Journal of Biometrics & Biostatistics
Jiang and Cook, J Biom Biostat 2018, 9:3
DOI: 10.4172/2155-6180.1000406
Research Article Open Access

Cost-effective Design of Growth Studies with Aggregation and Tracking

Shu Jiang* **and Richard J Cook**
Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada

Abstract

Studies of the development and growth of organisms are often conducted in laboratories where organisms maintained in tanks are examined repeatedly over time. Collection and recording of cross-sectional aggregate count data on stage occupancy is both less expensive and administratively more convenient than tracking the stages of each organism over time. In such settings tank to tank variation must also be taken into account as growth rates may be more similar among organisms within the same tank than for those in different tanks. We consider the cost effect design of a prospective developmental study of organisms based on a marginal Markov model which deals with between tank variation and within tank dependence. We develop a flexible design in which some tanks provide repeated cross-sectional aggregate data, and other tanks provide serial responses through tracking individuals. We assess the relative efficiency of aggregate and individual-level longitudinal data. The optimal cost-effective design is shown to depend on whether primary interest lies in transition intensities or associated cluster-level covariate effects.

Keywords: Aggregate data; Clustered data; Design; Heterogeneity; Interval censoring; Markov process; Multistate model

Introduction

In many growth and developmental studies organisms are arranged in tanks or other types of enclosure and repeatedly examined over time to acquire information on developmental stages. Examples include studies of plant growth, metamorphosis of fish or amphibians, or small arthropods [1-3]. The maturation process can usually be naturally modelled using multistate processes.

In some contexts it can be difficult to identify individual organisms. In studies of hornworms for example [4], the larvae are both mobile and indistinguishable. Gouno et al. also reported on a growth study of arabidopsis thaliana where the data are recorded in aggregated form [1]. In such cases the available data consists only of the counts of the number of organisms in the different developmental stages at each assessment time. This form of aggregation is also common when the only available data are published in tabular form.

There has been much discussion on methods for dealing with aggregate data. MacRae first introduced the nonlinear generalized least squares and briefly mentioned methods for exact maximum likelihood for aggregate data [5]. Kalbfleisch and Lawless introduced a weighted least squares approach for estimating transition intensities from aggregate data [6]. We develop build upon a likelihood approach in this paper and consider strictly progressive Markov processes appropriate for growth data. Computational challenges may arise as the number of assessment times and individuals increase, so we propose composite likelihood as an appealing alternative in such cases [7]. When organisms are organized in different tanks (i.e., clusters), tank-to-tank variation must be taken into consideration. Jiang and Cook [8] use composite likelihood to handle such data based on both marginal methods with robust variance estimation, and a random effects model.

The focus of this paper is on the optimal design for studies involving multiple tanks/clusters; we adopt the marginal approach of Jiang and Cook [8] for aggregate data. In some contexts tracking of individuals is possible but incurs a cost [9]. We also consider cost-effective design by addressing the situation in which some tanks contain organisms to be tracked individually over time, while other tanks may be designated to provide only aggregate counts in the different developmental stages at different assessment times. Sample size calculations are derived and cost-effective allocation of tanks to these two observation schemes is also considered.

The remainder of this paper is organized as follows. In the next section we define notation and describe a composite likelihood for clustered Markov processes which we use to characterize growth of individual organisms and to accommodate dependence in progression rate within tanks. Large sample results and methods of inference for both tracking and aggregate observation schemes are given. Sample size criteria are developed to meet design objectives and cost-effective allocation of tanks to the tracking and aggregate observation schemes are developed and discussed along with some simulation results. Concluding remarks are then made.

Notation and Likelihood

Composite likelihood for clustered panel data

We consider strictly progressive multistate models suitable for studying maturation processes. Suppose that observations are made on a group of individuals who act independently of one another, with each individual passing through states according to a multistate process with state space $\{1, 2,\ldots, K\}$. We let $Z_j(t)$ denote the state occupied by individual j at time t and $\{Z_j(s), 0<s\}$ be the multistate process.

