SSC Annual Meeting, May 2019
Proceedings of the Survey Methods Section

ESTIMATING POPULATION ABUNDANCE USING AUXILIARY POPULATION COUNTS

Matthew R. P. Parker and Laura L.E. Cowen[1]

ABSTRACT

We develop a new method for estimating population abundance for notoriously difficult to count populations. This is made possible using an easy to count auxiliary population with a known link to the target population. The new models require population specific domain knowledge, and can be easily applied using existing software to estimate population abundances where offspring are more readily counted than the adults. Applications could include parent/offspring, predator/prey, and symbiotic relations in which one population is more feasible to study than the other.

KEY WORDS: Binomial thinning, Hidden Markov model, N-Mixture model, Replicated count data.

RÉSUMÉ

Nous développons une nouvelle méthode d'estimation de l'abondance de la population pour les populations notoirement difficiles à dénombrer. Ceci est rendu possible en utilisant une population auxiliaire facile à dénombrer avec un lien connu à la population cible. Les nouveaux modèles nécessitent une connaissance du domaine spécifique à la population et peuvent être facilement appliqués à l'aide de logiciels existants. pour estimer l'abondance de la population où la progéniture est plus facilement dénombrée que les adultes. Les applications pourraient comprendre les relations parent/ enfant, prédateur/proie et symbiotiques dans lesquelles une population est plus facile à étudier que l'autre.
MOTS CLÉS : Amincissement binomial; Model de Markov caché; Modèle de mélange; Données de comptage par réplique

1 INTRODUCTION

Statistical methods of population abundance estimation play a critical role in species conservation efforts. Knowledge of current population levels, as well as both current and past population trends can be used to inform important policy and decision making processes. Accurate estimates of wildlife population abundances and trends are often essential for understanding ecosystems and for managing wildlife.

Developing less labour-intensive methods of producing accurate population estimates is an important goal of population biologists. Advancements in computing technologies are helping to realize this goal by allowing for more computationally intensive methods to be studied and applied. Evaluating the applicability and accuracy of these computationally intensive methods is vital in assessing their strengths and their shortcomings. This can be accomplished through simulation studies, and through comparisons against current methods. New statistical techniques should aim to improve on the established methods in both accuracy and precision of estimates, while remaining cognisant of the importance of minimizing the cost of data collection.

N-mixture modelling methods improve on estimates extrapolated from estimated mean abundances in two distinct ways. First, N-mixtures improve on accuracy by modelling across sites to reduce unexplained variability, and second, they improve on precision by assuming parametric distributions for population abundance and dynamics. N-mixture models are also less labour intensive, and ultimately less expensive than mark-recapture methods, as they do not necessitate batch marking and repeated captures. We develop here an extension of N-mixture models which enables the use of a separate, auxiliary population in estimating the abundance of a particular target population of interest. These auxiliary population models allow researchers to use data from easy to observe populations in order to estimate abundances for difficult or expensive to count populations of interest.

[1]Matthew Parker and Laura Cowen, Mathematics and Statistics, University of Victoria, PO BOX 1700 STN CSC, Victoria, BC, Canada V8W 2Y2, mrparker@uvic.ca, lcowen@uvic.ca

There are many potential applications for these auxiliary population models; however, they require specific domain knowledge to recognize and implement. Examples of population relations which can be leveraged include predator/prey, symbiotic, and parent/offspring relationships. Species can be difficult to count for numerous reasons, for example living in underwater or remote habitats, or having difficult to distinguish individuals which would increase the odds of double counting. The ability to link an auxiliary population which is more easily counted with one which would otherwise be difficult to count improves this situation. We focus on the parent/offspring relation here as it is the most readily implemented. This is because the associated model can be formulated to allow the use of existing software for model parameter fitting.

An example of a group of populations which can be studied using the auxiliary population models are marine mammals such as the Irish Grey Seals (Ó Cadhla et al., 2013). Irish Grey Seals spend the majority of their adult lives submerged, and leave their offspring on land at haul-out sites after birth. This makes obtaining population counts of the offspring significantly easier compared to the adults (for example, using aerial photography of the haul-out sites). Another example are burrow-nesting seabird populations, many species of which are nocturnal while on land, making individuals challenging to count (Major and Chubaty, 2012). One such seabird species are the Ancient Murrelet, who additionally spend most of their adult lives out on open water, increasing the challenge. However, the Ancient Murrelet chicks are highly precocial, rushing on mass from their burrows on land, out to the open ocean during the hatching season (Gaston, 1990). Thus the adult birds are difficult to obtain accurate counts of, while the chicks are relatively easy to count.

