# SORT

Statistics and Operations Research Transactions

Volume 27
Number 1, January - June 2003

ISSN: 1696-2281

Generalitat de Catalunya
Institut d'Estadística
de Catalunya

## Aims

SORT (*Statistics and Operations Research Transactions*) – formerly *Qüestiió* - is an international journal launched in 2003, published twice-yearly by the Institut d'Estadística de Catalunya (Idescat), co-sponsored by the Universitat Politècnica de Catalunya (UPC), Universitat de Barcelona, Universitat Autònoma de Barcelona and Universitat de Girona and with the co-operation of the Spanish Region of the International Biometric Society. SORT promotes the publication of original articles of a methodological or applied nature on statistics, operations research, official statistics and biometrics.

The journal is described in the *Encyclopedia of Statistical Sciences*, and referenced in the *Current Index to Statistics*, the *Índice Español de Ciencia y Tecnología*, *Statistical Methods and Abstracts*, as well as MathSci of the American Mathematical Society (*Current Mathematical Publications and Mathematical Reviews*).

SORT represents the third series of the *Quaderns d'Estadística i Investigació Operativa* (*Qüestiió*), published by Idescat since 1992 until 2002, which in turn followed from the first series *Quaderns d'Estadística, Sistemes, Informàtica i Investigació Operativa* (1977-1992). The three series of *Qüestiió* have their origin in the *Cuadernos de Estadística Aplicada e Investigación Operativa*, published by the UPC till 1977.

ISSN: 1696-2281
SORT 27 (1) January - June 1-137 (2003)

# SORT

## Statistics and Operations Research Transactions

Sponsoring institutions

*Universitat Politècnica de Catalunya*
*Universitat de Barcelona*
*Universitat de Girona*
*Universitat Autònoma de Barcelona*
*Institut d'Estadística de Catalunya*

Supporting institution

*Spanish Region of the International Biometric Society*

Generalitat de Catalunya
**Institut d'Estadística de Catalunya**

# SORT 27 (1) January-June 2003

## Contents

*Foreword*

*Articles*

SORT 27 (1) January-June 2003, 1-12

# Likelihood for interval-censored observations from multi-state models

Daniel Commenges*

*INSERM Team of Biostatistics*

## Abstract

We consider the mixed dicrete-continuous pattern of observation in a multi-state model; this is a classical pattern because very often clinical status is assessed at discrete visit times while time of death is observed exactly. The likelihood can easily be written heuristically for such models. However a formal proof is not easy in such observational patterns. We give a rigorous derivation of the likelihood for the illness-death model based on applying Jacod's formula to an observed bivariate counting process.

*MSC:* 62N01, 62N02, 62P10, 92B15, 62M05

*Keywords:* Multi-state models; illness-death; counting processes; ignorability; interval-censoring; Markov models

## 1 Introduction

Multi-state models are a generalisation of survival and competing risks models. In epidemiology, multi-state models are used to represent the evolution of subjects through different statuses, generally including clinical statuses and death. Clinical statuses of subjects are often observed at a finite number of visits. This leads to interval-censored observations of times of transition from one state to another. A classical reference for multi-state models is Andersen *et al.* (1993). This book however essentially treats right-censored observations: building estimators by decomposing the observed processes and equating to zero the martingale term is very elegant in that case but this does not work for interval-censored observations.

**Address for correspondence:* D. Commenges, ISPED, 146 rue Léo. Saignat, Bordeaux, 33076, France. daniel.commenges@isped.u-bordeaux2.fr

Received: June 2002
Accepted: January 2003

One first issue is whether the mechanism leading to these incomplete observations is ignorable. If this is the case, the likelihood can be written heuristically in terms of both transition probabilities and transition intensities. In homogeneous Markov models, transition probabilities can be expressed simply in terms of transition intensities but this is not the case in more general multi-state models. In addition, inference in homogeneous Markov models is easy because these are parametric models. Non-parametric approaches to non-homogeneous Markov models may follow two paths: one is the completely non-parametric approach and can be seen as a generalisation of the Peto-Turnbull approach (Turnbull, 1976); the other implies a restriction to smooth intensities models. In particular, the penalized likelihood method has been applied to this problem. A review of this topic can be found in Commenges (2002). However all these approaches are based on likelihoods which have been given only heuristically. In the complex setting of observations from multi-state models involving a mixed pattern of continuous and dicrete time observations it is important to have a rigorous derivation of the likelihood.

In Section 2 we describe the possible patterns of observation from multi-state models, especially those which are relevant in epidemiology, and then we give the heuristic formulas for the likelihood. We begin Section 3 by describing the theoretical basis of likelihood, Jacod's formula for the likelihood ratio for a counting process and a way to apply it to incomplete observations; we give a rigorous derivation of the likelihood for the illness-death model, based on a representation of this model by a bivariate counting process and applying Jacod's formula to an observed bivariate counting process.

## 2 Generalities on inference

### 2.1 Patterns of observation

Generally we will represent the status of a subject $i$ by a stochastic process $X_i$; $X_i(t)$ can take a finite number of values $\{0,1,\dots,K\}$ and we can make more or less stringent assumptions on the process, for instance, time homogeneity, Markov or semi-Markov properties. Multi-state processes are characterized by transition intensities or transition probabilities between states $h$ and $j$ that we will denote respectively by $\alpha_{hj}(t;\mathcal{F}_{t-})$ and $p_{hj}(s,t) = P(X(t) = j|X(s) = h,\mathcal{F}_{s-})$, where $\mathcal{F}_{s-}$ is the history before $s$; for Markov processes the history can be ignored.

We may consider that the state of the process $i$ is observed at only a finite number of times $V_0^i, V_1^i, \dots, V_m^i$. This typically happens in cohort studies where fixed visit times have been planned. In such cases the exact times of transitions are not known; it is only known that they occurred during a particular interval; these observations are said to be interval-censored. It is also possible that the state of the process is not exactly observed but it is known that it belongs to a subset of $\{0,1,\dots,K\}$.

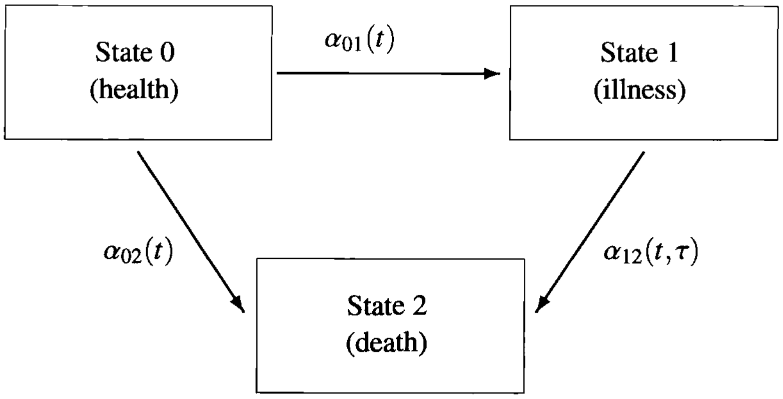

***Figure 1**: Illness-death model.*

The most common pattern of observation is in fact a mixing of discrete and continuous time observations. This is because most multi-state models include states which represent clinical status and one state which represents death: most often clinical status is observed at discrete times (visits) while the (nearly) exact time of death can be retrieved. This is the case in the study of dementia by Joly *et al.* (2002) where an irreversible illness-death model (see Figure 1) was used and dementia was assessed only at planned visits. Note that in the irreversible model no transition from state 1 to state 0 is possible, which is well adapted to modelling dementia, considered as an irreversible clinical condition.

In all cases we should have a model describing the way the data have been observed. For writing reasonably simple likelihoods, there must be some kind of independence of the mechanisms leading to incomplete observations relative to the process itself. A simple likelihood can be written if the observation times are fixed. More realistically, the observation process should be considered as random and intervene in the likelihood. The mechanism leading to incomplete data will be said to be ignorable if the likelihood treating the observation process as non-random leads to the same inference as the full likelihood. An instance where this works is the case of observation processes completely independent of the processes of interest $X_i$. A general approach for representing the observation of a process $X_i$ is to consider a process $R_i$ which takes value 1 at $t$ if $X_i(t)$ is observed, 0 otherwise. $R_i$ must satisfy certain independence properties relatively to $X_i$ in order to be ignorable; in that case one can write the likelihood as if $R_i$ was fixed. In the remaining of this paper we will assume that this is the case that the mechanism leading to incomplete observation is ignorable: we shall write the likelihood as if the discrete observation times and the right censoring variable were fixed.

### 2.2 Inference

The first interesting fact to be noted is that with continuous observation times, the inference problem in a multi-state model can be decoupled into several survival problems; with discrete-time observation (leading to interval-censoring), this is no longer possible. The likelihood for the whole observation of the trajectory must be written as in Joly and Commenges (1999); Joly *et al.* (2002) gave an example of the bias that occurs when one tries to treat interval-censored observation from an illness-death model as a survival problem.

We shall give the likelihood for interval-censored observations of a single process $X$ taken at $V_0, V_1, \ldots, V_m$, (treating the $V_j$ as fixed); for sake of simplicity we drop the index $i$. If we have a sample of size $n$ the processes $X$ and the observation times should be indexed by $i$; assuming the independence of the processes (the histories of the "subjects") the likelihood is the product of the individual likelihoods. For sake of simplicity we will also restrict to Markov models. So, for purely discrete-time observations this individual likelihood is as follows:

$$\mathcal{L} = \prod_{r=0}^{m-1} p_{X(V_r),X(V_{r+1})}(V_r, V_{r+1}),$$

where $p_{hj}(s,t) = P(X(t) = j | X(s) = h)$.

Variants of this likelihood can be written in cases of mixing of continuous and discrete-time observations. We give the likelihood when the process is observed at discrete times but time of transition towards one absorbing state, representing generally death, is exactly observed or right-censored, a common model and observational pattern in epidemiology. Denote by $K$ this absorbing state. Observations of $X$ are taken at $V_0, V_1, \ldots, V_L$ and the vital status is observed until $C$ ($C \geq V_L$); here $V_L$ is the last visit time of an alive subject. Let us call $\tilde{T}$ the follow-up time that is $\tilde{T} = \min(T, C)$, where $T$ is the time of death; we observe $\tilde{T}$ and $\delta = I\{T \leq C\}$. For continuous intensities model the likelihood can be written:

$$\mathcal{L} = \left[\prod_{r=0}^{L-1} p_{X(V_r),X(V_{r+1})}(V_r, V_{r+1})\right] \sum_{j \neq K} p_{X(V_L),j}(V_L, \tilde{T}) \alpha_{j,K}(\tilde{T})^{\delta}.$$

This likelihood can be understood intuitively as the "probability" of the observed trajectory but it is not so easy to prove that this is really the likelihood, as we shall see in the next section. For this likelihood to be useful, it must be expressed in term of the transition intensities which are the basic parameters of the model; so we must be able to express the transition probabilities in term of the transition intensities. This is particularly easy in the homogeneous Markov model. In other models it generally requires the computation of integrals.

Let us now specialize these formulas to the illness-death model, a model with the three states "health", "illness", "death" respectively labelled $0, 1, 2$. If the subject starts in state "health", has never been observed in the "illness" state and was last seen at visit $L$ (at time $V_L$) the likelihood is:

$$\mathcal{L} = p_{00}(V_0, V_L)[p_{00}(V_L, \tilde{T})\alpha_{02}(\tilde{T})^{\delta} + p_{01}(V_L, \tilde{T})\alpha_{12}(\tilde{T})^{\delta}]; \tag{1}$$

if the subject has been observed in the illness state for the first time at $V_J$ then the likelihood is:

$$\mathcal{L} = p_{00}(V_0, V_{J-1})p_{01}(V_{J-1}, V_J)p_{11}(V_J, \tilde{T})\alpha_{12}(\tilde{T})^{\delta}. \tag{2}$$

This equations are valid for the reversible as well as for the irreversible illness-death model. In Markov models, the transition probabilities are linked to the transition intensities by the Kolmogorov differential equations. For the irreversible illness-death model, to which we shall specialize from now on, the forward Kolmogorov equation gives:

$$\begin{aligned}
\frac{dp_{00}}{dt}(s,t) &= -p_{00}(s,t)[\alpha_{01}(t) + \alpha_{02}(t)] \\
\frac{dp_{11}}{dt}(s,t) &= -p_{11}(s,t)\alpha_{12}(t) \\
\frac{dp_{01}}{dt}(s,t) &= p_{00}(s,t)\alpha_{01}(t) - p_{01}(s,t)\alpha_{12}(t).
\end{aligned} \tag{3}$$

The solution of these equations are:

$$\begin{aligned}
p_{00}(s,t) &= e^{-A_{01}(s,t) - A_{02}(s,t)} \\
p_{11}(s,t) &= e^{-A_{12}(s,t)} \\
p_{01}(s,t) &= \int_s^t p_{00}(s,u)\alpha_{01}(u)p_{11}(u,t)du,
\end{aligned}$$

where $A_{hj}(s,t) = \int_s^t \alpha_{hj}(u)du$. These equations have been given for general compensators in Andersen *et al.* (1993).

Inference can be based on maximising the likelihood. If a parametric model is chosen, modified Newton-Raphson algorithms (such as the Marquardt algorithm) can be used for the maximisation (the simplest parametric model is the homogeneous Markov model, followed by the piece-wise homogeneous Markov model). Non-parametric approaches can take two paths: one is the unconstrained non-parametric approach in the spirit of Turnbull (1976) and this was developed by Frydman (1995), another one uses smoothing, for instance through penalized likelihood such as in Joly and Commenges (1999). In the former path the EM algorithm is attractive, in the latter the Marquard algorithm achieves a good speed of convergence. All the above approaches are based on the likelihood which has been derived heuristically. In complex problems such as the one at hand, it is important to have a rigourous derivation of the likelihood; this is the purpose of the next section.

## 3 Rigorous derivation of likelihood for illness-death

### *3.1 Generality on likelihood*

Consider a measurable space $(\Omega, \mathcal{F})$ and a family of measures $P^\theta$ absolutely continuous relatively to a dominant measure $P^0$. The likelihood ratio is defined by:

$$\mathcal{L}_{\mathcal{F}}(\theta) = \frac{dP^\theta}{dP^0}_{|\mathcal{F}}$$

where $\frac{dP^\theta}{dP^0}_{|\mathcal{F}}$ is the Radon-Nikodym derivative of $P^\theta$ relatively to $P^0$. Recall that $\frac{dP^\theta}{dP^0}_{|\mathcal{F}}$ is the $\mathcal{F}$-measurable random variable such that

$$P^\theta(F) = \int_F \frac{dP^\theta}{dP^0} dP^0, F \in \mathcal{F}$$

For instance, the likelihood ratio corresponding to the observation of a random variable $X$ (that is to the $\sigma$-algebra $\mathcal{X} = \sigma(X)$) can be written

$$\mathcal{L}_{\mathcal{X}}(\theta) = \frac{f_X^\theta(X)}{f_X^0(X)},$$

where $f_X^\theta(.)$ is the density of the law of $X$ relatively to a given measure: for instance, for a continuous variable, $f_X^\theta(.)$ is the probability density function. Since the denominator does not depend on $\theta$, inference can be based only on $f_X^\theta(X)$, which is the form of the likelihood which appears in statistical papers. It is sometimes overlooked that the likelihood is a random variable, being a composition of the probability density function and the random variable $X$ itself.

When dealing with complex problems such as inference based on incomplete observations of processes, such a simplification is not available and it is necessary to return to more fundamental theory. We are especially interested here in writing the likelihood for interval-censored observations from an illness-death model. We shall see that an illness-death model can be described as a bivariate counting process. We could find the likelihood for interval-censored observation of a unidimensional counting process relatively easily, for instance by considering that we have interval-censored observation of a random variable which represents the time of jump. However for a multivariate process this becomes much more difficult.

Consider the case of multivariate (or marked) point processes: $N = (N_h, h = 1, 2, \ldots)$. Denote $N_. = \sum N_h$ and $\Lambda_. = \sum \Lambda_h$, where $\Lambda_h$ are the compensators of $N_h$ (that is $N_h - \Lambda_h$ are martingales and $\Lambda_h$ are increasing predictable processes); when the compensators are continuous we define intensities $\lambda_h$ by $\Lambda_h = \int \lambda_h$. Consider also two probability measures $\tilde{P}$ and $P$ with $\tilde{P} \ll P$. Jacod (1975) has given the formula for the likelihood ratio of the process $N$; this formula is presented in Andersen *et al.* (1993) in term of product-integral, and supposing there is no information at time 0 it takes the form:

$$\frac{d\tilde{P}}{dP} = \prod_{t \le C} \prod_h \left( \frac{d\tilde{\Lambda}_h}{d\Lambda_h}(t) \right)^{\Delta N_h(t)} \frac{\prod_{t \le C : \Delta N_.(t) \neq 1} (1 - d\tilde{\Lambda}_.(t))}{\prod_{t \le C : \Delta N_.(t) \neq 1} (1 - d\Lambda_.(t))}$$

This is the likelihood ratio for the sigma-algebra $\mathcal{N} = \sigma(N(t), t \ge 0)$ with compensators relative to the filtration $\mathcal{N}_t = \sigma(N(u), u \ge 0, u \le t)$; thus we cannot directly apply the formula because we do not observe $\mathcal{N}$ but $\mathcal{O} \subset \mathcal{N}$.