Let $\mathscr{H}_j(t)=\{Z_j(s),\ 0 \le s<t\}$ denote the history of the process for individual j at time t and let

$$\lambda_k(t \mid \mathscr{H}_j(t)) = \lim_{\Delta t \downarrow 0} \frac{P(Z_j(t+\Delta t^-) = k+1 \mid Z_j(t^-) = k; \mathscr{H}_j(t))}{\Delta t} \qquad (1)$$

denote the $k \to k+1$ transition intensity, $k=1,\ldots, K-1$. For Markov processes the intensity does not depend on the history in which case we write the left hand side of eqn. (1) as $\lambda_k(t)$. Given a $K \times K$ transition intensity matrix $\Lambda(t)$ with $(k, k+1)$ entry $\lambda_k(t)$, diagonal entry $-\lambda_k(t)$, for

***Corresponding author:** Shu Jiang, Department of Statistics and Actuarial Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada, Tel: +15198884567; E-mail: s64jiang@uwaterloo.ca

Received August 18, 2018; **Accepted** August 27, 2018; **Published** August 31, 2018

Citation: Jiang S, Cook RJ (2018) Cost-effective Design of Growth Studies with Aggregation and Tracking. J Biom Biostat 9: 406. doi: 10.4172/2155-6180.1000406

k=1,…, K−1 and zeros elsewhere, by product integration [8] the $K \times K$ transition probability matrix is

$$P(\mathrm{s,t}) = \prod_{(\mathrm{s,t}]}\{1+\Lambda(u)du\} \tag{2}$$

with (k, l) entry $P(Z(t)=l|Z(s)=k)$ for k ≤ l. If observations are made at times $0=a_{j0}<a_{j1}<\cdots<a_{jR_j}$ for individual j, panel data denoted by $\{(Z_j(a_{jr}), a_{jr}), r=1, 2, ..., R_j\}$ are obtained. Kalbfleisch and Lawless [10] develop a Fisher-scoring algorithm for maximizing the likelihood which is implemented in the R function *msm* [11].

Now consider a setting with I tanks of organisms with n_i individuals in tank i, i=1,…, I. Let $0=a_{i0}\leq\ldots\leq a_{iR_i}$ denote the common assessment times for all j=1,…, J individuals in tank i, i=1,…, I. Diao and Cook [12] formulate a copula-based model for correlated Markov processes which accommodate dependence between processes within clusters and retain the marginal Markov property for each process. With progressive processes, within-cluster dependence can be modeled in terms of sojourn or state entry times through copula functions. We consider a class of Archimedian copulas [13] of the form $C\left(u_1,u_2,\ldots,u_{n_i};\eta\right)=G^{-1}\left(G(u_1;\eta)+\ldots+G(u_2;\eta)\right)$ where G : [0, 1] → [0,∞) is a continuous, strictly decreasing and convex generator function with dependence parameter η and $G(1;\eta)=0$ [13]. To induce a dependence, we select the first transition time (i.e., the entry time to state 2) and note that a dependence is induced within clusters for the subsequent state entry times. Specifically, we let T_{ij2} denote the entry time to state 2 for individual j in tank i and $T_{i2}=(T_{i12},\ldots, T_{ini2})'$ denote the vector of all state 2 entry times in tank i, i=1,…, I. We adopt the Clayton copula [13] and use Kendall's τ as a measure of dependence where

$$\tau = 1+4\int_0^1 \frac{G(u;\eta)}{G'(u;\eta)}du.$$

We formulate the joint survivor function for T_{i2} by linking all marginal survivor functions $\mathscr{F}_{ij}(t_{ij2};\lambda_1)=P(T_{ij2}>t_{ij2})\exp(-\lambda 1 t_{ij2})$ via the Clayton copula as

$$\mathscr{F}(t_{i2};\lambda_1,\eta)=(\mathscr{F}(t_{i12};\lambda_1)^{-\eta}+\ldots+\mathscr{F}(t_{ini2};\lambda_1)^{-\eta}-(\eta_i-1))^{1/n}.$$

Diao and Cook [12] describe an alternative approach where the association in the absorption times is modeled instead of earlier state entry or sojourn times, but the principle of inducing a dependence between multistate processes within a cluster by linking a particular time is in the same spirit.