In the following sections we develop the auxiliary population model to estimate population size using data from an auxiliary population. We do so under the framework of layered hidden Markov models, and show that the model can be applied using existing model fitting software.

2 N-MIXTURE MODELS

N-mixture models were originally developed to produce population estimates using only replicated count data on closed-populations where there are no births/deaths or immigration/emmigration (Royle, 2004). N-mixture models are being widely used for estimating population sizes and trends for many different groups of animals, such as birds (Lyons et al., 2012), reptiles (Ward et al., 2017), and large mammals (Belant et al., 2016). For N-mixture models, counts are replicated at a number of sites R and number of sampling occasions M, giving a total of RM observations (Royle, 2004). Sites are assumed to be independent, and spatially distinct. One of the key features of N-mixture models is their flexibility in model form through distributional choices. They accommodate a variety of detection characteristics, and allow for parameter dependency on covariates. As discussed by Dail and Madsen (2011), who extend these models to open-populations, population dynamics can be modelled in many ways, since different distributional choices for apparent recruitment and apparent survival are viable. These models are computationally very expensive, as they treat population abundance as a nuisance variable, requiring summations over very large ranges to integrate abundance from the likelihood function. Due to these large summations, computation time grows rapidly with increasing population abundances. Alternatively, N-mixtures can be framed as hidden Markov models (Cowen et al., 2017), however some computational issues remain. Barker et al. (2018) raised concerns regarding identifiability of abundance (N) and probability of detection (p) in closed-population N-mixture models, and we address these concerns in the discussion.

The observed counts are assumed to be the result of a binomial thinning on the current population (Fernández-Fontelo et al., 2016) with detection probability p_{it} for site i and time t. Other thinning operators can be considered, however we focus on the binomial thinning operator as a simple and robust detection mechanism. In open-population N-mixture models, population dynamics are accounted for by considering the population as two distinct groups. The population abundance at site i and time t is given by $N_{it} = S_{it} + G_{it}$ (Nichols et al., 2000), where S_{it} is the number of individuals at site i who have survived from time $t-1$ to t, and G_{it} is the number of individuals recruited for site i at time t. There are many choices involved in modelling S_{it} and G_{it}, and those choices are heavily dependent on the population under study. We will use the standard choices here, and note that justifications for these choices are case study dependent. We model S_{it} with apparent survival probability ω_{it} as Binomial(N_{it-1}, ω_{it}), and G_{it} as Poisson(γ_{it}) where γ_{it} is the recruitment rate into site i at time t. Initial abundance for each site i is modelled as a Poisson random variable with parameter λ_i, representing the initial mean abundance. Population abundance N_{it} is assumed to have the Markov property, so that $P[N_{it} = k|N_{i1}, N_{i2}, ..., N_{it-1}] = P[N_{it} = k|N_{it-1}]$. Each of the distributions mentioned are a model choice, and other choices do exist. For example a population

that shows over-dispersion of counts may be better modelled with a negative binomial distribution rather than a Poisson distribution, however, there has been discussion in the literature around problematic behaviour of the negative binomial (see for example Kéry 2017). The model parameters (λ_i, ω_{it}, γ_{it}, and p_{it}) can be given additional covariate structure using appropriate link functions (such as the logit link for probability parameters: $logit(p) = \beta_0 + \sum_j \beta_j x_j$, where x_j are the covariates) (Dail and Madsen, 2011). A simple model of this form would have each parameter constant over time and across sites ($\lambda_i = \lambda$, $\omega_{it} = \omega$, $\gamma_{it} = \gamma$, and $p_{it} = p$).