There are two strategies for applying this formula to our incomplete observation problem:

- Take the conditional expectation: $E[\frac{d\tilde{P}}{dP}|\mathcal{O}]$
- Apply the formula not on $N$ but on an observed process

As an example of the latter consider the one-dimensional (so $h = 1$) process $N^{\mathcal{O}}(t) = N(l(t))$, where $l(t) = \sup(u \le t : R(u) = 1)$. By definition this process is observed: $\mathcal{O} = \sigma(N^{\mathcal{O}}(t), t > 0)$, so that we can apply Jacod's formula. Consider the case of purely interval-censored data: $R(t) = 1$ for $t = V_0, V_1, \ldots, V_m$, $R(t) = 0$ otherwise. Then $N^{\mathcal{O}}$ has a discrete compensator with jumps at $V_0, V_1, \ldots, V_m$

$$\Delta\Lambda^{\mathcal{O}}(V_j) = P[N^{\mathcal{O}}(V_j) = 1 | N^{\mathcal{O}}(V_{j-1}) = 0] I_{\{N^{\mathcal{O}}(V_{j-1}) = 0\}}$$

It is easy to see that by applying Jacod's formula we get the expected result for the likelihood (expressed in term of the survival function $S$ of the jump time):

$$\mathcal{L} = d\tilde{P} = \tilde{S}(V_{J-1}) - \tilde{S}(V_J),$$

where the random variable $J$ is defined as $N^{\mathcal{O}}(V_J) - N^{\mathcal{O}}(V_{J-1}) = 1$; in this formula we have dropped the denominator which does not depend on the parameters.

### 3.2 Counting process model for illness-death

Consider one counting process $N_I$ for illness ($N_I(t) = 0$ if healthy at $t$, $N_I(t) = 1$ if subject became ill before $t$) with intensity $\lambda_I$ and one for death $N_D$ ($N_D(t) = 0$ if alive at $t$, $N_D(t) = 1$ if subject died before $t$) with intensity $\lambda_D$. Let us model the intensities (in the $\mathcal{N}_t$-filtration) as:

$$\lambda_I(t) = I_{\{N_I(t-)=0\}} I_{\{N_D(t-)=0\}} \alpha_{01}(t)$$

$$\lambda_D(t) = I_{\{N_D(t-)=0\}} [I_{\{N_I(t-)=0\}} \alpha_{02}(t) + I_{\{N_I(t-)=1\}} \alpha_{12}(t)] \tag{4}$$

If we define $X = N_I + N_D + N_D(1 - N_I)$, this defines a multi-state process taking values on $\{0, 1, 2\}$ and with transition intensities $\alpha_{01}(.)$, $\alpha_{02}(.)$ and $\alpha_{12}(.)$ between $(0,1)$, $(0,2)$ and $(1,2)$ respectively; there is identity between this multi-state (illness-death ) process and the bivariate counting process.

To $N_D$ we associate a response process $R_D(t) = 1$, for all $t \leq C$; to $N_I$, we associate a response process $R_I(t) = 1$ for $t = V_0, \dots, V_m$, $R_I(t) = 0$ otherwise. The observed process is $N^{\mathcal{O}} = (N_I^{\mathcal{O}}, N_D^{\mathcal{O}})$, with

$$N_I^{\mathcal{O}}(t) = N_I(l(t))$$

where $l(t) = \sup\{u \leq t : R_I(u) = 1\}$, and

$$N_D^{\mathcal{O}}(t) = N_D(t), \text{ for } t \leq C.$$

Jacod's formula can be applied if we know the compensator of $N^{\mathcal{O}}$ in the $\mathcal{O}_t$ filtration: although we observe $N_D$ its compensator is not the same on $\mathcal{N}_t$ and on $\mathcal{O}_t$. Thus, we need compute the compensators of $N_I^{\mathcal{O}}$ and $N_D^{\mathcal{O}}$ in the $\mathcal{O}_t$-filtration. It is easy to see that $N_I^{\mathcal{O}}$ has a discrete compensator which is null everywhere except possibly at observation times $V_j, j = 0, \dots, m$ where it is equal to :

$$\Delta\Lambda_I^{\mathcal{O}}(V_j) \quad = \quad P[N_I^{\mathcal{O}}(V_j) = 1 | N_I^{\mathcal{O}}(V_{j-1}) = 0, N_D^{\mathcal{O}}(V_j-) = 0] I_{\{N_I^{\mathcal{O}}(V_{j-1})=0\}} I_{\{N_D^{\mathcal{O}}(V_j-)=0\}}$$

It can be seen that $N_I^{\mathcal{O}}$ and $N_D^{\mathcal{O}}$ can be replaced by $N_I$ and $N_D$ and, reminding that $N_I$ and $N_D$ are not independent, we can write:

$$P[N_I(V_j) = 1 | N_I(V_{j-1}) = 0, N_D(V_j-) = 0] = \frac{p_{01}(V_{j-1}, V_j)}{p_{0.}(V_{j-1}, V_j)},$$

where $p_{0.}(.,.) = p_{00}(.,.) + p_{01}(.,.)$ (the probability of being still alive); of course the transition probabilities $p_{hj}(s,t)$ still have a meaning in terms of the bivariate counting process, for instance $p_{00}(s,t) = P[N_I(t = 0, N_D(t) = 0 | N_I(s) = 0, N_D(s) = 0]$.

As for $N_D$, it is observed in continuous time so we have $N_D^{\mathcal{O}}(t) = N_D(t)$, for $t \leq C$. However its compensator is not the same in the $\mathcal{N}_t$-filtration and in the $\mathcal{O}_t$-filtration: it is clear that the intensity given in formula (4) is not $\mathcal{O}_{t-}$-measurable. We may use the innovation theorem and compute the $\mathcal{O}_t$-intensity as:

$$\lambda_D^{\mathcal{O}}(t) = \mathrm{E}[\lambda_D(t) | \mathcal{O}_{t-}] = \mathrm{E}[I_{\{N_D(t-)=0\}}[I_{\{N_I(t-)=0\}}\alpha_{02}(t) + I_{\{N_I(t-)=1\}}\alpha_{12}(t)] | \mathcal{O}_{t-}].$$

In this formula, only $I_{\{N_I(t-)=0\}}$ is not $\mathcal{O}_{t-}$-measurable so the only problem is to compute

$$\mathrm{E}[I_{\{N_I(t-)=0\}} | \mathcal{O}_{t-}] = P[N_I(t-) = 0 | \mathcal{O}_{t-}].$$

If $N_D(t-) = 1$ we can take any arbitrary value for this probability; if $N_I(l(t-)) = 1$, this probability is null. The only non-trivial quantity is

$$P[N_I(t-) = 0 | N_D(t-) = 0, N_I(l(t-)) = 0] = \frac{p_{00}(l(t-), t-)}{p_{0.}(l(t-), t-)}.$$

Finally, the $\mathcal{O}_t$-intensity of $N_D$ is

$$\lambda_D^{\mathcal{O}}(t) = I_{\{N_D(t-)=0\}}[I_{\{N_I(l(t-))=0\}}\tilde{\alpha}_D(t) + I_{\{N_I(l(t-))=1\}}\alpha_{12}(t)],$$

where $\bar{\alpha}_D(t) = \frac{p_{00}(l(t-),t-)\alpha_{02}(t)+p_{01}(l(t-),t-)\alpha_{12}(t)}{p_{0.}(l(t-),t-)}$. This formula has a natural interpretation, the intensity being a weighting of the transition intensities from health and illness with the required probabilities conditional on what has been observed just before $t$; if the subject has been observed in the illness state, then the intensity is $\alpha_{12}$ (for an alive subject).

The likelihood ratio in Jacod's formula can be written as the product of three terms $\mathcal{L} = \mathcal{L}_I \mathcal{L}_D \mathcal{L}_{.}$. The first term is the contribution of observing a jump of $N_I$: it is equal to 1 if no jump has been observed and if a jump has been observed at $V_J$:

$$\mathcal{L}_I = \frac{\Delta\tilde{\Lambda}_I^{\mathcal{O}}(V_J)}{\Delta\Lambda_I^{\mathcal{O}}(V_J)} = \frac{\tilde{p}_{01}(V_{J-1},V_J)p_{0.}(V_{J-1},V_J)}{\tilde{p}_{0.}(V_{J-1},V_J)p_{01}(V_{J-1},V_J)}.$$

From now on we drop the denominator and the tilde and we will simply write: $\mathcal{L}_I = \frac{p_{01}(V_{J-1},V_J)}{p_{0.}(V_{J-1},V_J)}$

The second term is the contribution of observing a jump of $N_D$: it is equal to 1 if no jump has been observed; if a jump (that is death) has been observed at $T$,it is equal to $\lambda_D^{\mathcal{O}}(T)$. If the subject has been seen ill at $V_J$ the contribution is $\mathcal{L}_D = \alpha_{12}(T)$; if not it is

$$\mathcal{L}_D = \bar{\alpha}_D(T) = \frac{p_{00}(l(T-),T-)\alpha_{02}(T) + p_{01}(l(T-),T-)\alpha_{12}(T)}{p_{0.}(l(T-),T-)}.$$

The last term of the formula, the product integral over times where no jump happened, is the product of a dicrete and a continuous part: $\mathcal{L}_{.}\mathcal{L}_{.I}\mathcal{L}_{.D}$. The discrete part $\mathcal{L}_{.I}$ comes from the discrete compensator $\Lambda_I^{\mathcal{O}}$ and if a subject has been seen ill for the first time at $V_J$ is a simple product:

$$\mathcal{L}_{.I} = \prod_{j=1}^{J-1}(1 - \Delta\Lambda_I^{\mathcal{O}}(V_j)) = \frac{p_{00}(V_0,V_{J-1})}{p_{0.}(V_0,V_{J-1})};$$

the product stops at $V_{J-1}$ because there is a jump at $V_J$ and the compensator is constant after $V_J$; if the subject is never seen ill, the product goes until the last visit time. Finally the continuous part of the product integral is

$$\mathcal{L}_{.D} = \prod_{t \le \tilde{T}}(1 - d\Lambda_D^{\mathcal{O}}(t)) = e^{-\int_{V_0}^{\tilde{T}} \lambda_D^{\mathcal{O}}(t)dt}.$$

On $V_{j-1} < t < V_j$, where $N_I(V_{j-1}) = 0$ and $N_D(t-) = 0$ we have using the Kolmogorov equations (3)

$$\lambda_D^{\mathcal{O}}(t) = \bar{\alpha}_D(t) = -\frac{d\log p_{0.}(V_{j-1},t)}{dt}.$$

Thus for a subject who has not been seen ill we have:

$$\mathcal{L}_{.D} = e^{-\int_{V_0}^{\tilde{T}} \bar{\alpha}_D(t)dt} = p_{0.}(V_0,\tilde{T}),$$

and for a subject seen ill at $V_J$:

$$\mathcal{L}_{.D} = e^{-\int_{V_0}^{V_J} \bar{\alpha}_D(t)dt - \int_{V_J}^{\tilde{T}} \alpha_{12}(t)dt} = p_{0.}(V_0, V_J)p_{11}(V_J, \tilde{T}).$$

Finally for a subject not seen ill, calling $V_L = l(\tilde{T})$ the last visit time, we have

$$\mathcal{L}_{.I}\mathcal{L}_{.D} = p_{00}(V_0, V_L)p_{0.}(V_L, \tilde{T}).$$

Thus the likelihood is:

$$\mathcal{L} = p_{00}(V_0, V_L)p_{0.}(V_L, \tilde{T})\bar{\alpha}_D(\tilde{T})^{\delta},$$

where $\alpha_D(\tilde{T}) = \frac{p_{00}(V_L,T)\alpha_{02}(\tilde{T})+p_{01}(V_L,\tilde{T})\alpha_{12}}{p_{0.}(V_L,\tilde{T})}$, which is identical to (1).

For a subject seen ill at $V_J$, writing the likelihood as $\mathcal{L} = \mathcal{L}_{.I}\mathcal{L}_{I}\mathcal{L}_{.D}\mathcal{L}_{D}$ we have:

$$\mathcal{L} = \frac{p_{00}(V_0, V_{J-1})}{p_{0.}(V_0, V_{J-1})} \frac{p_{01}(V_{J-1}, V_J)}{p_{0.}(V_{J-1}, V_J)} p_{0.}(V_0, V_J)p_{11}(V_J, \tilde{T})\alpha_{12}(\tilde{T})^{\delta},$$

which is identical to (2).

Thus we have proved that the heuristic way of deriving the likelihood gives the correct result for the illness-death model with the mixed discrete-continuous time observation pattern.

## 4 References

Andersen, P.K., Borgan Ø., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New-York: Springer-Verlag.

Commenges, D. (2002). Inference for multi-state models from interval-censored data. *Statistical Methods in Medical Research*, 11, 167-182.

Frydman, H. (1995). Non-parametric estimation of a Markov "illness-death model" process from interval-censored observations, with application to diabetes survival data. *Biometrika* 82, 773-789.

Jacod, J. (1975). Multivariate point processes: predictable projection; Radon-Nikodym derivative, representation of martingales. *Z. Wahrsheinlichkeitsth*, 31, 235-253.

Joly, P. and Commenges, D. (1999). A penalized likelihood approach for a progressive three-state model with censored and truncated data: Application to AIDS. *Biometrics*, 55, 887-890.

Joly, P., Commenges, D., Helmer, C. and Letenneur, L. (2002). A penalized likelihood appproach for an illness-death model with interval-censored data: application to age-specific incidence of dementia. *Biostatistics*, 3, 433- 443.

Turnbull, B.W. (1976). The empirical distribution function with arbitrarily grouped, censored and truncated data. *Journal of the Royal Statistical Society, Series B*, 38, 290-5.

## Resum

Considerem un patró d'observació mixt: discret i continu en un model multi-estat; aquest patró és clàssic ja que molt sovint l'estatus clínic s'avalua en temps de visita discrets i el temps de la mort s'observa exactament. La versemblança es pot escriure, heurísticament, de forma senzilla per a aquests models. Nogensmenys, les demostracions formals no són senzilles amb aquest patrons observacionals. Donem una derivació rigorosa de la versemblança per al model de malaltia-mort basant-nos en l'aplicació de la fòrmula de Jacod a un procés comptador bivariat.

*MSC:* 62N01, 62N02, 62P10, 92B15, 62M05

*Paraules clau:* Censurament en un interval; ignorabilitat; malatia-mort; models de Markov; models multi-estat; processos comptadors

SORT 27 (1) January-June 2003, 13-30

# Cumulative processes related to event histories

Richard J. Cook*, Jerald F. Lawless and Ker-Ai Lee

*University of Waterloo*

## Abstract

Costs or benefits which accumulate for individuals over time are of interest in many life history processes. Familiar examples include costs of health care for persons with chronic medical conditions, the payments to insured persons during periods of disability, and quality of life which is sometimes used in the evaluation of treatments in terminally ill patients. For convenience, here we use the term costs to refer to cost or other cumulative measures. Two important scenarios are (i) where costs are associated with the occurrence of certain events, so that total cost accumulates as a step function, and (ii) where individuals may move between various states over time, with cost accumulating at a constant rate determined by the state occupied. In both cases, there is frequently a random variable T that represents the duration of the process generating the costs. Here we consider estimation of the mean cumulative cost over a period of interest using methods based upon marginal features of the cost process and intensity based models. Robustness to adaptive censoring is discussed in the context of the multi-state methods. Data from a quality of life study of breast cancer patients are used to illustrate the methods.

*MSC:* 62N01, 62N02, 62N05, 62P20

*Keywords:* Adaptive censoring, cumulative processes, marginal methods, multi-state processes, quality adjusted life-years, prevalence function.

## 1 Introduction

Costs or benefits that accumulate over time for individuals are of interest in many life history processes. Familiar examples include the cost of health care for persons with chronic medical conditions, the payments to insured persons during periods of disability, and cumulative quality of life measures which are sometimes used in the evaluation of treatments for terminally ill patients. Costs or benefits may be

* *Address for correspondence:* R. J. Cook. Department of Statistics and Actuarial Science. Waterloo, Canada, N2L 3G1. Phone: (519) 888-4567 ext. 5549, Fax: (519) 746-1875, rjcook@uwaterloo.ca

Received: October 2002
Accepted: January 2003

multivariate and may accrue for a variety of reasons. For example, in studies of persons with chronic obstructive pulmonary disease (e.g. Torrance *et al.*, 1999) costs were incurred by prescription of prophylactic or therapeutic medications, by hospitalizations, by time off work, and so on.

For convenience we will often use the term costs to refer to cost or other cumulative measures such as utility, profit, or quality of life, and let $C(t)$ denote a cumulative (univariate) cost for an individual over the time period $(0,t)$. There is typically also a random variable $T$ that represents the duration of the cumulative process, so the objects of interest are $T$ and $\{C(t),\ 0 \leq t \leq T\}$. Simple methods for the analysis of cumulative cost (e.g. Lin *et al.*, 1997; Zhao and Tsiatis, 1997) have focussed directly on them, or in some cases, just on the total lifetime cost, $C(T)$. However, a more informative approach is to consider the underlying event processes that generate costs, along with the costs themselves. For example, in a breast cancer trial (Gelber *et al.* 1995) discussed later in the paper, different utilities were assigned for periods in which patients were (i) subject to toxic effects of treatment, (ii) toxicity-free and relapse-free, and (iii) in a state of relapse. Cumulative utility was then used to define a quality of life measure, so that $C(T)$ can be thought of as a "quality-adjusted" lifetime.