Consider the case with a cluster level covariate xi, i=1,…, I and let

$$\lim_{\Delta t\downarrow 0}\frac{P(Z_{ij}(t+\Delta t^-)=k+1\,|\,Z_{ij}(t^-)=k;\mathscr{H}_j(t))}{\Delta t}=\lambda_k\exp(x_i'\beta)$$

which we denote more compactly as λ_{ik}, k=1,…, K − 1. If $\alpha_k=\log\lambda_k$, k=1,…, K − 1, $\alpha=(\alpha_1,\ldots,\alpha_{K-1})'$ and $\beta=(\beta'_1,\ldots,\beta'_{K-1})'$, we then let $\theta=(\alpha',\beta')'$. Under a working independence assumption and a panel observation scheme (i.e., with individual tracking) the composite likelihood is

$$L_1(\theta)\propto\prod_{i=1}^{I}\prod_{r=1}^{R_i}L_{1ir}(\theta) \tag{3}$$

where,

$$L_{1ir}(\theta)\propto\prod_{i=1}^{n_i}\prod_{k\leq l}P(Z_{ij}(a_{ir})=l\,|\,Z_{i,j}(a_{i,r-1})=k,x_i;\theta) \tag{4}$$

and $Z_{ij}(t)$ is the state occupied by individual j in tank i at time t. We then define

$$S_{1ir}(\theta)=\sum_{j=1}^{n_i}\sum_{k\leq l}\frac{\partial\log P(Z_{i,j}(a_{ir})=l\,|\,Z_{i,j}(a_{i,r-1})=k,x_i;\theta)}{\partial\theta} \tag{5}$$

and we let $S_{1i}(\theta)=(S_{1i1}(\theta),\ldots,S_{1iR_i}(\theta))$ be a $p\times R_i$ matrix. We let $\hat{\theta}$ denote the solution to $S_1(\theta)=\sum_{i=1}^{I}\sum_{r=1}^{R_i}S_{1ir}(\theta)=0$.

A robust sandwich variance estimate is required to ensure valid inference under this working independence assumption. Under standard regularity conditions [14]

$$\sqrt{I}(\hat{\theta}-\theta)\rightarrow N(0,\mathscr{A}_1^{-1}(\theta)\mathscr{B}_1(\theta)\mathscr{A}_1^{-1}(\theta)) \tag{6}$$

where, $\mathscr{A}_1(\theta)=-E\left\{\sum_{r=1}^{R_i}\partial S_{1ir}(\theta)\right\}$ and $\mathscr{B}_1(\theta)=E\left\{S_{1i}(\theta)S_{1i}'(\theta)\right\}$.

The matrices $\mathscr{A}_1(\theta)$ and $\mathscr{B}_1(\theta)$ can be estimated empirically by

$$\hat{A}_1=-I^{-1}\sum_{i=1}^{I}\sum_{r=1}^{R_i}\frac{\partial S_{1ir}(\theta)}{\partial\theta'}\Big|_{\theta=\hat{\theta}}$$

and

$$\hat{B}_1=-I^{-1}\sum_{i=1}^{I}S_{1i}(\theta)S_{1i}'(\theta)\Big|_{\theta=\hat{\theta}}$$

and tests regarding elements of θ or associated 95% confidence intervals are constructed based on the estimated covariance matrix $\hat{A}_1^{-1}\hat{B}_1^{-1}\hat{A}_1^{-1}$.

Composite likelihood for correlated aggregate data

Under the Markov property for a single individual process considered on its own, the stage occupied at time a_{ir} only depends on the stage occupied at $a_{i,r-1}$. With aggregate data we only need to consider two consecutive assessment times, and the joint distribution is built up as a product of the conditional probabilities. However, as the number of assessment times and individuals per tank increase, the likelihood becomes computationally challenging. That motivates use of a composite likelihood approach where we adopt a working independence assumption and consider contributions from the marginal frequency data observed at each time point as arising independently from the data at different time points from the same tank.

Here we consider data from the baseline assessment to each of the follow up assessment times. Thus for two assessment times a_{i0}=0 and a_{ir}, the missing information in the aggregate data are $N_i(a_{ir})$, the vector containing all counts $N_{i1l}(a_{ir})=\sum_{j=1}^{n_i}I(Z_{ij}(a_{ir})=l\,|\,Z_{ij}(a_{i0})=1)$ for l=1, ...,K and i=1,…, I. With a strictly progressive process and $P(Z_{ij}(a_{i0})=1)=1$ and we let $N_{i1l}(a_{ir})=M_{il}(a_{ir})$ corresponds to the number of individuals occupying state l at time a_{ir} in tank i. We can then obtain the composite likelihood

$$L_2(\theta)\propto\prod_{i=1}^{I}\prod_{r=1}^{R_i}L_{2ir}(\theta) \tag{7}$$

and

$$L_2(\theta)\propto P(M_i(a_{ir})\,|\,M_{i1}(a_{i0})=n_i,x_i;\theta)$$

where $M_i(a_{ir})=(M_{i1}(a_{ir}),\ldots,M_{ik}(a_{ir}))'$.