3 AUXILIARY POPULATION MODELS

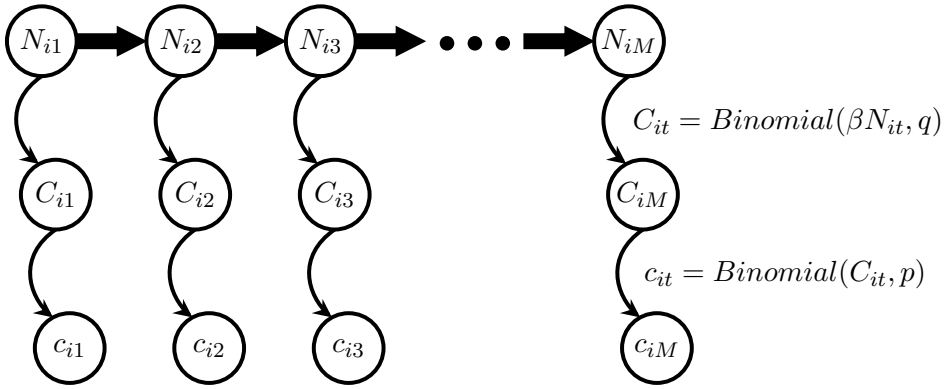
Data for the auxiliary population models are in the same form as for N-mixture models (replicated counts over sites and sampling occasions) except that the counts are of the auxiliary population rather than the target population of interest. A population link needs to be established between the auxiliary and target populations. The population link can take many forms, and relies on expert knowledge specific to the two populations at hand. For example, Irish Grey Seals have a litter size of 1, so that a particular female will give birth to either zero or one pup per breeding season (Ó Cadhla et al., 2013). This lends itself to considering a Bernoulli population link: (number of pups) $\sim \sum_{\text{number of females}} Bernoulli(\kappa)$, where κ is the unknown probability of a female producing a pup. We will be focussing on the adult/offspring population link, however, we note that the methods can be easily extended to other target/auxiliary populations through the judicious choice of model distributions and parameters based on specific domain knowledge of the population link.

In the case of an adult population with a maximum brood size, the link between the offspring and the adult breeding pairs is the maximum brood size β. Consider a population N_{it} of breeding pairs. Each breeding pair from N_{it} will have a potential brood size of β; however, not all of those offspring will be born, cared for, and able to survive to become observable. If we let q be the probability for each potential offspring to survive and become observable, and if we make the assumption that offspring survival is independent, then the auxiliary population can be formulated as $C_{it} = \sum_{j=1}^{\beta N_{it}} Bernoulli(q)$. This is equivalent to a binomial thinning, $C_{it} = q \circ (\beta N_{it})$. The observed counts are, by the N-mixture formulation, also due to a binomial thinning. If we let c_{it} be the observed auxiliary counts, and p the probability of detection, then $c_{it} = p \circ C_{it} = p \circ q \circ (\beta N_{it}) = pq \circ (\beta N_{it})$. Thus the probability of detection p and auxiliary survival probability q (related to productivity) are confounded. By using this auxiliary population we lose the ability to obtain estimates of probability of detection, and instead our estimates will be of the inseparable product $\pi = pq$. We will refer to this π as the auxiliary probability of detection, not to be confused with the probability of detection p.

The entire process can be viewed as a layered hidden Markov model (Oliver et al., 2004) with two hidden layers. Under this view, each site produces an independent unobserved Markov chain of length equal to the number of sampling occasions, which taken together form the first layer. Transitions in the first layer occur along sites, through the open N-mixture population dynamics for the adult breeding pairs, $N_{it} = S_{it} + G_{it}$. The first layer has symbols equal to the unobserved target population abundance (N_{it}). The second layer is connected to the first by the population link from target population to unobserved auxiliary population (C_{it}) as symbols. Under the adult/offspring relation, this link is a binomial thinning with thinning probability q. The third and final layer of the hidden Markov chain is formed by detection thinning on the unobserved auxiliary population. This produces observed auxiliary counts (c_{it}). The second and third layers are completely self-disconnected, with no intralayer connections. The layered hidden Markov process is illustrated for a single site in Figure 1. Since sites are assumed independent, the process is identical for each site.

For the proposed model, the population parameters being estimated are for the target population, and not for the auxiliary population (as the auxiliary population represents surrogate counts for the breeding pairs, and not a population in the sense of open-population N-mixture models). This can be seen easily from the layered hidden Markov model in Figure 1, where the only connections across sampling occasions occur in the first layer. The detection thinning process $\pi \circ \beta N_{it}$ is still binomial in nature, and directly analogous to the detection thinning process of the usual N-mixture models. This allows us to use the standard N-mixture model fitting software `unmarked` (Fiske and Chandler, 2011) for our model fitting, being careful to correctly identify π from the fitted model parameters. When using the auxiliary population to estimate breeding pairs using the standard N-mixtures software, the estimates will be of βN_{it} rather than N_{it}, and so we will be over-estimating by a factor of β. To correct for this, we simply divide the estimates by β.