Advantages of analyzing and modeling the event processes that generate costs include increased understanding; the ability to deal with observation schemes involving censoring, intermittent observation, or truncation; methods for predicting costs; a convenient separation of the underlying event process from costs which may be subjective, or subject to differing interpretations. The purpose of this paper is to review models on which analysis of cumulative costs can be based, and to discuss efficiency and robustness properties associated with these approaches. An analysis of data on the treatment of breast cancer patients (Gelber *et al.* 1995) will be used for illustration.

We now set some general notation and describe two frameworks that have been used to study cumulative processes.

The first framework assumes that for each individual $i$ in a study there is a cumulative cost (or quality) process $\{C_i(t),\ t \geq 0\}$, and a time $T_i$ at which the process terminates. For example, in a cost of treatment study $T_i$ would represent the duration of the treatment period for the individual. In studies of the utilization of health care resources among patients with terminal medical conditions, $T_i$ would represent the time of death. In many studies the value of $T_i$ may be right-censored at some censoring time $\tau_i$, in which case the cost process is unobserved for $t > \tau_i$. Considerable previous work has focussed on nonparametric estimation of the distribution of "total lifetime cost" $C_i = C(T_i)$, or just on $E(C_i)$; see for example Lin *et al.* (1997), Zhao and Tsiatis (1997), Bang and Tsiatis (2000), Ghosh and Lin (2000), and Strawderman (2000). In most realistic situations $T_i$ is not independent of the cost process; more specifically, if $\overline{C}_i(t) = \{C_i(u),\ 0 \leq u < t\}$ is the cost history to time $t$, then the termination time hazard function,

$$\lim_{\Delta t \longrightarrow 0} \frac{Pr(T_i < t+\Delta t | T_i \geq t, \overline{C}_i(t))}{\Delta t}, \tag{1.1}$$

depends on $\overline{C}_i(t)$. This implies that, even if the censoring time $\tau_i$ and $(T_i,\overline{C}_i(T_i))$ are independent, the censoring value $C_i^* = C(\tau_i)$ and the total lifetime cost $C_i = C(T_i)$ are not in general independent.

The second framework we discuss models the underlying multi-state process driving the costs. Suppose that at time $t$ an individual occupies one of $K$ life states $1,\ldots,K$. It is assumed that all individuals begin in state 1 at $t = 0$, that states $1,\ldots,K-1$ are transient and that state $K$ is an absorbing state. Letting $Y(t)$ represent the state occupied by an individual at time $t$, we assume that there is a cost rate function $V[Y(t),t]$ that determines the incremental cost over the short interval $(t,t+dt)$. The total cumulative cost up to time $t$ is then

$$C(t) = \int_0^t V[Y(u),u]du. \tag{1.2}$$

The process terminates upon entry to state $K$, which occurs at time $T$, so that $V[K,u] = 0$ for all $u > 0$.

Given $(Y(u),u)$, the cost rate function $V[Y(u),u]$ may in general be random, but we restrict consideration to cases where

$$V[Y(u),u] = v_j(u) \text{ if } Y(u) = j, \tag{1.3}$$

where $v_j(u)$ is a known (deterministic) function, $j = 1,2,\ldots,K$. In this case (1.2) gives

$$C(t) = \sum_{j=1}^{K-1} \int_0^t v_j(u)I[Y(u) = j]du \tag{1.4}$$

and

$$E[C(t)] = \sum_{j=1}^{K-1} \int_0^t v_j(u)p_j(u)du, \tag{1.5}$$

where

$$p_j(u) = Pr[Y(u) = j], \qquad j = 1,\ldots,K, \tag{1.6}$$

are prevalence functions. Gelber *et al.* (1995), Glasziou *et al.* (1990) and others have considered the case where $v_j(u) = v_j$ in connection with quality of life.

Note that in this framework $C(T) = C(\infty)$, and process termination is conveniently handled within the multi-state model. However, assumptions about the process $\{Y(t),\ t \geq 0\}$ are needed. In a completely general setting transition intensities might depend on prior cost history, but in the case of deterministic cost rate functions (1.3) we have

$$Pr[Y(t+\Delta t) = j|\overline{Y}(t),\overline{C}(t)] = Pr[Y(t+\Delta t) = j|\overline{Y}(t)] \tag{1.7}$$

so we merely need to model the multi-state process.

The remainder of the paper is as follows. Section 2 reviews strategies for estimation of cost distributions and addresses the multi-state framework in more detail. Methods based on full models are compared with those based on recently proposed robust nonparametric estimates of prevalence functions $p_j(t)$. Section 3 examines the various methods in the context of quality of life assessments in an IBCSG Breast Cancer Trial (e.g. Gelber *et al.* 1995). Section 4 presents conclusions and discusses additional problems.

## 2 Strategies for estimation

### 2.1 Marginal methods

Suppose interest lies in the distribution of $C = C(T)$. Furthermore assume $\tau$ is independent of $(T, \overline{C}(T))$ with corresponding survivor function $K(t) = P(\tau > t)$. where $h(t|\cdot)$ is the hazard function for the In this setting we observe $X = T \wedge \tau$, $\Delta = I(T \leq \tau)$ and $\overline{C}(X) = \{C(u),\ 0 \leq u < X\}$ which we denote as $(X, \Delta, \overline{C}(X))$. If $n$ individuals are under observation and their responses are independently distributed, then we observe $n$ independent replicates $\{(X_i, \Delta_i, \overline{C}_i(X_i)), i = 1, \ldots, n\}$.

Glasziou *et al.* (1990) point out that even when $\tau$ is independent of $(T, C)$, the $C$-censoring value, $C^* = C(\tau)$, and $C$ are correlated. As a result, the assumption of independent censoring for $C$ is violated, and

$$\lim_{\Delta c \downarrow 0} \frac{Pr\{C < c + \Delta c | C \geq c, C^* \geq c\}}{\Delta c} \neq \lim_{\Delta c \downarrow 0} \frac{Pr\{C < c + \Delta c | C \geq c\}}{\Delta c}.$$

Zhao and Tsiatis (1997, 1999), Bang and Tsiatis (2000), and others suggest the use of "inverse probability of censoring-weighted" estimating equations (e.g. Robins and Rotnitzky, 1995) for estimation of the survivor function $Pr(C \geq c)$ to adjust for the dependent censoring induced by $\tau$. Specifically they propose the estimate

$$\widehat{Pr}(C \geq c) = \frac{1}{n} \sum_{i=1}^{n} \frac{\Delta_i}{\widehat{K}(T_i)} I(C_i \geq c) \tag{2.1}$$

where $\widehat{K}(t)$ is a consistent estimate of the censoring time survivor function.

The expected cost up to time $t$,

$$\mu(t) = E\{C(T \wedge t)\} \tag{2.2}$$

is often of interest, as is the expected lifetime cost $E(C) = \mu(\infty)$. These can be estimated using the fact that

$$\mu(t) = \int_0^\infty Pr[C(T \wedge t) > c] dc. \tag{2.3}$$

Cook and Lawless (1997), Lin *et al.* (1997) and Ghosh and Lin (2000) discuss alternative estimators based on the fact that

$$\mu(t) = \int_0^t S(u) dM(u), \tag{2.4}$$

where $S(u) = Pr(T_i \geq u)$ and $dM(u) = E\{dC_i(u)|T_i \geq u\}$. We can estimate $S(u)$ with an ordinary Kaplan-Meier estimate, and $dM(u)$ as

$$d\hat{M}(u) = \sum_{i=1}^{n} I(X_i \geq u)dC_i(u) / \sum_{i=1}^{n} I(X_i \geq u). \tag{2.5}$$

Strawderman (2000) discusses and compares (2.5) and estimators based on (2.1) and (2.3).

The estimators above were developed under the assumption that censoring times $\tau_i$ are independent of $(T_i, \bar{C}(T_i))$, $i = 1,\ldots,n$. This is sometimes violated, as we discuss below. The next two sections deal with methods based on multi-state models and how to deal with non-independent censoring.

### 2.2 Methods based on multi-state models

We consider methods based on specific models for multi-state processes below. First, we describe an alternative approach based on marginal Kaplan-Meier estimates which was suggested by Glasziou *et al.* (1990) and developed more formally by Pepe *et al.* (1991). Let $T_j^{(\ell)}$ and $W_j^{(\ell)}$ denote the times of the $\ell$'th entry and exit from state $j$, respectively. The prevalence functions can then be written for $j = 1,\ldots,K-1$ as

$$p_j(t) = \sum_{\ell=1}^{\infty} [Pr(T_j^{(\ell)} \leq t) - Pr(W_j^{(\ell)} \leq t)] \tag{2.6}$$

and the expected total time spent in state $j$ over the interval $(0,t)$ as

$$\begin{aligned} \mu_j(t) &= \int_0^t p_j(u)du \\ &= \sum_{\ell=1}^{\infty} [E(W_j^{(\ell)} \wedge t) - E(T_j^{(\ell)} \wedge t)] \qquad j = 1,\ldots,K-1. \end{aligned} \tag{2.7}$$

The example in Section 3 involves the progressive model shown in Figure 1. In this setting we let $T_k^{(1)} = T_k$, $k = 1,2,\ldots,K$ and $W_k^{(1)} = W_k$, $k = 1,2,\ldots,K-1$ since each state can be visited only once, and note that $T_1 = 0$, $T_2 = W_1$, $T_3 = W_2$, and $T_4 = W_3$. Then (2.6) gives

$$p_j(t) = Pr(W_j \geq t) - Pr(W_{j-1} \geq t) = S_j(t) - S_{j-1}(t), \;\; t > 0, \;\; j = 1,2,3, \tag{2.8}$$

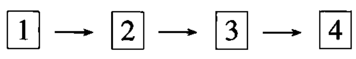

*Figure 1*: *A Progressive Model.*

where $W_0 = 0$. Assuming that censoring times $\tau_i$ are independent of $W_1$, $W_2$, $W_3$, we can estimate the $S_j$'s with standard Kaplan-Meier estimates. Specifically, let $N_{ij}(t)$ denote the counting process recording the number of transitions out of state $j$ (and into state $j+1$) by subject $i$ up to time $t$ and let $\delta_{ij}(t) = I(t \leq \min(W_{ij}, \tau_i))$ denote the "at risk" indicators for exit from state $j$. Then if $w_{\ell j}$ ($\ell = 1, 2, \ldots, L_j$) denote the distinct times of transitions from state $j$ to state $j+1$ across all individuals, the Kaplan-Meier estimate for $S_j(t) = Pr(W_j \geq t)$ is

$$\hat{S}_j(t) = \prod_{w_{\ell j} \leq t} \left(1 - \frac{dN_{.j}(w_{\ell j})}{\delta_{.j}(w_{\ell j})}\right), \tag{2.9}$$

where dots indicate summation over $i = 1, \ldots, n$ and $dX(t) = X(t) - X(t-)$ for any right-continuous process.

The approach just described uses the progressive nature of the process in Figure 1, but is readily extendible to any other multi-state processes of the type considered here, through (2.7) and the use of Kaplan-Meier estimates for the survivor functions of the random variables $T_j^{(\ell)}$ and $W_j^{(\ell)}$. Pepe (1991), Pepe *et al.* (1991), and Couper and Pepe (1997) discuss specific types of processes, and variance and covariance estimates for the Kaplan-Meier estimates.

Prevalence functions can also be estimated by developing a full probabilistic model for the multi-state process. This can be done by specifying transition intensities, denoted here by

$$\lambda_{kk'}(t|\overline{Y}(t)) = \lim_{\Delta t \downarrow 0} \frac{P(Y(t+\Delta t) = k'|\overline{Y}(t), Y(t) = k)}{\Delta t} \qquad k \neq k'.$$

Methods based on Markov models where $\lambda_{kk'}(t|\overline{Y}(t)) = \lambda_{kk'}(t)$ are well known, and nonparametric estimation of transition probabilities is given by the Aalen-Johansen estimates (Andersen *et al.* 1993, Section 4.4). Couper and Pepe (1997), Aalen *et al.* (2001) and Datta and Satten (2001) point out that the Aalen-Johansen estimator of the prevalence functions, while formally justified under a Markov assumption, in fact provides a consistent estimate of the state occupancy probabilities (prevalence functions) for non-Markov processes. To show this, Datta and Satten (2001) consider the "partially conditioned transition rate" (Pepe and Cai, 1993),

$$\alpha_{kk'}(t) = \lim_{\Delta t \downarrow 0} \frac{P(Y(t+\Delta t) = k'|Y(t) = k)}{\Delta t}, \quad k \neq k',$$

with $\alpha_{kk}(t) = -\sum_{k' \neq k} \alpha_{kk'}(t)$, as well as a corresponding integrated transition rate $A_{kk'}(t) = \int_0^t \alpha_{kk'}(u)du$ with matrix form $A(t) = \{A_{kk'}(t)\}$. The $\alpha_{kk'}(t)$ are also the transition intensity functions for Markov models but not for non-Markov models. Let $N_{ikk'}(t)$ denote the cumulative number of transitions from state $k$ to $k'$ over $(0, t]$ for subject $i$ and $N_{.kk'}(t) = \sum_{i=1}^n I(t \leq \tau_i) N_{ikk'}(t)$. Let $Y_{ik}(u) = I(Y_i(u^-) = k)$, and $Y_{.k}(u) = \sum_{i=1}^n I(u \leq \tau_i) Y_{ik}(u)$. The Markov process estimator is the Nelson-Aalen estimate of

$A_{kk'}(t)$,

$$\hat{A}_{kk'}(t) = \int_0^t \frac{I(Y_{.k}(u) > 0)\, dN_{.kk'}(u)}{Y_{.k}(u)}.$$

Product integration gives the Aalen-Johansen estimate of the transition probability matrix over $(0,t)$ as

$$\widehat{P}(0,t) = \prod_{(0,t]} (I + d\hat{A}(u)). \tag{2.10}$$

If $p(0) = (p_1(0), \ldots, p_K(0))'$ is the initial probability vector, then the prevalences (1.6) at time $t$ are estimated as $p(0)'\hat{P}(0,t)$. The estimate (2.10) is not robust to departures from the Markov model, but Couper and Pepe (1997), Aalen *et al.* (2001), and Datta and Satten (2001) show that the estimates of the prevalence functions $p_j(t)$ are robust to departures from the Markov model under the assumption that censoring times $\tau_i$ are independent of the multi-state processes. Glidden (2002) discusses variance estimation for the $\hat{p}_j(t)$'s.

If there is no censoring until after time $t$, then the Glasziou-Pepe estimates of $p_j(t)$ based on (2.6) and the Markov (Aalen-Johansen) estimates based on (2.10) are identical and equal to the observed prevalences $\sum_{i=1}^n I(Y_i(t) = j)/n$. Obtaining variance estimates for

$$\hat{\mu}(t) = \sum_{j=1}^{K-1} \int_0^t v_j(u)\hat{p}_j(u)du \tag{2.11}$$

in the general case is messy via delta method techniques (see Praestgaard 1991 for similar calculations) and bootstrap methods seem the best approach. We also note that the Cook-Lawless (1997) estimate of $\mu(t)$ based on (2.4) and (2.5) uses, under (1.3),

$$\begin{aligned} d\hat{M}(u) &= \frac{\sum_{i=1}^{n}\sum_{j=1}^{K-1} v_j(u) I(X_i \geq u) I[Y_i(u) = j] du}{\sum_{i=1}^{n} I(X_i \geq u)} \\ &= \sum_{j=1}^{K-1} v_j(u) \left\{ \frac{\sum_{i=1}^{n} I(X_i \geq u) I[Y_i(u) = j]}{\sum_{i=1}^{n} I(X_i \geq u)} \right\} du \end{aligned}$$

and so differs from (2.11) by the use of an empirical prevalence estimate $\tilde{p}_j(u)$ instead of the Glasziou-Pepe or Aalen-Johansen estimates. It is identical to the other estimates when there is no censoring until after time $t$, but might be expected to be less efficient with censored data.

The methods in this section assume that censoring is completely independent of the multi-state process. Thus, for example, if censoring were state-dependent, bias could occur. The next section discusses ways of dealing with this.

### 2.3 State-dependent censoring

The assumption of general independent censoring (e.g. Andersen *et al.* 1993, pp. 139-40; Kalbfleisch and Prentice 2002, pp. 194-5) implies that at any time the transition intensities for individuals that are under observation are representative of those in the population of interest at that time. This allows consistent estimation of hazard functions or state transition intensities. Thus, the transition probability matrix (2.10) and associated prevalence estimators are consistent when there is general independent censoring, provided the multi-state model is Markov. This means that censoring does not have to be fully independent of the multi-state process, but could be state-dependent. However, the Glasziou-Pepe estimators based on (2.8) and (2.9) are not valid under state-dependent censoring. To illustrate this, consider the estimate for $S_3(t)$ in (2.8), which from (2.9), has jumps determined by

$$\begin{aligned} d\hat{H}_3(t) &= \frac{dN_{\cdot 3}(t)}{\delta_{\cdot 3}(t)} \\ &= \frac{dN_{\cdot 34}(t)}{Y_{\cdot 1}(t)+Y_{\cdot 2}(t)+Y_{\cdot 3}(t)}, \end{aligned} \tag{2.12}$$

where in the second expression we switch to the multi-state notation. If the model is Markov with censoring intensities $\lambda_{jc}(t)$, $j=1,2,3$ from states 1, 2, and 3 then $d\hat{H}_3(t)$ does not estimate

$$dH_3(t) = dA_{34}(t)\left[\frac{p_3(t)}{p_1(t)+p_2(t)+p_3(t)}\right] \tag{2.13}$$

in general, but instead estimates $dA_{34}(t)P_3^*(t)$, where $P_3^*(t)$ is the probability an individual is in state 3, given that they are in states 1, 2, or 3 (and thus uncensored). The quantity $dA_{34}(t)P_3^*(t)$ equals (2.13) only if $\lambda_{1c}(t)=\lambda_{2c}(t)=\lambda_{3c}(t)$.