Robust sandwich variance estimates are adopted to ensure valid inference. The estimating equations corresponding to the composite likelihood is

$$S_2(\theta)=\sum_{i=1}^{I}\sum_{r=1}^{R_i}S_{2ir}(\theta)$$

Citation: Jiang S, Cook RJ (2018) Cost-effective Design of Growth Studies with Aggregation and Tracking. J Biom Biostat 9: 406. doi: 10.4172/2155-6180.1000406

where $S_{2ir}(\theta)=\partial\log L_{2ir}(\theta)/\partial$. Since the contributions of eqn. (7) are valid likelihood contributions $E\{S_2(\theta)\}=0$ and the solution is denoted by $\tilde{\theta}$. Again, under standard regularity conditions [14], we can then construct the robust sandwich variance as

$$\sqrt{I}(\tilde{\theta}-\theta)\rightarrow N(0,\mathscr{A}_2^{-1}(\theta)\mathscr{B}_2(\theta)\mathscr{A}_2^{-1}(\theta)) \quad (8)$$

where, $\mathscr{A}_2(\theta)=-E\left\{\sum_{r=1}^{R_i}\partial S_{2ir}(\theta)\right\}$ and $\mathscr{B}_2(\theta)=E\left\{S_{2i}(\theta)S_{2i}^{'}(\theta)\right\}$ with $S_{2i}(\theta)=(S_{2i1}(\theta),\ldots,S_{2iR_i}(\theta))$. The matrices $\mathscr{A}_2(\theta)$ and $\mathscr{B}_2(\theta)$ can be estimated empirically by

$$\hat{A}_2=-I\,|^{-1}\sum_{i=1}^{I}\sum_{r=1}^{R_i}\frac{\partial S_{2ir}(\theta)}{\partial\theta'}|_{\theta=\tilde{\theta}}$$

and

$$\hat{B}_2=-I^{-1}\sum_{i=1}^{J}S_{2i}(\theta)S_{2i}^{'}(\theta)|_{\theta=\tilde{\theta}}.$$

Study Design

In this section we discuss the cost-effect design of a prospective study in which a Markov model can characterize dynamic features of the process with some clusters providing repeated aggregate data, and others providing longitudinal responses at the individual level. Note that the expected information for both panel and aggregate data will be computed in a robust sandwich form due to the working independence assumption from the composite likelihood (see above sections). We let I_1 denote the number of tanks for assigned to the panel observation scheme and I_2 denote the number of tanks providing only repeated aggregate data. Without loss of generality we suppose tanks 1,…, I_1 are under the panel and tanks I_1+1,…, I_1+ I_2 are aggregate observation schemes. The composite likelihood resulting from pooling the data from the panel and aggregate data observation schemes is $L(\theta)=L_1(\theta)L_2(\theta)$ where $L_1(\theta)=\prod_{i=1}^{I_1}\prod_{r=1}^{R_i}L_{1ir}(\theta)$ and $\mathscr{G}(\theta)=\mathscr{A}(\theta)^{-1}\mathscr{B}(\theta)\mathscr{A}(\theta)^{-1}$. We let $f=I_1/I$ denote the proportion of tanks that are under panel observation scheme. We let n denote the number of individuals per tank which is fixed and common across all tanks. The cost of observation per individual is C_1 and C_2 for panel and aggregate data observation schemes respectively. The asymptotic robust variance of the maximum composite likelihood estimator is then

$$\mathscr{G}(\theta)=\mathscr{A}(\theta)^{-1}\mathscr{B}(\theta)\mathscr{A}(\theta)^{-1}$$

where

$$\mathscr{A}(\theta)=f\mathscr{A}_1(\theta)+(1-f)\mathscr{A}_2(\theta),$$

$$\mathscr{B}(\theta)=f\mathscr{B}_1(\theta)+(1-f)\mathscr{B}_2(\theta)$$

with the component matrices from eqns. (6) and (8) and

$$\sqrt{I}(\bar{\theta}-\theta)\sim N(0,\mathscr{G}(\theta))$$

with $\bar{\theta}$ being the estimate of θ. Given a target parameter of interest represented by the qth element of θ, the optimal cost-effective design involves allocation of tanks subject to the cost constraint B satisfying

$$\min[\mathscr{G}(\theta)]_{qq}+\rho\left[nI(fC_1+(1-f)C_2)-B\right] \quad (9)$$

where is a Lagrange multiplier. If the interest lies in more than one parameter, we can adopt other optimal allocation methods such as the D-optimality which is widely used in experimental design studies [15].