Figure 1: Layered hidden Markov chain for adult breeding population N_{it} with maximum brood size β. Illustrates the link to the auxiliary offspring population C_{it}, and to the observed offspring counts c_{it} for a single site i. Transitions between N_{it-1} and N_{it} are through the open N-mixture population dynamics for the adult breeding pairs, $N_{it} = S_{it} + G_{it}$.



4 METHODS

We estimated the total breeding population in two stages. First we used N-mixture models to estimate annually the number of offspring (the auxiliary population). From those estimates we then used the adult/offspring population link to convert to abundance estimates of adult breeding pairs within observation sites.

Open-population N-mixture models implemented using the R package `unmarked` (Fiske and Chandler, 2011), are maximum likelihood parameter estimates. We can allow the model parameters to have various covariate structures for example they can vary by location or time; however, one must consider parameter redundancy in a fully time-vary model. Bayesian information criterion (Wit et al., 2012), BIC, can be used to rank the models, and this ranking can be used to select the best model. In calculation of BIC, we suggest sample size is conservatively taken to be the number of sites rather than the product of number of sites with number of time replicates. The conservative sample size was chosen given the following. Each of the RM observations contributes to the degrees of freedom available for model fitting for parameters and their covariates that gain information from observations alone (p and any associated covariates for example). However, since the population dynamics parameters gain information only from pairs of adjacent sampling occasions, each site only provides $M - 1$ degrees of freedom for the dynamics parameters and their covariates (γ and ω for example). As well, the parameter λ gains information only from the first observation from each site, limiting the degrees of freedom available for estimating λ and any associated covariates to the number of sites R. The total degrees of freedom available for estimating model parameters, covariates, and error is thus bounded below by $R(M - 1)$. Conservatively, the total covariates and parameters in the model should number less than $R(M - 1)$, with the extra condition that λ and its covariates should number less than R.

The fitted models resulted in estimates of the auxiliary population abundance for each site and each year, $\hat{C}_{it}$. These were then divided by the auxiliary population link $\beta = 2$, resulting in the estimated number of breeding pairs within the sample areas ($\hat{N}_{i,t}$).

If one had estimates of the actual colony areas, the estimated numbers of breeding pairs $\hat{N}_{T,t}$ in the whole colony can be calculated using an area expansion of the estimated number of breeding pairs in the sampled locations. One would have to assume that the sampled area abundances ($\hat{N}_{i,t}$) are proportionally representative of the total colony abundances. However, our method allows for different population densities at each site, which should produce more accurate area expansion estimates of total colony abundance than using a single colony mean (as is done in typical area expansion estimators; Lemon 2007).

5 SIMULATION STUDY

We performed a simulation study to illustrate the viability of making population abundance estimates using the auxiliary population models, and showed that the estimated populations closely match the actual populations.

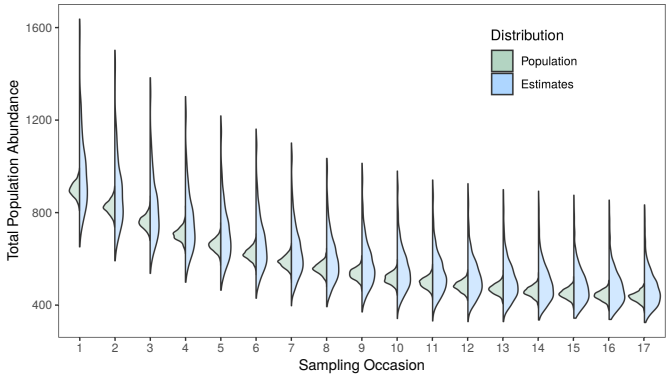


Figure 2: Simulation study investigating the validity of using an auxiliary offspring population to estimate an adult population using a maximum brood size of $\beta = 2$ as a population link. Left hand distributions show the actual breeding pair population size distributions, right hand distributions show the distributions of the β-corrected N-mixture estimates of breeding pairs.