Robins (1993), Satten *et al.* (2001) and others have suggested a way to adjust estimators for adaptive censoring, by identifying internal time-dependent covariates, denoted by $Z(t)$, with history $\overline{Z}(t)=\{Z(u), 0\le u<t\}$, such that at time $t$ the censoring intensity satisfies

$$\lambda_C(t|\overline{Z}(t), T\ge t) = \lambda_C(t|\overline{Z}(t)).$$

They propose the use of "inverse probability of censoring weighted" estimates for survival probabilities and other quantities. In the context of survival times $T$, let

$$K_i(t) = \prod_{s\le t}[1-d\Lambda_C(s|\overline{Z}_i(s))]$$

where $d\Lambda_C(t|\overline{Z}_i(t)) = \lambda_C(t|\overline{Z}_i(t))dt$. Robins (1993) and Satten *et al.* (2001) consider $\bar{N}(t)=\sum_{i=1}^{n} I(t_i\le \min(t,\tau_i))/K_i(t_i-)$ and $\bar{\delta}(t)=\sum_{i=1}^{n} I(t\le \min(t_i,\tau_i))/K_i(t^-)$. They prove that $E(\overline{N}(t))=E(N^*(t))$, where $N^*(t)=\sum_{i=1}^{n} I(t_i\le t)$, and $E(\overline{\delta}(t))=E(\delta^*(t))$, where $\delta^*(t)=\sum_{i=1}^{n} I(t\le t_i)$. As a result, if $K_i(t)$ is known, and $w_1,\dots,w_L$ denote the $L$

unique failure times,

$$\bar{S}(t) = \prod_{w_k \le t} \left(1 - \frac{d\bar{N}(w_k)}{\bar{\delta}(w_k)}\right) \tag{2.14}$$

is consistent for $Pr(T \ge t)$. Replacing $K_i(t)$ with a consistent empirical estimate, $\widehat{K}_i(t)$, gives

$$\widehat{N}(t) = \sum_{i=1}^{n} I(t_i \le \min(t, \tau_i))/\widehat{K}_i(t_i-)$$

and

$$\widehat{\delta}(t) = \sum_{i=1}^{n} I(t \le \min(t_i, \tau_i))/\widehat{K}_i(t^-).$$

$\delta(t) = \sum_{i=1}^{n}$ then with known $K_i(t)$, Satten *et al.* (2001) show that

$$\widehat{S}(t) = \prod_{w_k \le t} \left[1 - \frac{d\widehat{N}(w_k)}{\widehat{\delta}(w_k)}\right]$$

is then a consistent estimate of the marginal survivor function, $S(t) = Pr(T \ge t)$.

In the context of progressive multi-state models such as the one represented in Figure 1, inverse probability of censoring-weighted methods can be used to obtain consistent estimates of the distributions for the time to entry/exit of each state. As a result estimates of the state occupancy probabilities based on (2.8) can be corrected for state-dependent censoring.

This method of inverse probability of censoring-weighted estimation was generalized to deal with multi-state processes in Datta and Satten (2002) where the focus was on marginal transition rates. Let $s_{i1}, \ldots, s_{ir_i}$ denote the $r_i$ transition times for subject $i$. Then let

$$\bar{N}_{ikk'}(t) = \sum_{r=1}^{r_i} \frac{I(s_{ir} \le \min(\tau_i, t)) dN_{ikk'}(s_{ir})}{K_i(s_{ir_i}-)}$$

and

$$\bar{Y}_{ik}(t) = Y_{ik}(t) I(t \le \tau_i)/K_i(t-) \ .$$

Replacing $K_i(t)$ with a consistent estimate gives $\widehat{N}_{ikk'}(t)$ and $\widehat{Y}_{ik}(t)$, which give

$$\widehat{N}_{.kk'}(t) = \sum_{i=1}^{n} I(t \le \tau_i) \widehat{N}_{ikk'}(t)$$

and

$$\widehat{Y}_{.k}(t) = \sum_{i=1}^{n} I(t \le \tau_i) Y_{ik}(t)\}.$$

These can in turn be used to compute weighted Nelson-Aalen estimates of the integrated partially conditioned transition rates as

$$\tilde{A}_{kk'}(t) = \int_0^t \frac{I(\widehat{Y}_{.k}(u) > 0) d\widehat{N}_{.kk'}(u)}{\widehat{Y}_{.k}(u)}, \tag{2.15}$$

and an estimate of the transition probability matrix by (2.10). These give estimates of state occupancy probabilities that are robust to departures from the Markov model when there is adaptive or state-dependent censoring.

The weighted Glasziou-Pepe approach and the weighted Aalen-Johansen approach enable one to adjust for state-dependent censoring in estimating state prevalence functions. In general, however, considerable effort may be required to identify a suitable model for $\lambda_C(t|Z(t))$; see Datta and Satten (2002). If censoring is only state- and time-dependent, then $\lambda_C(t|Z(t)) = \lambda_{jC}(t)$ and adjustments are readily made; we consider this in Section 3.

Instead of modelling the censoring intensity, an alternative approach is to specify intensities for the state transitions, thus rendering the censoring process ignorable. State prevalence functions can then be estimated from this model. Intensity based methods for multi-state processes raise issues of model specification and diagnostic checks are essential to assess fit, because the prevalence estimates may be non-robust to departures from the model. Advantages of this approach include a more thorough understanding of the process of interest, and the ability to use the model for prediction. If interest lies solely in the prevalence functions, however, then censoring-adjusted Markov-based estimation is appealing.

## 3 An example

To illustrate the methodology and various points discussed in the preceding sections, we consider a randomized clinical trial of adjuvant chemotherapy for breast cancer that was conducted by the International Breast Cancer Study Group (IBCSG). This study investigated the effectiveness of short duration (one month) and long duration (six or seven months) chemotherapy (e.g., see The Ludwig Breast Cancer Study Group 1988, Gelber *et al.* 1995). A total of 1,229 patients were randomized to treatment: 413 to the short duration treatment and 816 to the long duration treatment. Median folllow-up time was about seven years.

These data have been the subject of various quality of life analyses, based on a four state progressive model as displayed in Figure 1. In this case the four states were 1: Toxicity, 2: Toxicity-free and symptom-free, 3: Relapse, and 4: Death. Quality of life utilities for states 1-4, such as $v_1 = 0.5$, $v_2 = 1.0$, $v_3 = 0.5$, $v_4 = 0$, have been used by many authors; Gelber *et al.* (1995) provide references. We focus here on estimation of the prevalence functions $p_j(t) = P_{1j}(t)$, $j = 1,2,3,4$ for patients in the two treatment groups, but will discuss total quality of life at the end of the section. For the analyses here, we dropped 16 patients for whom one or more state transition times were missing, leaving 411 and 802 subjects in the short and long duration groups, respectively.

Figure 2 shows Kaplan-Meier estimates of the survivor functions $S_j(t)$, $j = 1,2,3$ for the times $T_j$ at which the sojourn in state $j$ ends, for the two treatment groups. The

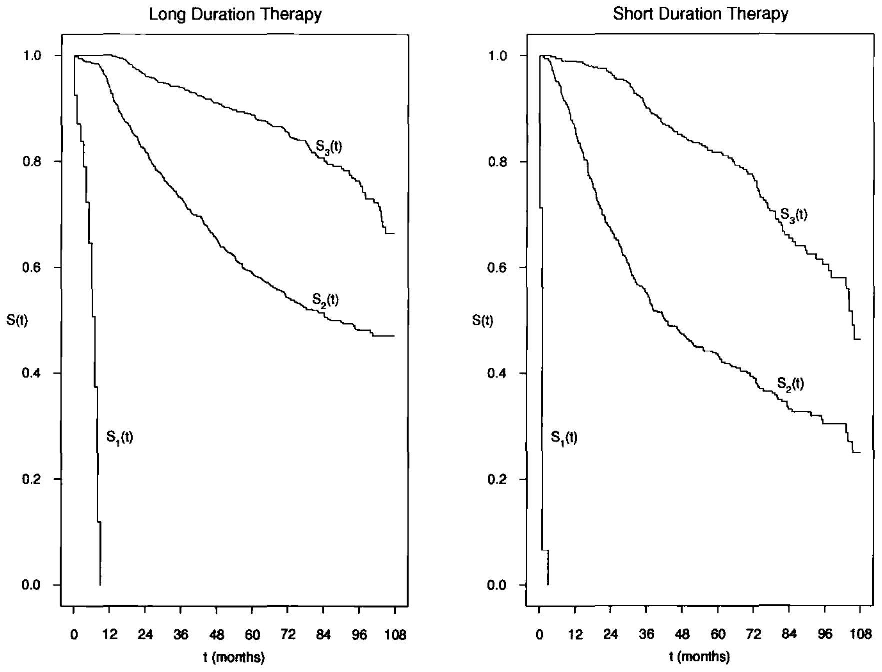

***Figure 2**: Kaplan-Meier estimates for distributions of exit times from states 1, 2, and 3.*

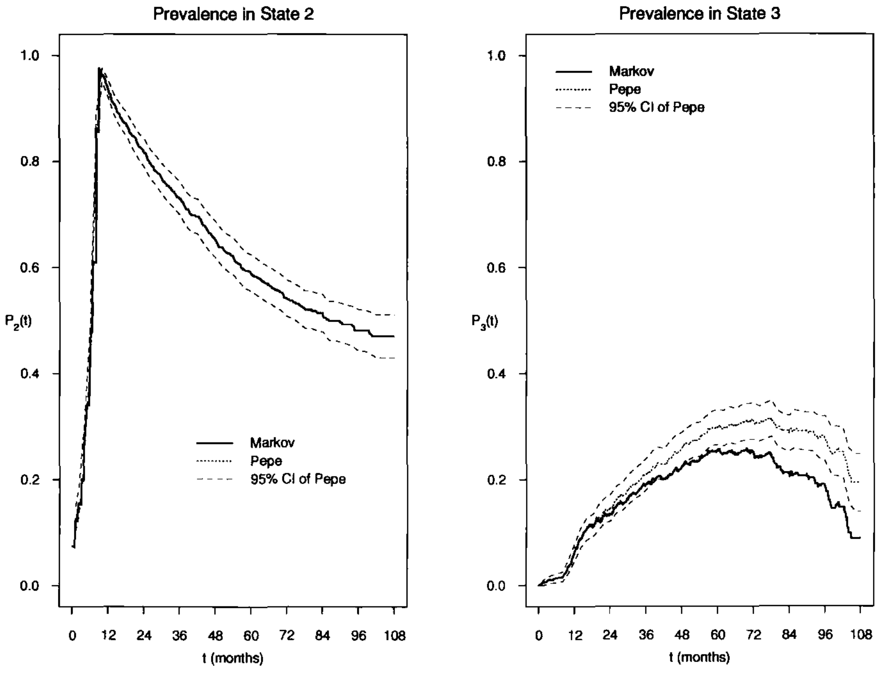

***Figure 3**: Prevalence estimates for states 2 and 3 for the long duration chemotherapy group.*

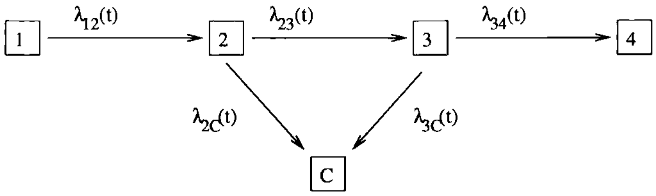

***Figure 4**: The Disease and Censoring Process.*

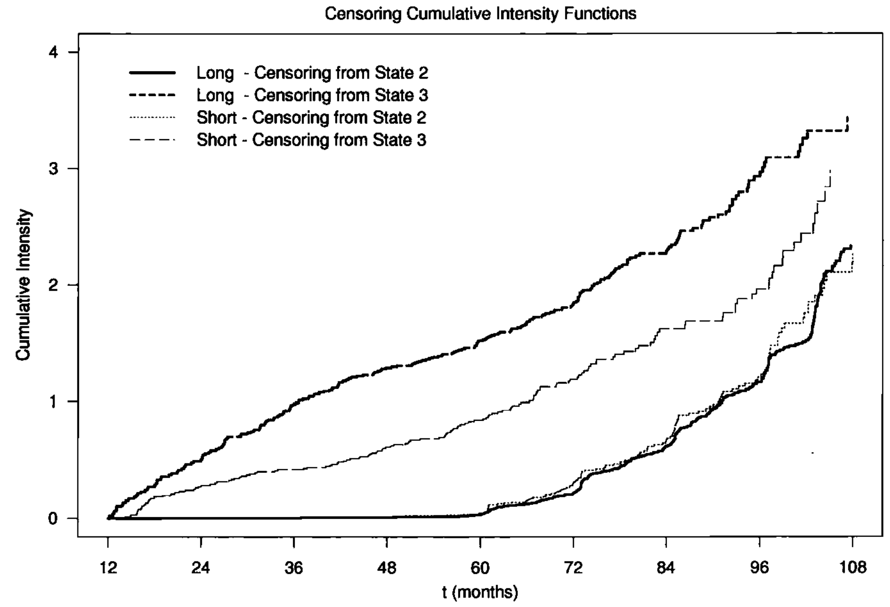

***Figure 5**: Markov cumulative intensity functions for censoring from states 2 and 3.*

prevalence functions can be estimated from (2.8); these estimates are apparent from the figure. Figure 3 shows prevalence estimate $\hat{P}_{12}(t)$ and $\hat{P}_{13}(t)$ based on both (2.8) and the Markov (Aalen-Johansen) estimator (2.10), for the long duration group. The two estimates for $P_{12}(t)$ are virtually identical, but those for $P_{13}(t)$ differ substantially. Pointwise .95 confidence limits for the Glasziou-Pepe estimator (2.8), obtained via 500 nonparametric bootstrap samples, are also shown. Each bootstrap sample was a sample of 802 subjects, drawn with replacement from the 802 long duration chemotherapy subjects. The confidence limits are the estimated prevalence plus or minus 1.96 standard errors, which were estimated from the 500 bootstrap samples.

As discussed in Section 2, the estimates represented in Figure 3 are robust, provided that the censoring mechanism is completely independent of the multi-state process. The Markov estimate is also valid under more general independent censoring (Andersen *et al.* 1993) if the multi-state process is actually Markov. However, the estimates may be biased if these assumptions are not met. It is therefore advisable to assess the censoring process and also the state transition intensities.

An examination of censoring suggests that it is not completely independent of the multi-state disease process. In particular, the censoring intensity at time $t$ on study differs according to whether an individual is in state 1, 2 or 3 at time $t$. First, all individuals spend only a short time (9 months or less) in state 1, and no one in the study was censored while in state 1. Figure 5 shows estimated cumulative censoring intensities $\hat{\Lambda}_{2C}(t)$ and $\hat{\Lambda}_{3C}(t)$ for the Markov model portrayed in Figure 4, where the state $C$ stands for "censored", or withdrawn from the study. Because no individuals were censored from state 2 until well after 12 months on study, and because there were only a very few subjects who progressed to state 3 before 12 months, we have shown the Nelson-Aalen estimates $\hat{\Lambda}_{2C}(t) - \hat{\Lambda}_{2C}(12)$ and $\hat{\Lambda}_{3C}(t) - \hat{\Lambda}_{3C}(12)$ in the figure. Two features are apparent: (1) the censoring intensities from states 2 and 3 are very different for $t \leq 60$ months, and (2) up to about $t = 48$ months the censoring intensity for state 3 is substantially higher for the long duration group than for the short duration group. These features suggest that some individuals were withdrawn from the study after relapse and that this was more pronounced in the long duration chemotherapy group.