Here we give an example of the cost-effective design for clustered data under a specified setting. Let n=10 for each tank i and let $X_i\sim$ Bern (0.5) be a tank level covariate, i=1,…, I. We use a 5-state progressive process as in the case of the maturation stages of Northern rock sole. We assume 4 follow-up assessment times (not including a_0) and the assessment times are evenly spaced between 0 and 1. We set λ_{12} such that $P(Z_{ij}(1)=1|Z_{ij}(0)=1)=0.135$. We then set $\lambda_{23}=\lambda_{12}\omega$, $\lambda_{34}=\lambda_{12}\omega^2$, and $\lambda_{45}=\lambda_{12}\omega^3$ with ω=1.1 indicating an increasingly rapid progression through the more advanced states, and set β=log1.2. The data is generated such that the entry times to state 2 within each tank are correlated under a copula model (see above section); the subsequent sojourn times are generated from an exponential distribution. In this example we adopt the Clayton copula with Kendall's τ set to 0 (for independence) or 0.2.

Under the above setting, we now consider the case where the interest lies in study design with the goal is to achieve a pre-specified precision set to 0.01 for the estimator of the regression coefficient. Given the pre-specified variance, Figure 1 shows the percentage of aggregate tanks needed to achieve that variance as a function of the cost when Kendall's τ is 0 (left column) and Kendall's τ is 0.2 (right column). Note that when the cost ratio is 1, λ_{12} increases then decreases again. This is due to the fact that our model is strictly progressive and all units start in state 1. Another thing to notice is that it's the number of aggregate clusters that give us more information rather than the number of individuals per cluster. Hence we see the concave shape for λ_{12} for example. Under this particular situation, aggregate data gives similar amount of information as panel data. Moreover, we see that λ_{34} has a strictly increasing curve for cost ratio=1 which corresponds to the fact that aggregate data is losing information comparing to panel data. Note that increasing Kendall's τ increases the cost to achieve the pre-specified variance. Figure 2 displays the trade-off between the optimal allocation of tanks and the associated asymptotic variance asymptotic variance when we decrease the budget but keep the constraint that the total number of tanks is the same. The number of tanks is fixed at a number such that we can achieve the pre-specified variance under panel observation scheme. Here we used cost ratio C_2/C_1=0.5 for illustration purposes. Again, we plot the results in Figure 2 for Kendall's τ 0 (left column) and Kendall's τ 0.2 (right column) [16].

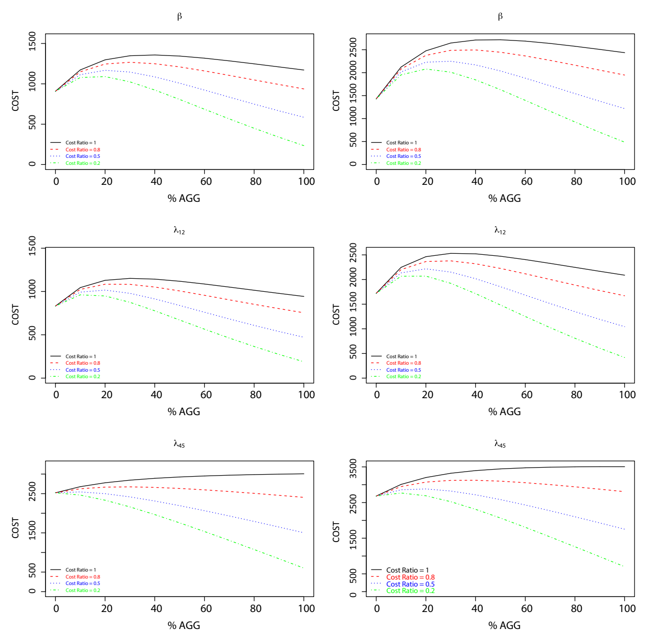

Figure 1: Plot of the trade-off between cost and % of aggregated tanks need to achieve a pre-specified variance with Kendall's τ=0 on the left column and Kendall's τ=0.2 on the right column.