First we chose a set of population parameters. We set the number of sites $R = 6$, number of sampling occasions $M = 17$ (one/year), initial mean abundance per site $\lambda = 150$, mean recruitment $\gamma = 10$, survival probability $\omega = 0.85$, and probability of detection $p = 0.8$. We also chose $\beta = 2$ (after a population that had the potential to produce 2 and only 2 offspring as is the case with the Ancient Murrlet example; Gaston 1990), and to use a productivity of $\alpha = 1.5$, but note that since q and p are confounded, this choice of α does not impact the simulation results.

The effect of increasing sites and/or sampling occasions would be to decrease the variability in the estimates as more data is available, and the effect of changing β is to simply change the scale factor for the estimates. For a very different population link, such as a complex predator/prey relation, a separate simulation study would need to be conducted to verify the validity of using the proposed auxiliary population in making target population estimates. The simulation steps are listed below:

1. Generate ground truth data:
   (a)
   - Generate breeding pairs from Poisson(λ), for all sites i and $t = 1$
   - Generate breeding pairs for $t > 1$ using Binomial(ω)+Poisson(γ)

   (b) Generate auxiliary population $C_{it} = \sum_{k=1}^{N_{it}}(Bin(\beta, q = \alpha/\beta))$. In this way, each adult breeding pair produces 0, 1, or 2 osffspring, with mean offspring per pair of α.
2. Generate observational data:
   - Generate observed offspring counts from the potential offspring using $c_{it} = Bin(C_{it}, p)$.
3. Auxiliary population model fitting:
   - Fit open population N-mixture model to the observed offspring counts
   - Switch from auxiliary to breeding pair estimates by correcting by the factor β.
4. Repeat simulation steps 1000 times.

The simulation results, shown in Figure 2, show that the trend in population size is closely matched between the ground truth and the estimates. The distributions have similarly located modal peaks, with the estimated distributions having larger variability than the ground truth distributions.

6 DISCUSSION

We developed a novel layered hidden Markov model that can be employed in situations for which a link to an auxiliary population exists. The choice of model distributions, as well as parameter covariates, allow an enormous

variety of populations to be studied. One possible example is the breeding population of grey seals in Ireland (Ó Cadhla et al., 2013). The adult grey seals are difficult to count, whereas the pups remain near haul-out locations while young, making them much easier to count from aerial photos of the haul-out sites. In this case the auxiliary population would be the pups, where the known link is that there is at most one pup per female seal per breeding season.

Our simulation study investigated the validity of using the auxiliary population model for a well understood auxiliary population. The results show strong agreement between the distributional modes of the true population abundances and the estimated population abundances. The estimated population distributions are right-skewed, with both the variability and the skewness decreasing with increasing sampling occasion. It is of interest that the variability in the estimates reduces as the sampling occasion increases, so that the more recent estimates are the most precise. This is strong evidence that the number of sampling occasions should be as large as possible to ensure the best possible estimates of population abundance, and that even with a large number of sampling occasions, the earlier abundance estimates will be intrinsically less precise. This can be understood from considering the population flow through time. Later sampling occasions draw much informative data from every previous sampling occasion; however, the earlier occasions are products of only relatively few population updates, lending to higher variability in the possible initial states.

Barker et al. (2018) raised concerns about identifiability of abundance in N-mixture models. It is important to note that the Barker paper dealt specifically with the closed-population formulation of N-mixture models, whereas we deal with the open-population formulation. Barker et al. (2018) showed that if the probability of detection p is allowed to vary with time, then the models become over-specified. We do not consider in this paper any models with time varying p, and so this is not a concern for us. Barker et al. (2018) make note that data generated from an N-mixture process can be indistinguishable from data generated from a model where N is not identifiable (or not even a parameter of the model). Barker et al. (2018) considered the limiting case of a $Binomial(N, p)$ random variable, which converges to $Poisson(Np)$ when $N \to \infty$, $p \to 0$ and Np is kept constant. While this situation does produce unidentifiable N and p, this is not a reasonable situation in cases where probability of detection is not expected to be very small. In our case we assume our auxiliary probability of detection π to be very high. Further, Barker et al. (2018) indicate that the issues are most problematic when data quality is low, or when the count data is sparse. Users must be wary of their data and assess these issues on a case by case basis.