Diagnostic checks on the Markov model, for which the transition intensities $\lambda_{12}(t)$, $\lambda_{23}(t)$ and $\lambda_{34}(t)$ in Figure 4 are functions of time on study only, did not show serious departures from the model for either the short or long duration groups. These checks included the introduction of terms in multiplicative models for $\lambda_{j,j+1}(t|\bar{Y}(t))$ that

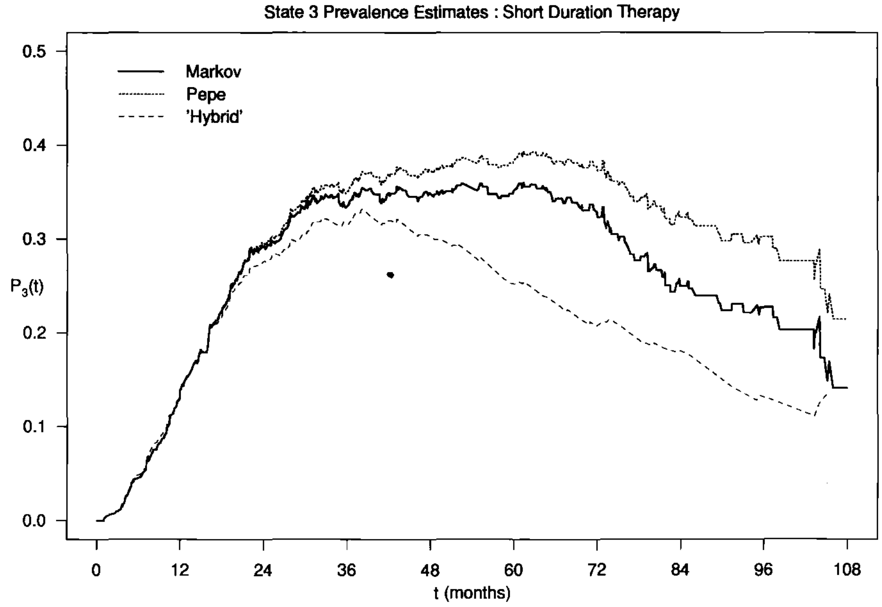

*Figure 6*: *Markov, Glasziou-Pepe, and hybrid model estimates of state 3 prevalence for short duration chemotherapy group.*

represented time since entry to the current state, and sojourn times in previous states. This suggests that the Markov prevalence estimates should be fairly robust to the state-dependent censoring. As a further check, we used censoring-related weights as defined in (2.14) and (2.15), with the censoring intensity $d\Lambda_C(s|\overline{Z}_i(s)) = d\Lambda_C(s|Y_i(s))$ at time $s$ depending on the state occupied. This made very little difference in the Markov estimates for either $P_{12}(t)$ or $P_{13}(t)$. Interestingly, the use of weighted Kaplan-Meier estimation for the Glasziou-Pepe prevalence estimates based on (2.8) also made very little difference, even though the unweighted estimates are affected by state-dependent censoring. The overestimation of state 3 prevalences in Figure 3 (and in Figure 6 below) is as suggested by a comparison of (2.13) with $dA_{34}(t)P_3^*(t)$. In the setting here, $P_3^*(t)$ underestimates the term in square brackets in (2.13), so the estimate (2.9) for $j = 3$ is biased up, as is the estimate of $p_3(t)$ from (2.8).

The transition intensities from states 3 to 4 might be expected to depend on time in state 3 (i.e. time since relapse). Although the checks on the Markov model did not indicate any such dependence, we also fitted a model for which

$$\lambda_{34}(t|\overline{Y}(t)) = \lambda_0(t - t_2)e^{\beta t_2}. \tag{3.1}$$

This model gave a reasonably satisfactory fit, especially for the short duration chemotherapy group. The effect of $t_2$ in (3.1) was highly significant, with $\hat{\beta} > 0$ indicating a negative association between time spent in the toxicity and toxicity-free states, and the relapse state. This effect is sometimes seen in other cancer treatment studies, where patients with longer times to relapse tend to have somewhat shorter survival after relapse. Our preference here is for the simpler Markov model, but we note that if (3.1) is adopted the prevalence estimate for $P_{13}(t)$ becomes

$$\hat{P}_{13}(t) = \int_0^t \exp[-e^{\hat{\beta}u}\hat{\Lambda}_0(t-u)](-d\hat{S}_2(u)). \tag{3.2}$$

Figure 6 shows this estimate along with the Gelber-Pepe and Markov estimates for the short duration group. We see that the new estimate falls substantially below the other two. There is no obvious explanation for this, except that prevalence estimates from semi-Markov models appear to be quite non-robust to model departures (e.g. Couper and Pepe 1997). Another possibility that would also affect the Markov estimates is that some persons were withdrawn from the study because of factors related to future prognosis. This would render the censoring non-independent.

Quality of life (QOL) utilities used by Glasziou *et al.* (1990) and others for this study were $v_1 = 0.1$, $v_2 = 0.5$, $v_3 = 0.1$. Because the Glasziou-Pepe and Markov prevalence estimates differed substantially only for the low-utility state 3, the corresponding estimates of cumulative quality of life, which from (1.5) are

$$\hat{\mu}(t) = \sum_{j=1}^{3} v_j \int_0^t \hat{p}_j(u)du, \tag{3.3}$$

do not differ much. In particular, the estimated mean QOL $\mu(t)$ at $t = 84$ months (as discussed by previous authors) for the long duration group is 29.29 (Markov estimate) or 29.54 (Glasziou-Pepe estimate), with standard errors estimated by 500 bootstrap samples of about 0.43. For the short duration group the corresponding estimates for $\mu(84)$ are 26.20 and 26.35, with standard errors of about 0.65.

## 4 Discussion

Multi-state models often provide an effective way to deal with cumulative cost or quality processes. As indicated, robust estimation of state prevalence functions is possible in many settings; this provides estimates of expected cumulative cost in the settings discussed here. However, the development of full probabilistic models for a multi-state process has the added advantages of providing (in conjunction with the cost model) estimates of the distribution of costs, prediction, and ways of dealing with incomplete data due to intermittent observation or selective sampling of subjects.

There are several areas that deserve further attention. One concerns efficiency and robustness trade-offs among the methods of prevalence function and expected cost estimation discussed in Sections 2 and 3. Limited simulation studies carried out by us and others (e.g. Couper and Pepe 1997, Datta and Satten 2002) for specific multi-state models suggest that the Markov (Aalen-Johansen) estimates are both more efficient and more robust to adaptive censoring than the Glasziou-Pepe estimates. They also suggest that estimates based on semi-Markov models are highly susceptible to departures from the model even under random censoring. Interestingly, the use of censoring-based weighting as described in Section 2.3 seems to have a relatively small effect in many situations involving adaptive or state-dependent censoring. Further study is needed, but it may be that censoring has to be highly adaptive for the weighting to make much difference. In practice there is, of course, the problem of having to model the censoring process in order to produce weights, and the effects of model misspecification here have not been investigated.

It would also be worthwhile to study the estimation of cost distributions, and variance estimation and confidence interval procedures for cost distribution characteristics. Nonparametric bootstrap methods based on resampling individual data histories seem to the most feasible approach at present.

In many applications it may not be feasible to define states in such a way that the cost processes are linear with rates $v_j$, or even deterministic, given the state occupied. A more general cost process that has some degree of tractability is that the cumulative cost up to a duration $s$ for a sojourn in state $j$ is $v_j s + Z_j(s)$, where $\{Z_j(s),\ s \geq 0\}$ is a stochastic process with independent increments. It seems important for tractability and interpretability that we define states so that (1.7) holds, that is, so that state transition intensities are independent of cost history, given the multi-state history. In some cases we may want to stratify individuals or add covariates to the multi-state process in order to achieve this.

### Acknowledgements

This work was supported by grants from the Natural Sciences and Engineering Research Council of Canada and the Canadian Institutes of Health Research (CIHR). R.J. Cook is an Investigator of the CIHR. The authors gratefully acknowledge the International Breast Cancer Study Group for providing the data from the breast cancer trial.

### References

Aalen, O., Borgan, O., and Fekjaer, H. (2001). Covariate adjustment of event histories estimated from Markov chains: the additive approach. *Biometrics*, 57, 993-1001.

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. Springer-Verlag: New York.

Bang, H. and Tsiatis, A.A. (2000). Estimating medical costs with censored data. *Biometrika*, 87, 329-343.

Cook, R.J. and Lawless, J.F. (1997). Marginal analysis of recurrent events and a terminating event. *Statistics in Medicine*, 16, 911-924.

Couper, D. and Pepe, M.S. (1997). Modelling prevalence of a condition: chronic graft-versus-host disease after bone marrow transplantation. *Statistics in Medicine*, 16, 1551-1571.

Datta, S. and Satten, G.A. (2001). Validity of the Aalen-Johansen estimators of stage occupation probabilities and Nelson-Aalen estimators of integrated transition hazards for non-Markov models. *Statistics and Probability Letters*, 55, 403-411.

Datta, S. and Satten, G.A. (2002). Estimation of integrated transition hazards and stage occupation probabilities for non-Markov systems under dependent censoring. *Biometrics*, 58, 792-802.

Gelber, R.D., Cole, B.F., Gelber, S. and Goldhirsch, A. (1995). Comparing treatments using quality-adjusted survival: the Q-TWIST method. *Amer. Statistician*, 49, 161-9.

Ghosh, D. and Lin, D.Y. (2000). Nonparametric analysis of recurrent events and death. *Biometrics*, 56, 554-562.

Glasziou, P.P., Simes, R.J. and Gelber, R.D. (1990). Quality adjusted survival analysis. *Statistics in Medicine*, 9, 1259-1276.

Glidden, D.V. (2002). Robust inference for event probabilities with non-Markov event data. *Biometrics*, 58, 361-368.

Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data*, 2nd edition. John Wiley and Sons: New York.

Lin, D.Y., Feuer, E.J., Etzioni, R. and Wax, Y. (1997). Estimating medical costs from incomplete followup data. *Biometrics*, 53, 419-434.

Pepe, M.S. (1991). Inference for events with dependent risks in multiple endpoint studies. *Journal of the American Statistical Association*, 86, 770-778.

Pepe, M.S. and Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association*, 88, 811-820.

Pepe, M.S., Longton, G. and Thornquist, M. (1991). A qualifier $Q$ for the survival function to describe the prevalence of a transient condition. *Statistics in Medicine*, 10, 413-421.

Prestgaard, J. (1991). Nonparametric estimation of actuarial values. *Scandanavian Journal of Statistics*, 2, 129-143.

Robins, J.M. (1993). Information recovery and bias adjustment in proportional hazards regression analysis of randomized trials using surrogate markers. In *Proc. Amer. Statist. Assoc. Biopharmaceutical Section*, 24-33.

Robins, J.M. and Rotnitzky, A. (1995). Semiparametric regression estimation in the presence of dependent censoring. *Biometrika*, 82, 805-20.

Satten, G.A., Datta, S. and Robins, J. (2001). Estimating the marginal survival function in the presence of time dependent covariates. *Statistics and Probability Letters*, 54, 397-403.

Strawderman, R. (2000). Estimating the mean of an increasing stochastic process at a censored stopping time. *Journal of the American Statistical Association*, 95, 1192-1208.

The Ludwig Breast Cancer Study Group (1988). Combination adjuvant chemotherapy for node-positive breast cancer: inadequacy of a single perioperative cycle. *New England Journal of Medicine*, 319, 677-683.

Torrance, G., Walker, V., Grossman, R., Mukherjee, J., Vaughan, S., LaForge, J. and Lampron, N. (1999). Economic evaluation of Ciprofloxacin compared to usual care for the treatment of acute exacerbation of chronic bronchitis followed for 1 year. *Pharmacoeconomics*, 16, 499-520.

Zhao, H. and Tsiatis, A.A. (1997). A consistent estimator for the distribution of quality-adjusted survival time. *Biometrika*, 84, 339-348.

Zhao, H. and Tsiatis, A.A. (1999). Efficient estimation of the distribution of quality adjusted survival time. *Biometrics*, 55, 231-236.

## Resum

Els beneficis o costos acumulats al llarg del temps per als individus són aspectes d'interès en molts processos sobre la història dels esdeveniments. Exemples familiars inclouen el cost mèdic per a persones amb una malaltia crònica, els pagaments a les persones assegurades durant els períodes de discapacitat, i la qualitat de vida usada de vegades en l'avaluació del tractament en pacients terminals. Usarem aquí el terme cost per a referir-nos al cost o a d'altres mesures acumulades. Hi ha dos escenaris importants: (i) aquell en què els costos estan associats amb l'ocurrència de certs esdeveniments, i en aquests el cost total s'acumula com una funció esglaonada, i (ii) aquell en què els individus es mouen entre diferents estats al llarg del temps, amb un cost que s'acumula a una taxa constant determinada per l'estat que s'ocupa. En ambdós casos, acostuma a definir-se una variable aleatòria T que representa la duració del procés que genera els costos. Considerarem aquí l'estimació del cost miitjà acumulat al llarg d'un període d'interès usant mètodes basats en aspectes marginals dels processos i models d'intensitat. Es discuteix la robustesa dels mateixos per esquemes de censurament adaptatiu en el context de mètodes multi-estat. Els mètodes s'il·lustren amb dades d'un estudi de qualitat de vida amb pacients amb càncer de pit.

*MSC:* 62N01, 62N02, 62N05, 62P20

*Paraules clau:* Anys-vida ajustats per qualitat; censurament adaptatiu, funció de prevalença; mètodes marginals; processos acumulats; processos multi-estat

SORT 27 (1) January-June 2003, 31-40

# A sensitivity analysis for causal parameters in structural proportional hazards models

E. Goetghebeur* and T. Loeys

*Ghent University*

## Abstract

Deviations from assigned treatment occur often in clinical trials. In such a setting, the traditional intent-to-treat analysis does not measure biological efficacy but rather programmatic effectiveness. For all-or-nothing compliance situation, Loeys and Goetghebeur (2003) recently proposed a Structural Proportional Hazards method. It allows for causal estimation in the complier subpopulation provided the exclusion restriction holds: randomization per se has no effect unless exposure has changed. This assumption is typically made with structural models for noncompliance but questioned when the trial is not blinded. In this paper we extend the structural PH model to allow for an effect of randomization per se. This enables analyzing sensitivity of conclusions to deviations from the exclusion restriction. In a colo-rectal cancer trial we find the causal estimator of the effect of an arterial device implantation to be remarkably insensitive to such deviations.

*MSC:* 62N01, 62N02, 62P10, 92B15

*Keywords:* Causal inference; compliance; exclusion restriction; proportional hazards, randomized clinical trials, sensitivity analysis

## 1 Introduction

While the randomized clinical trial remains the gold standard design for causal inference, a thorough analysis of the impact of an intervention should consider treatment actually received besides treatment assigned. The distance between intended and materialized treatments can indeed vary widely, first inside the trial and later under less controlled

**Address for correspondence:* E. Goetghebeur. Dept. of Applied Mathematics and Computer Science, Ghent University, Krijgslaan 281, S9, 9000 Gent, Belgium. els.goetghebeur@rug.ac.be

Received: September 2002
Accepted: January 2003

conditions. Hence the challenge to estimate the effect of different levels of treatment that occur in practice.

The hazard ratio has become the most popular measure of the effect of treatment on survival. Intention-to-treat results are typically cast in those terms, the theory has been well developed and resulting estimators well understood. On the other hand, structural accelerated failure time models (SAFT) as proposed by Robins and Tsiatis (1991) have become the usual tools for 'causal survival analysis' conditional on observed exposures. These models express how survival time can be shrunk or expanded by a parametric function of observed exposures to yield potential treatment-free survival times. In the absence of a direct effect of randomization, potential treatment-free survival times are by design equally distributed between randomized arms. To tap into the proportional hazards tradition and allow for a smooth exchange of information, Loeys and Goetghebeur (2003) developed structural proportional hazards models. They analyzed ECOG-trial E9288 (Kemeny *et al.*, 2002), a randomized clinical trial in colorectal cancer patients with liver metastases. This ECOG multi-centre trial was initiated because the long-term outcome of resection of hepatic metastases remained poor and arterial chemotherapy regimens targeted to the liver had demonstrated high potential. Patients were randomly assigned to either surgical resection alone (control arm, 56 patients) or surgical resection followed by chemotherapy (experimental arm, 53 patients). Interest focused on comparing 5-year survival with and without the implantation. The multi-centric nature of the study made preoperative randomization the practical option. As a result, ten patients who were randomized to receive experimental treatment, did not receive the arterial device implantation, possibly for reasons related to their survival chance.

At the First Barcelona Workshop on Survival Analysis, Ross Prentice raised questions concerning the exclusion restriction given the unblinded nature of the study. It is not inconceivable for instance that the bad news of not being able to receive the implant once it was planned had a negative effect on the patients outcome. Likewise, surgery involving an intended implant might be scheduled earlier in the day, which may have its own impact on survival etc. In response to such concerns this paper sets out to conduct a sensitivity analysis as follows.

In Section 2 we provide a rationale for causal methodology in a proportional hazards framework. Section 3 details the structural proportional hazards approach under the exclusion restriction. In Section 4 we extend the model and adapt the estimation procedure to allow for a sensitivity analysis and examine the impact of violations of the exclusion restriction on the causal PH-estimator. In Section 5 we move on to investigate the joint effect of assignment in both the noncompliant ('the exclusion effect') and compliant ('the causal effect') subpopulation. The methodology is applied to the E9288-data in Section 6 and discussed in Section 7.

## 2 Rationale for a causal proportional hazards estimator

When clinical trial participants fail to adhere to their assigned treatment, a straightforward but naive estimator of the effect of treatment actually received compares patients who were observed to receive the experimental exposure with those who did not. Consider specifically the Cox model

$$h(t \mid E_i) = h_0(t)\exp(\beta_0 E_i), \tag{1}$$

where $E_i$ is the all-or-nothing exposure indicator and $h(t \mid E_i)$ is the hazard rate for failure at time $t$ given exposure. When compliance is selective, i.e. when individuals who comply are prognostically different from those who do not, the parameter $\beta_0$ carries no causal interpretation.