Note that we have also superimposed a blue line mimicking the asymptotic variance from a simulation study under the same setting as outputted from Figure 2. Moreover, we have done 100 simulation to assess the empirical biases (Ebias), empirical standard errors (ESE) and the robust standard errors (ASE) for β. We see a good agreement between the blue lines (simulated ASE) broadly match of the red (expected asymptotic variance) from Figure 3.

Discussion

We have described a cost-effective optimal design method based on clustered panel and aggregate data. Aggregate data may be subjected to a lower cost and effort when monitoring organisms. Having aggregate data may also prevent possible misclassification or measurement error when the organisms are hard to identify. The method proposed here gives insight on the trade-off between number of aggregate tanks and

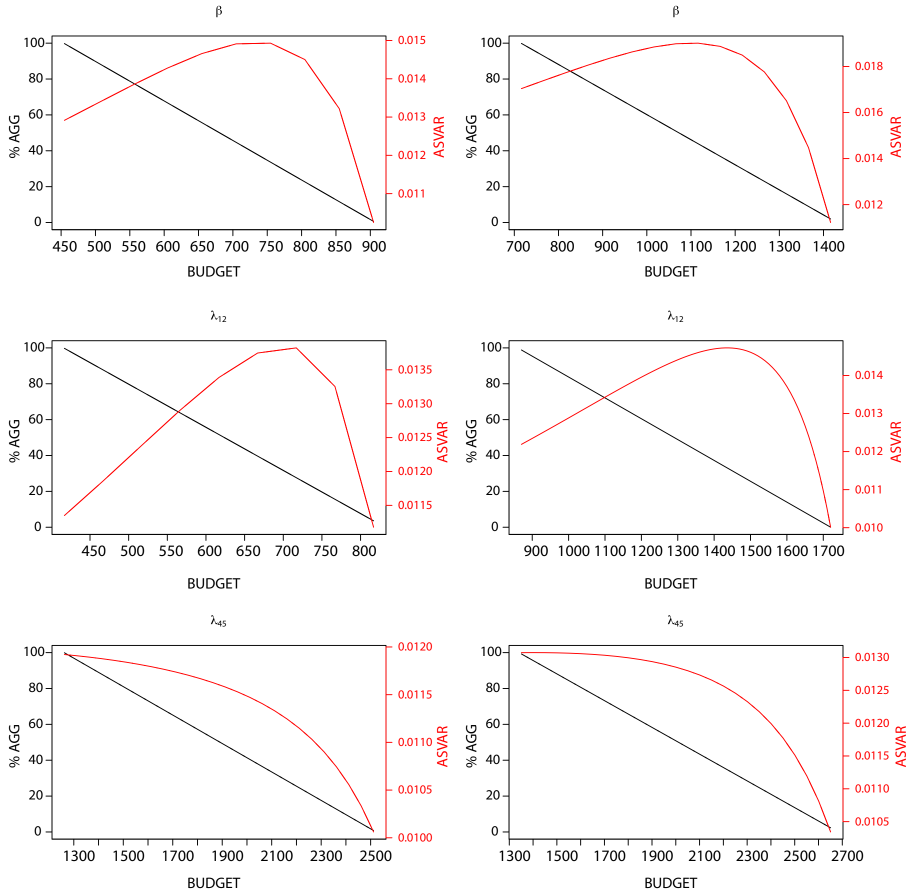

Figure 2: Plot of the trade-off between the optimal allocation of % aggregated tanks and their associated asymptotic variance subject to a fixed budget and number of tanks with a cost ratio of 0.5 with Kendall's τ=0 on the left column and Kendall's τ=0.2 on the right column.

