford University Press, Oxford.
2. Charnov, E.L. (1990) On evolution of age of maturity and the adult lifespan. *J. Evol. Biol.*, **3**, 139–144.
3. Forslund, P. & Part,T. (1995) Age and reproduction in birds – hypotheses and tests. *TREE*, **10**, 374–378.
4. Arnason, A.N. (1971) Migration models for animal populations. PhD Thesis, University of Edinburgh.
5. Arnason, A.N. (1972) Parameter estimates from mark–recapture experiments on two populations subject to migration and death. *Res. Popul. Ecol.*, **13**, 97–113.
6. Brownie, C., Hines, J.E., Nichols, J.D., Pollock, K.H. & Hestbeck, J.B. (1993) Capture-recapture studies for multiple strata including non-Markovian transitions. *Biometrics*, **49**, 1173–1187.
7. Nichols, J.D., Brownie, C., Hines, J.E., Pollock, K.H. & Hestbeck, J.B. (1993) The estimation of exchanges among populations or subpopulations, In *Marked Individuals in the Study of Bird Population* (eds J-D. Lebreton & P.M. North), pp. 265- 279. Birkhauser-Verlag, Basel.
8. Schwarz, C.J., Schweigert, J.F. & Arnason, A.N. (1993) Estimating migration rates using tag-recovery data. *Biometrics*, **49**, 177–193.
9. Nichols, J.D., Hines, J.E., Pollock, K.H., Hinz, R.L. & Link, W.A. (1994) Estimating breeding propor-

tions and testing hypotheses about costs of reproduction with capture-recapture data. *Ecology*, **75**, 2052–2065.

10. Nichols, J.D. & Kendall, W.L. (1995) The use of multi-state capture-recapture models to address questions in evolutionary ecology. *J. Appl. Stat.*, **22**, 835–846.
11. Kalbfleisch, J.D., Lawless, J.F. & Vollmer, W.M. (1983) Estimation in Markov models from aggregate data. *Biometrics*, **39**, 907–919.
12. Kalbfleish, J.D. & Lawless, J.F. (1984) Least-squares estimation of transition probabilities from aggregate data. *Can. J. Stat.*, **12**, 169–182.
13. Lawless, J.F. & McLeish, D.L. (1984) The information in aggregate data from Markov chains. *Biometrika*, **71**, 419–430.
14. Hawkins, D.L., Han, C.P. & Eisenfeld, J. (1996) Estimating transition probabilities from aggregate samples augmented by haphazard recaptures. *Biometrics*, **52**, 625–638.
15. Jaynes, E. (1957) Information theory and statistical mechanics. *Phys. Rev.*, **106**, 105–112.
16. Willekens, F. (1977) The recovery of detailed migration patterns from aggregate data: an entropy maximizing approach. Research Report RR-77–58, International Institute for Applied Systems Analysis, Laxenberg, Austria.
17. Willekens, F., Por, A. & Raquillet, R. (1981) Entropy, multiproportional, and quadratic techniques for inferring patterns of migration from aggregate data. In *Advances in Multiregional Demography* (ed. A. Rogers). Research Report RR-81–6, International Institute for Applied Systems Analysis. Laxenberg, Austria.
18. Johnson, N.L. & Kotz, S. (1969) *Discrete Distributions*, pp. 145–146. Houghton Mifflin, Boston, MA.
19. Kagan, A.M. Linnik, Y.V. & Rao, C.R. (1973) *Characterization Problems in Mathematical Statistics*, pp. 424–425. Wiley, New York.
20. Burnham, K.P. (1993) A theory for combined analysis of ring recovery and recapture data. In *Marked Individuals in the Study of Bird Population* (eds J-D. Lebreton & P.M. North), pp. 199- 214. Birkhauser-Verlag, Basel.