Therefore the most commonly used approach is an intent-to-treat analysis. In proportional hazard terms:

$$h(t \mid R_i) = h_0(t)\exp(\gamma_0 R_i), \tag{2}$$

where $R_i = 1$ indicates the experimental arm. The advantage of this approach is its validity under the null. When experimental treatment has no effect, survival distributions coincide on both randomized arms and $\gamma_0 = 0$ corresponds to the true model. However, in the presence of non-compliance, $\gamma_0$ does not generally measure the biological effect of treatment but rather mixes the effect on compliers with the absence of effect on non-compliers.

To estimate the causal effect of treatment actually received, structural models can be used. Loeys and Goetghebeur (2003) consider

$$h(t \mid R_i = 1, U_i = u) = h(t \mid R_i = 0, U_i = u)\exp(\psi_0 u) \tag{3}$$

where $U_i$ is the potential all-or-nothing exposure for the $i$th subject, that is the exposure that would have been observed had subject $i$ been randomized to experimental treatment. $U_i$ is observed on the experimental arm but latent on the control arm. The Causal Proportional Hazards Effect of Treatment (C-PROPHET) is the log hazard ratio $\psi_0$ in model (3). It compares survival under experimental and potential control conditions in the treatable subgroup $\{U_i = 1\}$. A negative (respectively positive) $\psi_0$ implies a beneficial (respectively harmful) effect of implantation in the treatable subset.

In the subgroup $\{U_i = 0\}$ that would not have been treated when assigned to experimental treatment, no effect of assignment on survival is assumed. Imbens and Rubin (1997) call this assumption the 'exclusion restriction', while Pearl (2002) calls this 'the absence of indirect effect'. The main challenge for inference in model (3) stems from $U_i$ being unobserved in the control arm. We summarize in the next section how $\psi_0$ can be estimated under the exclusion restriction despite ignoring this potential compliance information.

## 3 Inference for the C-PROPHET estimator

If potential receivers of the experimental exposure were known at baseline in both arms, one would fit a proportional hazards model in the subgroup $\{U_i = 1\}$. Denote then Breslow's cumulative baseline hazard estimator for survival in the $\{R_i = 0, U_i = 1\}$-group by $\hat{H}_{01}(t)$. Within the $\{U_i = 1\}$-subset the partial likelihood score equation can be rewritten - in the absence of ties - as

$$\sum_{t_{(j)}} \left[ R_{(j)} - \left\{ \widehat{H}_{01}(t_{(j)}) - \widehat{H}_{01}(t_{(j-1)}) \right\} n_{11j} e^{\psi_0} \right] = 0 \tag{4}$$

where $t_{(j)}$ is the j-th ordered failure time in the treatable subset and $R_{(j)}$ the corresponding assignment indicator. It thus suffices to estimate the jumps of the cumulative hazard for the unobserved subset of compliers in the control arm. $H_{01}$ can be estimated via the corresponding survival estimator $\hat{S}^*_{01}(t)$,

$$\widehat{S}^*_{01}(t) = \{\widehat{S}_0(t) - (1 - \widehat{\pi})\widehat{S}_{10}(t)\}/\widehat{\pi}, \tag{5}$$

with $S_r(t) := \Pr(T_i > t \mid R_i = r)$, $S_{ru}(t) := \Pr(T_i > t \mid R_i = r, U_i = u)$, and $\widehat{\pi}$ the observed compliance proportion in the experimental arm. The Kaplan-Meier estimates $\widehat{S}_0(t)$ and $\widehat{S}_{10}(t)$ will be consistent under independent censoring or the weaker assumption that censoring is non-informative for the control arm as a whole, while in the experimental arm censoring is non-informative conditional on treatment exposure. Frangakis and Rubin (1999) argue that it is sometimes more reasonable to assume non-informative censoring on potential treatment exposure in both arms, and showed that even under this scenario $S_{01}(t)$ is identifiable. Because $\widehat{S}^*_{01}(t)$ is not necessarily monotonic decreasing and found to be a poor estimator for $S_{01}(t)$, Loeys and Goetghebeur (2002) suggested to improve on the proposed estimator via isotonic regression and the 'Pool-Adjacent-Violators' Algorithm (Barlow *et al.*, 1972). To avoid ties in bootstrap samples (Efron, 1981), the jackknife procedure is proposed for variance estimation. Simulation revealed that this procedure provides a somewhat conservative variance estimator in this setting.

## 4 The exclusion restriction: a sensitivity analysis

The exclusion restriction implied by model (3) disallows an effect of assignment for (potential) non-receivers. While this is plausible in double-blind settings, our motivating example E9288 was unblinded.

Consider therefore the following pair of causal models:

$$\begin{cases} h(t \mid R_i = 1, U_i = 0) = h(t \mid R_i = 0, U_i = 0)\exp(\eta_0) \\ h(t \mid R_i = 1, U_i = 1) = h(t \mid R_i = 0, U_i = 1)\exp(\psi_0) \end{cases} \tag{6}$$

In the treatable subset $\{U_i = 1\}$ we consider a proportional hazards effect of exposure as before, but in the $\{U_i = 0\}$-subset we no longer require equality in distribution between randomized arms. Instead, a positive $\eta_0$ in (6) implies that omission of an implantantion that was assigned is bad news added to bad news, i.e. observing the inability of an implant on the experimental arm deteriorates the already bad survival prognosis compared to not observing this on the control arm. This additionally imposed proportional hazard assumption in the $\{U_i = 0\}$-subset can be rewritten in terms of the survival distributions:

$$S_{00}(t) = S_{10}(t)^{\exp(-\eta_0)}. \tag{7}$$

Using equality (7), we obtain a treatment-free survival curve for potential compliers from

$$\widehat{S}^*_{01}(t;\eta_0) = \{\widehat{S}_0(t) - (1-\widehat{\pi})\widehat{S}_{10}(t)^{\exp(-\eta_0)}\}/\widehat{\pi}. \tag{8}$$

As before, the pointwise estimator need not be monotone and isotonic regression on time yields our estimator $\hat{S}_{01}(t;\eta_0)$. Upon substituting the monotonized $\hat{S}_{01}$ to obtain $\tilde{H}_{01}(t;\eta_0) = -\log \widehat{S}_{01}(t;\eta_0)$, we estimate $\psi_0$ in function of $\eta_0$ as

$$\exp(\hat{\psi}_0(\eta_0)) = \frac{\sum_{t_{(j)}} R_{(j)}}{\sum_{t_{(j)}} \{\tilde{H}_{01}(t_{(j)};\eta_0) - \tilde{H}_{01}(t_{(j-1)};\eta_0)\} n_{11j}}. \tag{9}$$

## 5 Joint estimation of effect in the compliant and noncompliant subpopulation

In the previous section $\psi_0$ is estimated as a function of a fixed sensitivity parameter $\eta_0$. Next we investigate joint estimation of $\eta_0$ and $\psi_0$.

The estimation procedure outlined in Section 3 is restricted to all-or-nothing compliance. Indeed, when several levels of compliance are involved, the survival distribution $S_{01}(t)$ is no longer identified without strong additional assumptions. Recently, Loeys and Goetghebeur (2002) proposed an alternative estimation procedure that overcomes this limitation. Specifically, they backtransform observed survival distributions in the experimental arm by exponential functions of the measured exposures to obtain treatment-free survival distributions. Averaging these over all complier subgroups yields an unconditional treatment-free survival curve in the treatment arm. Under the exclusion restriction this should match the corresponding curve on the control arm. Allowing now for an effect of assignment in the $\{U_i = 0\}$-group as in model (6), the idea is to check whether the distribution of observed survival times in the control arm is close to the new mixture of backtransformed survival distributions observed in the experimental arm:

$$S_1 \longrightarrow {}_0(t;\eta,\psi) = \hat{S}_{10}(t)^{\exp(-\eta)}(1-\hat{\pi}) + \hat{S}_{11}(t)^{\exp(-\psi)}\hat{\pi}. \tag{10}$$

Parameter values $\eta$ and $\psi$ which 'equalize' the treatment-free survival distributions between randomized arms are point estimators for $\eta_0$ and $\psi_0$. Because we are estimating two parameters here, two estimating equations are needed. Loeys and Goetghebeur

(2002) combine a logrank and weighted logrank test statistic which are built as sums of 'pseudo' martingale residuals. As the test statistic $Q(\eta,\psi)$ is approximately $\chi^2(2)$, a 95% confidence region for $(\eta_0,\psi_0)$ is formed by the set of $(\eta,\psi)$-values for which $Q(\eta,\psi)$ is below 6.0.

In practice the information may be weak and identification of both $\eta_0$ and $\psi_0$ over-ambitious. Small scale simulations confirm that with limited selectivity (i.e. receivers and non-receivers having a comparable baseline survival prognosis) identifiability problems indeed occur. However with an increasing selection effect both 'causal' effects were reasonably well identified in the simulation setting. Results on the dataset E9288 are described next.

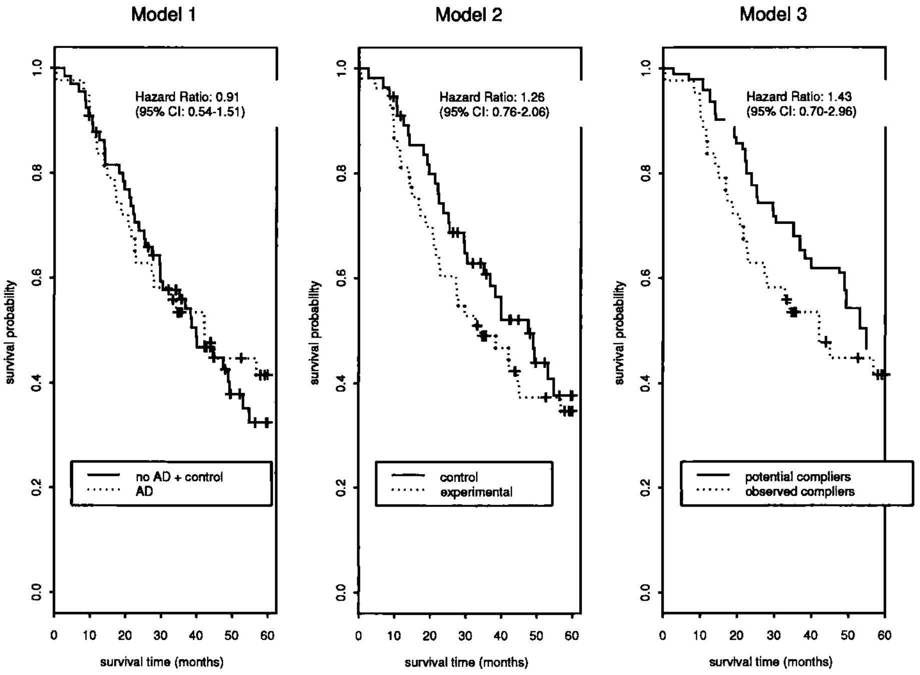

*Figure 1: 3 approaches: (1) As treated (2) Intent-to-treat (3) C-PROPHET. Note: AD stands for Arterial Device).*

### 6 The causal effect of an arterial device implantation

In this section we analyze the E9288-data. Within the 5-year follow-up, 30 patients (54%) died on the control arm compared to 33 patients (62%) on the experimental arm. Figure 1 shows results following models (1), (2) and (3). The as-treated analysis estimates a non-significant beneficial effect of implantation but is only valid under the assumption that non-receivers on the experimental arm form a random subset from the entire population. The estimated effect under the intent-to-treat analysis reveals a non-significant harmful effect of assignment, but does not capture the effect of the implantation actually received. The structural analysis reveals a 43% increase in hazard associated with arterial device implantation in the treatable subset. The latter is derived under the exclusion restriction. Under this assumption, it is further shown in Loeys and Goetghebeur (2003) how the selectivity of patients getting the intervention can be presented by contrasting $\widehat{S}_{01}$, as estimated in (5) with $\widehat{S}_{10} = \widehat{S}_{00}$, as observed in non-compliers in the experimental arm. That plot revealed that patients who would not have received the intervention when assigned to it have a much worse intervention-free survival prognosis than patients that would have received the intervention. This selection effect is also seen in the as-treated estimator, which mixes the causal treatment effect in the treatable subgroup with a diluted selectivity effect, and thus differs from the C-Prophet estimator.

Following Section 4 we can now investigate how sensitive the C-PROPHET estimator is to violations of the exclusion restriction. From Figure 2, we learn that under the assumption of a 50% decrease (respectively increase) in hazard associated with implantation assignment in the untreatable subset, the estimated causal effect equals 1.53 with 95% confidence interval ranging from 0.72 to 3.25 (respectively 1.36, with 95% CI: 0.68 – 2.77). We thus observe that quite substantial deviations from the exclusion restriction have rather limited impact on the estimated causal hazard ratio. Relative to the width of the 95% confidence interval, the change in causal effect as a function of the sensitivity parameter is indeed quite small.

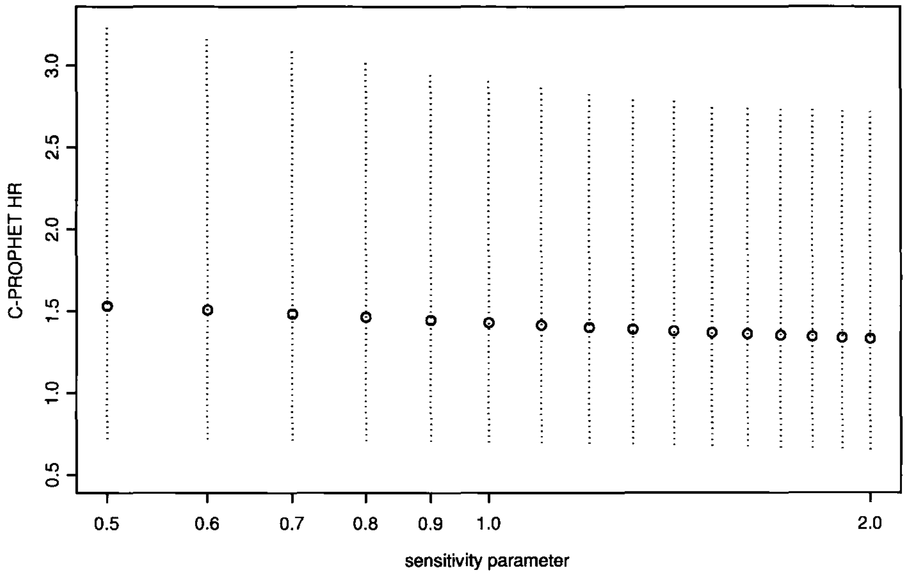

***Figure 2**: The causal hazard ratio* $\exp(\psi_0)$ *(with 95% confidence interval) as a function of the sensitivity parameter* $\exp(\eta_0)$.

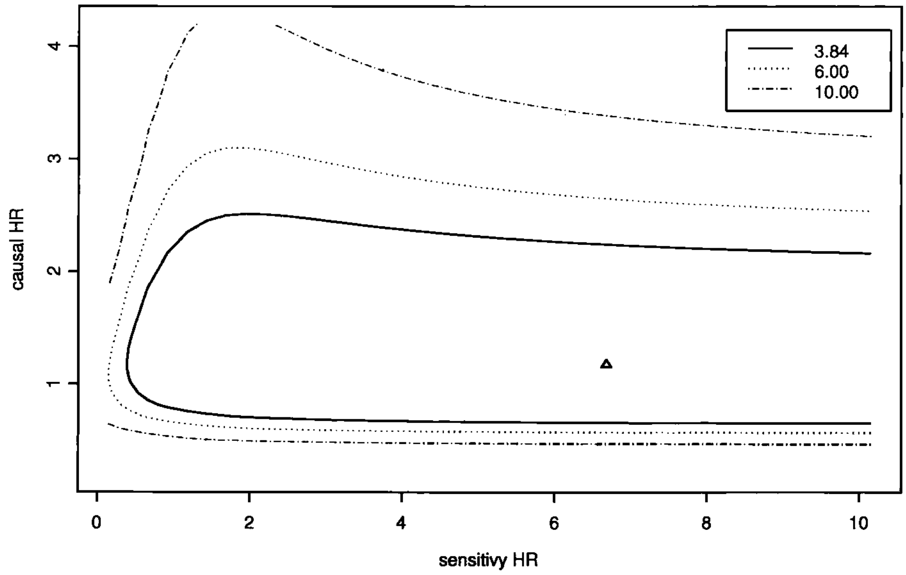

***Figure 3**: Joint estimation of the sensitivity hazard ratio* $\exp(\eta_0)$ *and causal hazard ratio* $\exp(\psi_0)$.

Results of joint estimation of $\eta_0$ and $\psi_0$ as outlined in Section 5 are summarized in Figure 3. Contour plots show the value of the $\chi^2(2)$-test statistic $Q(\eta,\psi)$ as a function of $\exp(\eta)$ and $\exp(\psi)$. Point estimators for $\exp(\eta_0)$ and $\exp(\psi_0)$ equal 6.63 and 1.19. As the 95% confidence region does not close for increasing values of $\eta$, identifiability of $\eta_0$ is rather poor, in contrast to that of $\psi_0$. Nevertheless the data appear to favour the region of $\eta$ suggesting that bad news add more bad news. A substantial positive effect of an intended but absent implant is excluded as a possibility. Surprisingly however we see negligeable impact on estimation of the primary parameter. A marginal 95% confidence interval for $\eta_0$ and $\psi_0$ can be found upon projecting the 3.84-contour on the axes (Robins and Greenland, 1994).