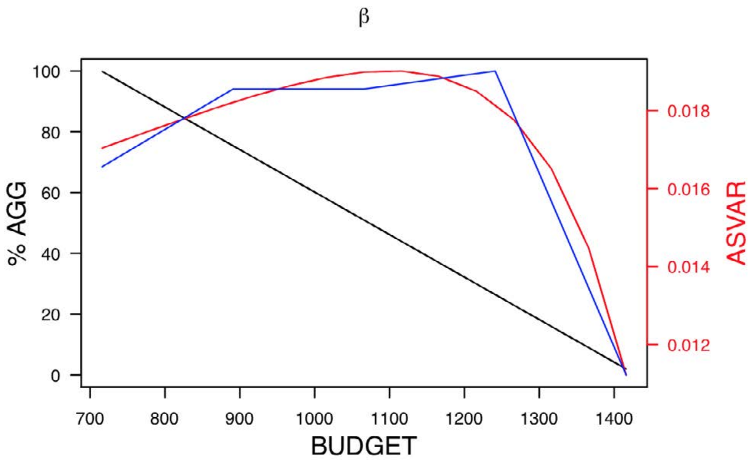

Figure 3: Empirical performance of estimators for 100 simulations under a mixture of panel and composite likelihood via marginal model according to a proportion vs. the expected asymptotic variance.

panel tanks needed in order to achieve a user-desired variance tolerance. Design can also be considered in terms of power of tests of the cluster level covariate effects, or other features of the multistate process such as mean sojourn times or median time to maturation. Depending on the cost ratio and the user desired variance tolerance, one can gain insights on such prospective study with the optimal cost-effective design.

The framework we have described can be generalized in a number of ways. In some settings it may only be possible to record aggregate data at certain phases of the development process (i.e., at the larval stage) but it may be possible to tag or otherwise identify organisms when they are more developed. In this case aggregate data may be available at early stages but tracking of individuals may yield panel observations once a certain stage of the life cycle has been reached. Another interesting variation of this design is to allow timing of assessments to differ between tanks. Some tanks, for example, may be examined more frequently at the early stages of the life cycle and others may be examined more frequently at later stages. Optimal allocation of the tanks to these observation schedules can also be considered.

We restrict attention here to progressive multistate processes with time homogeneous transition intensities. Calculations are easily adapted to deal with piecewise constant transition intensities as done in Jiang and Cook [8]. Extensions may be developed for recurrent processes or processes involving a terminal (e.g., death) state which can be entered at any time during the maturation process, but settings involving multistate models with reversible transitions are much more difficult to handle even under the panel observation scheme.

Acknowledgements

This research was supported by grants from the Natural Sciences and Engineering Research Council of Canada (RGPIN 155849) and the Canadian Institutes for Health Research (FRN 13887). Richard Cook is a Tier I Canada Research Chair in Statistical Methods for Health Research.

References

1. Gouno E, Courtrai L, Fredette M (2011) Estimation from aggregate data. Computational Statistics and Data Analysis 55: 615-626.
2. Laurel B, Danley C, Haines S (2014) The effects of temperature on growth, development and settlement of northern sole larvae (Lepidopsetta polyxystra). Fisheries oceanography 23: 495-505.
3. Munholland PL, Kalbfleisch JD (1991) A semi-Markov model for insect life history data. Biometrics 47: 1117-1126.
4. Borror DL, White RE (1970) A Field Guide to the Insects of America North of Mexico. Boston: Houghton Mifflin.
5. MacRae EC (1977) Estimation of time-varying Markov processes with aggregate data. Econometrica 45, 183-198.
6. Kalbfleisch JD, Lawless JF (1983) Estimation in Markov models from aggregate data. Biometrics 39: 907-919.
7. Varin C, Reid N, Firth D (2011) An overview of composite likelihood methods. Statistica Sinica.
8. Cook RJ, Lawless JF (2018) Multistate Models for the Analysis of Life History Data.
9. Jiang S, Cook RJ (2018) Analysis of aggregate data from clustered multistate processes via composite likelihood.
10. Kalbfleisch JD, Lawless JF (1985) The analysis of panel data under a Markov assumption. Journal of the American Statistical Association 80: 863-871.
11. Jackson CH (2011) Multi-state models for panel data: The msm package for R. Journal of Statistical Software 38: 1-28.
12. Diao L, Cook RJ (2014) Composite likelihood for joint analysis of multiple multistate processes via copulas. Biostatistics 15: 690-705.
13. Nelsen RB (2006) An Introduction to Copulas. Springer, New York.
14. White H (1982) Maximum likelihood estimation of misspecified models. Econometrica 50: 1-25.
15. John RCS, Draper NR (1975) D-Optimality for regression designs: A review. Technometrics 17: 15-23.
16. Hawkyard M, Laurel B, Landon C (2014) Rotifers enriched with taurine by microparticulate and direct enrichment methods influence the growth and metamorphic stage of northern rock sole (Lepidopsetta polyxystra) larvae. Aquaculture 425: 157-151.