N-mixture methods provide a powerful tool for simultaneously estimating abundance and probability of detection. Open-population formulations allow estimation of abundance trends over time, including estimates of apparent survival and apparent recruitment. Novel use of auxiliary populations allow N-mixture models to be applied to populations for which it is otherwise difficult to obtain accurate counts, at the cost of losing identifiability of probability of detection. Comparing the results of N-mixture models to currently established methods is a vital component of validating these new modelling techniques, and uncovering their strengths as well as their weaknesses.

ACKNOWLEDGEMENTS

Analyses were run on Westgrid/Compute Canada with assistance from Belaid Moa. This work was partly funded by NSERC grant 327025 to LC. MP would like to acknowledge the Visual and Automated Disease Analytics (VADA) Program for funding support during his MSc program.

REFERENCES

Barker, R. J., Schofield, M. R., Link, W. A., and Sauer, J. R. (2018). On the reliability of N-mixture models for count data. *Biometrics* **74,** 369 – 377.

Belant, J. L., Bled, F., Wilton, C. M., Fyumagwa, R., Mwampeta, S. B., and Beyer, D. E. (2016). Estimating lion abundance using N-mixture models for social species. *Scientific Reports* **6,** 35920.

Cowen, L. E., Besbeas, P., Morgan, B. J. T., and Schwarz, C. J. (2017). Hidden Markov models for extended batch data. *Biometrics* **73,** 1321–1331.

Dail, D. and Madsen, L. (2011). Models for estimating abundance from repeated counts of an open metapopulation. *Biometrics* **67,** 577–587.

Fernández-Fontelo, A., Cabaña, A., Puig, P., and Moriña, D. (2016). Under-reported data analysis with INAR-hidden Markov chains. *Statistics in Medicine* **35,** 4875–4890.

Fiske, I. and Chandler, R. (2011). unmarked: An R package for fitting hierarchical models of wildlife occurrence and abundance. *Journal of Statistical Software* **43,** 1–23.

Gaston, A. J. (1990). Population parameters of the Ancient Murrelet. *The Condor* **92,** 998–1011.

Kéry, M. (2017). Identifiability in N-mixture models: A large-scale screening test with bird data. *Ecology* **99,** 281–288.

Lemon, M. J. F. (2007). East Limestone Island Ancient Murrelet Colony Survey, June 2006. In Gaston, A. J., editor, *Laskeek Bay Research 15*, pages 67–86. Laskeek Bay Conservation Society, Queen Charlotte City, B.C.

Lyons, J. E., Royle, J. A., Thomas, S. M., Elliott-Smith, E., Evenson, J. R., Kelly, E. G., Milner, R. L., Nysewander, D. R., and Andres, B. A. (2012). Large-scale monitoring of shorebird populations using count data and N-mixture models: Black Oystercatcher (*Haematopus bachmani*) surveys by land and sea. *The Auk* **129,** 645–652.

Major, H. and Chubaty, A. (2012). Estimating colony and breeding population size for nocturnal burrow-nesting seabirds. *Marine Ecology Progress Series* **454,** 83–90.

Nichols, J. D., Hines, J. E., Lebreton, J. D., and Pradel, R. (2000). Estimation of contributions to population growth: A reverse-time capture-recapture approach. *Ecology* **81,** 3362–3376.

Ó Cadhla, O., Keena, T., Strong, D., Duck, C., and Hiby, L. (2013). Monitoring of the breeding population of grey seals in Ireland, 2009 – 2012. In *Irish Wildlife Manuals No. 74.* National Parks and Wildlife Service, Department of the Arts, Heritage and the Gaeltacht, Dublin, Ireland.

Oliver, N., Garg, A., and Horvitz, E. (2004). Layered representations for learning and inferring office activity from multiple sensory channels. *Computer Vision and Image Understanding* **96,** 163–180.

Royle, J. A. (2004). N-mixture models for estimating population size from spatially replicated counts. *Biometrics* **60,** 108–115.

Ward, R. J., Griffiths, R. A., Wilkinson, J. W., and Cornish, N. (2017). Optimising monitoring efforts for secretive snakes: a comparison of occupancy and N-mixture models for assessment of population status. *Scientific Reports* **7,** 18074.

Wit, E., Heuvel, E., and Romeijn, J. (2012). 'All models are wrong...': an introduction to model uncertainty. *Statistica Neerlandica* **66,** 217–236.