## 7 Discussion

In this paper we presented two approaches to investigate violations against the exclusion restriction in a causal proportional hazards framework. While the approach presented in Section 4 is limited to all-or-nothing compliance, the approach of Section 5 allows for several compliance levels. As identifiability under the second approach relies on the unobserved selectivity, one should be careful when interpreting its results. In a missing data setting Scharfstein (2002) recently discovered that when the signal for inference on $\eta_0$ is weak, it is dangerous to believe the $\eta$-estimate. He therefore favors a sensitivity approach as proposed in Section 4. Interestingly, Scharfstein (2002) found narrow confidence intervals conditional on $\eta$ relative to an enormous range of point estimates for $\psi$ as $\eta$ varies. In contrast we obtain in our motivating example, despite a non-negligeable portion of observed non-compliers on the experimental arm (10 out of 53), changes in the outcome distribution of potential non-receivers with a relatively small impact on the causal parameter estimates.

Further research would be welcomed on the role baseline predictors for treatment-free survival and/or exposure. In as treated model (1) we cannot expect to capture the dependence between treatment-free survival and potential experimental compliance by conditioning on these baseline covariates, and the as treated approach will still give biased results. Loeys and Goetghebeur (2002) propose an estimation procedure allowing to identify population-averaged Causal PROPortional Hazards Effects of Treatment at observed exposure and covariate levels. Conditioning on baseline covariates in model (3) can then address confounding (in the presence of imbalance between randomized arms), conservatism and/or help to keep censoring non-informative.

Finally it is worth remembering that our approach studies survival models conditional on exposure status. Marginal proportional hazards models have recently been introduced by Hernan, Brumback and Robins (2000).

### Acknowledgements

We thank the organizers and other participants of the BWSA for this Golden opportunity for stimulating discussions on survival, and in particular Ross Prentice for posing the question that prompted this analysis. We are grateful to Bob Gray for permission to explore the E9288-data.

### References

Barlow, R.E., Bartholomew, D.J., Bremner, J.M. and Brunk, H.D. (1972). *Statistical Inference under Order Restrictions.* New York: Wiley.

Efron, B. (1981). Censored data and the bootstrap. *Journal of the American Statistical Association,* 76, 312-319.

Frangakis, C. and Rubin, D. (1999). Addressing complications of intention-to-treat analysis in the combined presence of all-or-none treatment noncompliance and subsequent missing outcomes. *Biometrika,* 86, 365-379.

Imbens, G. and Rubin, D. (1997). Bayesian inference for causal effects in randomized experiments with noncompliance. *Annals of Statistics,* 25, 305-327.

Hernßn, M.A., Brumback, B. and Robins, J.M. (2000). Marginal structural models to estimate the causal effect of zidovudine on the survival of HIV-positive men. *Epidemiology,* 11(5), 561-570.

Kemeny, M.M., Adak, S., Gray, B., Macdonald, J.S., Smith, T., Lipsitz, S., Sigurdson, E.R., O'Dwyer, P.J. and Benson, A.B. (2002). Combined-modality treatment for resectable metastatic colorectal carcinoma to the liver: surgical resection of hepatic metastases in combination with continuous infusion of chemotherapy-an intergroup study. *Journal of Clinical Oncology,* 20, 1499-1505.

Loeys, T. and Goetghebeur, E. (2003). A causal proportional hazards estimator for the effect of treatment actually received in a randomized trial with all-or-nothing compliance. *Biometrics*, 59, 100-105.

Loeys, T. and Goetghebeur, E. (2002). Causal proportional hazards models for the effect of treatment actually received in a randomized trial with selective noncompliance. *Technical Report Centrum voor Statistiek, Gent.*

Pearl, J. (2002). Causal inference in the health sciences: a conceptual introduction. In *Health Ser. Outcomes Res. Method* (in press).

Robins, J.M. and Tsiatis, A.A. (1991). Correcting for noncompliance in randomized trials using rank preserving structural failure time models. *Communications in Statistics A*, 20 (8), 2609-2631.

Robins, K.M. and Greenland, S. (1994). Adjusting for differential rates of prophylaxis for PCP in high-versus low-dose AZT treatment arms in an AIDS randomized trial. *Journal of the American Statistical Association*, 89, 737-749.

Scharfstein, D. (2002). Frequentist and bayesian inference for potentially non-ignorable non-response in randomized clinical trials. *Proceedings IBC 2002.*

Sun, J. and Kalbfleisch, J.D. (1993). The analysis of current status data on point processes. *Journal of the American Statistical Association*, 88, 1449-1455.

## Resum

Les desviacions del tractament assignat són comuns en assajos clínics. En aquest context, l'anàlisi tradicional per intenció de tractar no mesura l'eficàcia biològica sinó l'eficàcia de la programació. Per a aquelles situacions on el compliment és total o nul, Loeys and Goetghebeur (2003) proposen un mètode estructural de riscos proporcionals. Aquest mètode permet l'estimació causal en la subpoblació que compleix, sempre que es verifiqui la restricció d'exclusió: l'aleatorització per se no te cap efecte llevat que es canviï l'exposició. Aquesta premissa s'usa en general amb models estructurals per a incompliment però es qüestiona quan l'assaig no és cec. En aquest treball estenem el model estructural de riscos proporcionals de manera que admeti un efecte d'aleatorizació per se. Això ha de permetre analitzar la sensibilitat de les conclusions a les desviacions de la restricció d'exclusió. A un assaig clínic sobre càncer colo-rectal trobem que l'estimador causal de l'efecte de la implantació d'un dispositiu arterial és remarcablement insensible a aquestes desviacions.

*MSC:* 62N01, 62N02, 62P10, 92B15

*Paraules clau:* Anàlisi de sensibilitat; assajos clínics aleatoritzats; compliment; inferència causal; restricció d'exclusió; riscos proporcionals

SORT 27 (1) January-June 2003, 41-64

# Survival analysis with coarsely observed covariates

Søren Feodor Nielsen*

*University of Copenhagen*

## Abstract

In this paper we consider analysis of survival data with incomplete covariate information. We model the incomplete covariates as a random coarsening of the complete covariate, and an overview of the theory of coarsening at random is given. Various ways of estimating the parameters of the model for the survival data given the covariates are discussed and compared.

*MSC:* 62N01, 62N02

*Keywords:* Incomplete data, ignorability, efficiency, EM algorithm, martingale estimating function, inverse probability weighting

## 1 Introduction: incomplete covariates

Statistics is often used to investigate the effect of one or more covariates, $X$, on an outcome of interest, $T$. In order to do this, a conditional model for the distribution of $T$ given $X$, typically in the form of a (linear, generalised linear, hazard, ...) regression, is fitted to the data. If we only look at observations of $(T,X)$ for $X$s fulfilling certain restrictions, which do not depend on $T$ (e.g. $X$ larger than 7, integer valued,...), the conditional distribution of $T$ given $X$ is not affected. In particular, if $X$ is sometimes incompletely observed, restricting attention to cases with $X$ completely observed does not change the conditional distribution of $T$ given $X$ as long as the incompleteness does not depend on the outcome. Thus a complete

*Address for correspondence:* S. Feodor Nielsen. Department of Applied Mathematics and Statistics. University of Copenhagen. Universitetsparken 5. DK-2100 Copenhagen Ø. DENMARK. Phone: (+45) 35 32 07 95. Fax: (+45) 35 32 07 72 feodor@stat.ku.dk

Received: September 2002

Accepted: January 2003

case analysis will lead to correct inference (consistent estimators, valid tests, etc). However, there is clearly a loss of information, as we are effectively reducing the sample size. In some cases this reduction may be considerable. Moreover, if $X$ is multivariate and only partly missing or more generally incomplete so that $X$ is known to lie in a restricted set $y$, a lot of good information may be lost. Consequently, we would like to incorporate cases with incomplete covariates.

If the incompleteness of $X$ depends on $T$, then we have a problem. Imagine for instance that $X$ is incompletely observed if $T \notin y$ for some set $y$. Then a complete case analysis amounts to looking only at cases with $T \in y$. But as $\mathcal{L}(T|X, T \in y) \neq \mathcal{L}(T|X)$, a complete case analysis will fail to give consistent estimators. Here more complicated methods are necessary.

In this paper we will mainly be concerned with the first type of incompleteness, i.e. incompleteness of covariates unrelated to the outcome. In prospective studies where the outcome is measured/recorded at a later stage than the covariate, this would often be case. We will however also touch upon the more general case, where the incompleteness of the covariates is related to the outcome as well as or instead of the covariate. This will often be the case in retrospective studies but also in some prospective studies, for instance if incompleteness is related to prognosis of outcome.

In the next section we give an introduction to coarse data and the concept of coarsening at random in discrete sample spaces. We will use this to model the incomplete covariates. In Section 3 we show how coarsening at random allows us to estimate a conditional survival function and look at the EM algorithm. In Section 4 we extend the concepts of Section 2 to general sample spaces and give a discussion of when censoring is ignorable if the covariates are coarsely observed. Different methods of estimation –including two likelihood based methods and weighted martingale estimation functions– in survival models with coarsely observed covariates are discussed in Section 5. Some conclusions are given in the final section.

## 2 Coarsening at random–discrete case

### 2.1 Coarse data

Let us start of with a simple example of an incomplete covariate. Let $X$ denote smoking status coded as "non-smoker", "light smoker", and "heavy smoker". Imagine that for some individuals we only know that they are not non-smokers. Then the covariate is incompletely observed rather than missing for these individuals, since we do have some information on the value of the covariate: It is either "light smoker" or "heavy smoker".

Let $X$ be a random variable with values in a finite space $E$. If $X$ is not completely observed, it means that all we know is that certain values of $X$ are possible, i.e. that $X \in Y$ for some subset $Y$ of $E$. This subset may be randomly determined as in the example above: It is only for some smokers we do not have a complete observation of smoking status. We call such a random subset $Y$ of $E$ a *coarsening* of $X$. As $Y$ is to represent the possible values of $X$, we will require that $X \in Y$ with probability 1; in particular, $Y \neq \emptyset$.

**Example 1** *Two examples of coarse data are missing data and heaped data:*

- *Missing data. We either observe $X$ or nothing at all:*

$$Y = \begin{cases} \{X\} & \text{if } X \text{ is observed} \\ E & \text{if } X \text{ is missing} \end{cases}$$

- *Heaped data. The covariate is either completely observed or rounded, i.e. we observe either $X$ or $c\lfloor \frac{X}{c} \rfloor$ for some given $c$. Thus*

$$Y = \begin{cases} \{X\} & \text{if } X \text{ is observed} \\ \{c\lfloor \frac{X}{c} \rfloor, c\lfloor \frac{X}{c} \rfloor + 1, \dots, c\lfloor \frac{X}{c} \rfloor + c - 1\} & \text{if } X \text{ is heaped} \end{cases}$$

*A typical example is self-reported smoking; some individuals report a number of cigarettes, others a number of packs smoked.*

*See Heitjan (1993) for more examples.* □

### 2.2 Coarsening at random

Let us go back to the smoking status example. Suppose that the probability of not observing which kind of smoker a person is, does not depend on whether the person is a "light smoker" or a "heavy smoker". Then the incompleteness is ignorable in the sense that it tells us nothing beyond the fact that this person is a smoker. This idea is formalised in the notion of coarsening at random.

**Definition 1** *We say that $Y$ is a* random coarsening *of $X$ if for all $y \subseteq E$*

$$P\{Y = y|X = x\} = q_y \quad \text{for all } x \in y. \tag{1}$$

We shall refer to this condition (1) as CAR (for *coarsening at random*). $(q_y)_{y\subseteq E}$ is called the *coarsening mechanism*; we note that $\sum_{y \ni x} q_y = 1$ for every $x \in E$. Equivalent to condition (1) in Definition 1 is

$$\begin{aligned} P\{Y = y|X \in y\} =& \sum_{x \in y} P\{Y = y|X = x\}P\{X = x|X \in y\} \\ =& P\{Y = y|X = x\} \text{ for all } x \in y \end{aligned}$$

or conditional independence of the coarse observation $Y = y$ and the complete observation $X = x$ given the fact that $X \in y$:

$$\{Y = y\} \underset{\{X \in y\}}{\perp} \{X = x\}, \quad x \in y. \tag{2}$$

Hence observing $Y = y$ tells us nothing more about the unobserved value of $X$ than the fact that $X \in y$. This is the essence of the ignorability mentioned above. For future reference we note that this conditional independence is between three events, which are tied together by $y$; $Y$ is not conditionally independent of $X$ given $X \in y$. A third equivalent condition is that

$$P\{X = x|Y = y\} = P\{X = x|X \in y\} \text{ for all } x \in y.$$

This follows directly from the conditional independence (2) or from the definition (1) and Bayes's theorem.

### 2.3 Estimating $p$

If CAR holds, the likelihood factors into a product of "the likelihood ignoring the incompleteness", $\sum_{x \in y} p(x)$, and the coarsening mechanism, $q_y$:

$$P\{Y = y\} = \sum_{x \in y} P\{Y = y|X = x\} P\{X = x\} = q_y \sum_{x \in y} p(x), \tag{3}$$

where $p(x) = P\{X = x\}$. Hence to estimate $p$ by maximum likelihood we can maximise the naïve log-likelihood

$$L(p) = \sum_{y \subseteq E} n_y \cdot \log \left( \sum_{x \in y} p(x) \right) \tag{4}$$

where $n_y$ is the number of observed subsets of type $y$. In other words, we may ignore the coarsening mechanism when estimating $p$ (by maximum likelihood). This log-likelihood can be maximised using the EM algorithm (Dempster, Laird and Rubin).

Given the marginal distributions of $X$ and $Y$ we can always write $P\{Y = y\} = q_y \sum_{x \in y} p(x)$ by defining $q_y = P\{Y = y\} / \sum_{x \in y} p(x)$. Of course, this will not ensure that $\sum_{y \ni x} q_y = 1$. Gill, van der Laan and Robins (1997, p. 262) seem to suggest that a factorisation such as (3) with $\sum_{y \ni x} q_y = 1$ implies CAR. The reader is invited to show that this is indeed the case for the smoking status example. However, the following example shows that it is not the case in general.

**Example 2** *Let $E = \{1,2,3,4\}$ and $p(x) = \frac{1}{4}$ for $x \in E$. Let $P\{Y = y\} = \frac{1}{4}$ for $y = \{1,2\}, \{1,3\}, \{2,4\}, \{3,4\}$ and 0 for all other subsets of $E$. Then the factorisation holds* *with* $q_y = \frac{1}{2}$ *for* $y$ *with* $P\{Y = y\} > 0$. *In particular,* $\sum_{y \ni x} q_y = 1$ *for every* $x$. *However it is quite possible to have, say,*

$$P\{Y = \{1,2\}|X = 1\} = \frac{2}{3} \qquad P\{Y = \{1,2\}|X = 2\} = \frac{1}{3}$$

*and so on, so that* $P\{Y = y|X = x\} \neq q_y$ *and CAR does not hold.* □

However, it is true that given the marginal distribution of $Y$ there is a marginal distribution of $X$ so that the factorisation (3) holds with $\sum_{y \ni x} q_y = 1$. In other words, we cannot test the hypothesis of coarsening at random. This is not unexpected since CAR is a condition on the distribution of what is observed given what is missing. Gill *et al.* (1997) prove this result by showing that the desired factorisation is obtained by maximising the naïve log-likelihood (4). Moreover, they show that the factorisation is unique in the sense that for any $y \subseteq E$ with $P\{Y = y\} > 0$, $\sum_{x \in y} p(x)$ (and $q_y$) is uniquely determined. In particular, assuming CAR the distribution of $X$ is determined from the distribution of $Y$ if for instance $P\{Y = \{x\}\} > 0$ for all $x$. This is however not a necessary condition, as the following example shows. Another sufficient condition is that the set $y$ with $P\{Y = y\} >$ is a $\pi$-system generating the $\sigma$-field consisting of all subsets of $E$ (e.g. Billingsley 1979), but as the example below shows this is not a necessary condition, either. As an example of $p$ not being identified from the distribution of $Y$ consider the example above. A necessary and sufficient condition for the identifiability of $p$ does not appear to be known and may not exist.

**Example 3** *Let* $E = \{1,2,3,4\}$ *and suppose that* $P\{Y = y\} > 0$ *for* $y = \{1,2\}, \{2,3\}, \{2,4\}$ *only. Then we can identify* $p(1) + p(2)$, $p(2) + p(3)$, *and* $p(2) + p(4)$, *and from this we get*

$$p(2) = 1 - \big(p(1) + p(2) + p(2) + p(3) + p(2) + p(4)\big)/3.$$

*Now the rest follows easily: For instance,* $p(1) = p(1) + p(2) - p(2)$. □

### *2.4 Discrete?*

The "discrete case" in the title of Section 2 refers to the discreteness of the joint distribution of $(X,Y)$. We note that almost everything discussed above carries through to the case where $E$ is countable if the support of the distribution of $Y$ is also at most countable. The only exception is the result about the existence of the CAR factorisation. Indeed, Gill *et al.* (1997) give a counter example (due to Y. Ritov) showing that if the support of the distribution of $Y$ is countable there may be no such factorisation. However since all observed data sets are finite, it is still fair to say that the CAR hypothesis cannot be tested.

### Notes

Coarsening at random was first defined by Heitjan and Rubin (1991) as a generalisation of Rubin's (1976) "missing at random". Their treatment was essentially restricted to the discrete case considered here. Jacobsen and Keiding (1995), Gill *et al.* (1997), and Nielsen (2000) extend their original idea to general sample spaces (see Section 4). Gill *et al.* (1997) also consider the discrete case in detail, and the present presentation is based on their work.

## 3 Survival data

We shall apply the ideas discussed in the previous section to the analysis of survival data with incomplete covariate information. Thus the data we are considering is in the form of a survival time, $T$, a censoring time, $C$, and a covariate, $X$, for each individual. This is the complete data. The observed data is the censored survival time, $T \wedge C$, the censoring indicator, $1_{\{T \leq C\}}$, and a coarsening, $Y$, of $X$. We will work under the assumption of random censoring in the sense of either independence of $T$ and $C$ or conditional independence of $T$ and $C$ given $X$. In many applications the latter assumption is more reasonable: If $T$ is time to death of a specific cause and the censoring includes "death of other diseases", both will usually depend on life style risk factors such as smoking. However, we shall see that conditional independence given $X$ causes problems for many of the methods we will consider.

### 3.1 Estimating the survival function

We will first consider estimating the conditional survival function $\bar{F}(t|x) = P\{T > t|X = x\}$ of $T$ given $X$ based on the censored survival times and the coarsened covariates. Suppose that $Y$ is a random coarsening of $X$ and that it is independent of $T$ given $X$. Then the conditional survival function given $Y = y$ is given by

$$\begin{aligned}\bar{F}(t|y) &= E[1_{\{T>t\}}|Y = y] = E\left[E[1_{\{T>t\}}|X,Y = y]\middle|Y = y\right] \\ &= E\left[E[1_{\{T>t\}}|X]\middle|Y = y\right] = E[\bar{F}(t|X)|Y = y] = \frac{\sum_{x\in y}\bar{F}(t|x)p(x)}{\sum_{x\in y}p(x)}\end{aligned}$$

Thus if the censoring is independent of $T$ and $X$, we may estimate $\bar{F}(t|y)$ by the usual Kaplan-Meier estimator and $p$ from $Y$, plug-in and minimise the sum of squares to obtain an estimator of $\bar{F}(t|x)$. Thus

$$\left[\widehat{\bar{F}}(t|x)\right]_x = (W^\top W)^{-1}W^\top \left[\widehat{\bar{F}}(t|y)\right]_y \text{ where } W = \left[\frac{\hat{p}(x)}{\sum_{x\in y}p(x)}\right]_{y,x}$$

We note that $\widehat{\bar{F}}(t|\{x\})$ actually estimates $\bar{F}(t|x)$; it is just the complete case estimator discussed in the introduction. However, the weighted estimator derived above uses all the data and should therefore be more efficient.

**Example 4** *To illustrate we simulate 4000 datasets with 200 survival times such that $T$ given $X = x$ is exponential with intensity $x$. $X$ is uniform on $\{1,2,3\}$ and coarsened as in the smoking status example such that*

$$P\{Y = y|X = x\} = \begin{cases} 1 & \textit{if } y = \{1\}, x = 1 \\ \frac{1}{2} & \textit{if } y = \{2\}, \{3\}, \textit{ or } \{2,3\} \textit{ and } x \in y \\ 0 & \textit{otherwise} \end{cases}$$

*The censoring is exponential with intensity 2 independent of $T$ and $X$. We give results for $t = 0.2$ in Table 1.*

*Table 1: Estimation of $\bar{F}(t|x)$ for $t = 0.2$: Complete case estimators and weighted estimators (standard deviations in parentheses). Efficiency is of the complete case estimator compared to the weighted estimator.*

| | *True value* | *Complete cases* | | *Weighted estimates* | | *Efficiency* |
|---|---|---|---|---|---|---|
| $X = 1$ | *0.8187* | *0.8185* | *(0.0530)* | *0.8185* | *(0.0530)* | *100%* |
| $X = 2$ | *0.6703* | *0.6693* | *(0.0909)* | *0.6690* | *(0.0809)* | *79%* |
| $X = 3$ | *0.5488* | *0.5484* | *(0.0979)* | *0.5487* | *(0.0874)* | *80%* |

*We see that both complete cases and the weighted estimators are unbiased, and that the weighted estimators are more efficient as we expected. Of course for $X = 1$ the weighted estimator and the complete case estimator are the same.* □

This weighting approach can be used for estimating any conditional functional of the survival distribution, e.g. the conditional hazard.

### 3.2 Maximum likelihood estimation

Under random censoring –in the sense of conditional independence of $T$ and $C$ given $X$– the distribution of the observed data is given by

$$P\{T \wedge C \leq t, 1_{\{T \leq C\}} = \delta, Y = y\} = \sum_{x \in y} P\{T \wedge C \leq t, 1_{\{T \leq C\}} = \delta, Y = y|X = x\}p(x)$$

Hence, if $X$ is coarsened at random and the coarsening $Y$ is independent of the survival data given $X$, the likelihood for the observed data is

$$q_y \cdot \sum_{x \in y} \left(f(t|x)P\{C > t|X = x\}1_{\{T \leq C\}} + h(c|x)P\{T > c|X = x\}1_{\{T < C\}}\right) p(x)$$

where $f$ is the density of $T$ given $X = x$ and $h$ the density of $C$ given $X = x$. The assumption that the coarsening only depends on $X$ may be dispensed with but we leave this case to Section 5.1.

We observe that generally censoring will not be ignorable in the sense of dropping out of the likelihood unless $C$ is actually independent of $X$. Thus if $C$ is only conditionally independent of $T$ given $X$, then we need to specify a model for the censoring in order to maximise the likelihood of $f$ when $X$ is coarsened.

Assuming that $C$ is independent of $(X,T)$ and $Y$ is independent of $(T,C)$ given $X$ we can maximise the likelihood using the EM algorithm. The E-step becomes

$$\sum_{x\in y} \log L_{T\wedge C,1_{\{T\leq C\}}|x}(f)p(x|y,T\wedge C,1_{\{T\leq C\}}) + \sum_{x\in y} \log p(x)p(x|y,T\wedge C,1_{\{T\leq C\}})$$

where

$$L_{T\wedge C,1_{\{T\leq C\}}|X}(f) = f(t|x)1_{\{T\leq C\}} + P\{T > c|X = x\}1_{\{T\leq C\}}$$

and $p(x|y,T\wedge C,1_{\{T\leq C\}})$ is the conditional probability of $X = x$ given $Y = y, T\wedge C, 1_{\{T\leq C\}}$. By the (conditional and unconditional) independence assumptions made we see that

$$p(x|y,T\wedge C,1_{\{T\leq C\}}) = \begin{cases} \dfrac{f(t|x)p(x)}{\sum_{x\in y} f(t|x)p(x)} & \text{when } 1_{\{T\leq C\}} = 1 \\ \dfrac{P\{T > c|X = x\}p(x)}{\sum_{x\in y} P\{T > c|X = x\}p(x)} & \text{when } 1_{\{T\leq C\}} = 0 \end{cases}$$

so that we may ignore the censoring mechanism as well as the coarsening mechanism when estimating the marginal distribution of $X$. It must be stressed that the assumption that $C$ is independent of $X$ in many practical applications will be an unreasonable assumption. In that case we need to estimate the censoring mechanism as well in order to find the maximum likelihood estimator of $f$.

It is tempting, though probably not fully efficient to estimate $p$ by the marginal MLE based on $Y$ and use this estimator in the EM algorithm.

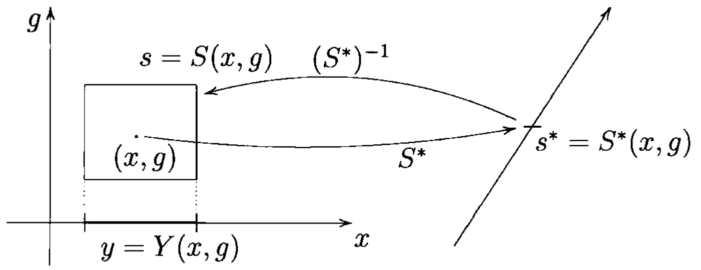

*Figure 1*: *Transformation of $(X,G)$ to $S$.*

*Table 2: Complete case estimator and estimators derived from the EM algorithm: EM I uses a plug-in estimator for p. Standard deviations in parentheses.*

| Method | Complete cases | | EM I | | EM II | |
|---|---|---|---|---|---|---|
| Mean | 1.012 | (0.1323) | 1.008 | (0.1057) | 1.008 | (0.1057) |

**Example 5** We apply these two versions of the EM algorithm to the data simulated in the previous example. We fit a parametric model and include the complete case estimator for comparison. As we expected there is (virtually) no difference between the two EM algorithms (see Table 2). We also note that the efficiency of the complete case analysis compared to the full maximum likelihood estimation is only 64%. In other words, using a complete case analysis and thus discarding on average one third of the observations, we loose a little more than one third of the available information even though the discarded observations are incomplete.

That the loss of information is larger than the fraction of missing observations is due to the differential missingness; the incompleteness only affects observations with $X > 1$. As the incomplete observations are very informative about the true unobserved value of the covariates a lot of information may be regained by a full maximum likelihood estimation. In fact the efficiency of the maximum likelihood estimator based on the observed data compared to the estimator obtained from the uncoarsened data (not shown) is 97.6%. In other words, the coarsening results in an almost negligible loss of information about the regression parameter. □

## 4 Coarsening at random–general case

### 4.1 Extending to general sample spaces

When discussing coarsenings with uncountable support, it seems to be useful to introduce some extra structure on the coarsening. Hence we will assume that there is a *coarsening variable* $G$ deciding the degree of coarseness with which $X$ is observed. For instance, if $X$ is censored, $G$ could be the censoring time. If $X$ is missing, $G$ could be a response indicator taking the value of 1 if $X$ is observed, 0 if $X$ is missing. Typically $G$ may not be completely observed either; if $G$ is a censoring time, then it is only observed if $X$ is censored.

We assume that $G$ is chosen so that there is no additional randomness in the incompleteness mechanism, i.e. that what we actually observe is a non-random function $S^*$ of $(X,G)$. Now, the possible values of $(X,G)$ will be the subset $S = \{(x,g) : S^*(x,g) = S^*(X,G)\}$ of the joint sample space of $(X,G)$. The possible values of $X$ are then represented by the subset $Y$ of $X$'s sample space, $E$, obtained by projecting $S$ onto $E$; see Figure 1.

We note that the extra structure introduced is not a restriction. If we only observe $Y$, "the possible values of $X$", then $Y$ can be used as the coarsening variable $G$ in which case $S = Y \times \{Y\}$. And for $S^*$ we may without loss of generality use $S$, "what is observed about $(X,G)$".

We shall in this paper focus on coarsenings $S$ which takes the form of a product set $S = Y \times H$, where $H$ is a subset of $G$s sample space. Gill *et al.* (1997) calls this a *Cartesian coarsening.*

Extending the theory of CAR to a general sample space is not straightforward. To see why, recall that in the discrete case, CAR was equivalent to a conditional independence of the events $\{Y = y\}$ and $\{X = x\}$ given $\{X \in y\}$, not of $Y$ and $X$ given $X \in y$. Thus in the discrete case CAR is a pointwise, distributional condition. With a general sample space these events will typically have probability 0 making this condition difficult to generalise.

A pointwise formulation is however easily obtained by replacing the condition on the conditional probability (1) by a similar condition on the conditional density of $S$ given $X = x$:

**Definition 2** *$S$ is a random coarsening of $X$ if the conditional density, $k(s|x)$, of $S$ given $X = x$ can be chosen to be independent of $x \in y$, where $y$ is the projection of $s$ onto $X$'s sample space.*

A general expression for the conditional density, $k(s|x)$, may be found in Nielsen (2000).

Densities require a reference measure, and it turns out that different reference measures lead to different conditions. To avoid measure theoretic difficulties we will in this paper use a reference measure, $P_0$, which is a probability measure, and which makes $X$ and $G$ independent. As shown by Jacobsen and Keiding (1995) in a slightly different set-up any product reference measure leads to the same CAR-condition. Expectations with respect to $P_0$ are denoted $E_0$.

**Example 6** *Consider a right censored variable $X$ with censoring time $C$. Here*

$$S = \begin{cases} ]G;\infty[\times\{G\} & \textit{if } X \textit{ is censored} \\ \{X\} \times [X;\infty[ & \textit{if } X \textit{ is observed} \end{cases}$$

*It can be shown (see e.g. Nielsen 2000) that*

$$k(s|x) = \begin{cases} \dfrac{dP\{G \in \cdot|X = x\}}{dP_0\{G \in \cdot\}}(g) & \textit{if } X \textit{ is censored; here } y = ]g;\infty[ \\ \dfrac{P\{G > x|X = x\}}{P_0\{G > x\}} & \textit{if } X \textit{ is observed; here } y = \{x\} \end{cases} .$$

*We see that $S$ is CAR if the conditional density of $G$ given $X = x$ (relative to $P_0$) can be chosen so that it does not depend on $x > g$.* □

As in the discrete case, CAR implies a factorisation of the likelihood.

**Theorem 1** *The density of a random Cartesian coarsening $S$ is $L_f \cdot k(s|x)$, where*

$$L_f = E_0[f(X)|Y = y] = \begin{cases} f(x) & \text{if } y = \{x\} \\ \dfrac{\int_y f(x) dP_0(x)}{P_0\{X \in y\}} & \text{if } P_0\{X \in y\} > 0 \\ ? & \text{otherwise} \end{cases}$$

*where $f$ is the marginal density of the distribution of $X$ with respect to the reference measure $P_0$.*

The question mark is meant to convey the impression that unless $y = \{x\}$ or $P_0\{X \in y\} > 0$ no general expression for the conditional mean is available, not that one cannot be calculated in concrete examples.

### Notes

In general sample spaces Gill *et al.* (1997) distinguish between two type of CAR-conditions, an absolute and a relative. The relative condition, CAR(REL), is formulated in terms of densities as in Definition 2 whereas the absolute, CAR(ABS), is formulated in term of probabilities. The absolute CAR condition is harder to formulate and understand than the relative CAR condition used in this paper. It does however generalise the conditional independence (2) to some extent: If $S$ is an absolute random coarsening, and $s$ is a set such that its projection $y$ has $P\{X \in y\} > 0$, then $X$ is independent of $S = s$ given $X \in y$ (Nielsen 2000, Lemma 4).

Actually for Cartesian coarsenings, independence of $X$ and $G$ implies CAR(ABS). Moreover, any measure which has a density fulfilling Definition 2 with respect to a measure which is CAR(ABS), is in itself CAR(ABS). Thus, even if we have chosen a relative or pointwise formulation in this paper, our choice of reference measure implies the stronger CAR(ABS).

It is clear that by using $Y$ as a coarsening variable, any coarsening may be turned into a Cartesian coarsening. But as $X \in Y$ with probability 1, a product reference measure seems out of the question if $G = Y$. Furthermore, if $S$ is not a Cartesian coarsening, then reducing the observation to $Y$ may not only reduce the available information significantly but it may also make a random coarsening non-random; see Nielsen (2000) for an example.

Extending results of Jacobsen and Keiding (1995), Nielsen (2000) shows that given a statistical model for $(X, G)$ any dominating measure chosen in the model leads to the same CAR(REL) condition. In this sense, choosing a reference measure inside the model is a canonical choice.

### 4.2 Survival data with coarsely observed covariates

We can view survival data with incomplete covariates as a joint coarsening of the survival time $T$ and the covariate $X$. For instance we could let $S^*(T,C,X) = (T \wedge C, 1_{\{T \leq C\}}, Y)$ where $C$ is the censoring time and $Y$ is the coarsening of $X$. Suppose for simplicity that the sample space of $X$ is finite and that $Y$ is a random coarsening of $X$ and independent of $(T,C)$ given $X$. Thus we observe

$$S = \tilde{Y} \times Y \times H = \begin{cases} ]C;\infty[ \times Y & \times \{C\} \quad \text{if } T \text{ is censored} \\ \{T\} \times Y & \times [T;\infty[ \quad \text{if } T \text{ is observed} \end{cases},$$

i.e. a Cartesian coarsening. $S$ is a random coarsening if there exists functions, $K$ and $k$, such that

$$K(t,y) = P\{C > t|X = x, T = t\}P\{Y = y|X = x\} \text{ for all } x \in y$$

and

$$k(c,y) = h(c|t,x)P\{Y = y|X = x\} \text{ for all } x \in y, t > c$$

where $h$ is the conditional density of $C$ given $T$ and $X$. Thus

$$\frac{K(t,y)}{P\{Y = y|X = x\}} = P\{C > t|X = x, T = t\} = 1 - \int_0^t \frac{k(c,y)}{P\{Y = E|y = x\}} dc$$

for all $t \geq 0$ and all $x \in y$, and hence

$$\int_0^t k(c,y)dc = P\{Y = y|X = x\} - K(t,y) \text{ for all } x \in y, t \geq 0.$$

Thus, $P\{Y = y|X = x\}$ cannot depend on $x \in y$, which is equivalent to $Y$ being an random coarsening of $X$. Moreover, it follows that $C$ may only depend on $X$ through $Y$. In particular, if $Y = \{X\}$ almost surely, CAR is just random censoring in the sense that $T$ and $C$ must be essentially independent given $X$ whereas if $Y = E$ almost surely, $C$ must be independent of $X$ given $T$. In the smoking status example, censoring may depend on whether the person is a smoker or not but not on whether the person is a light or a heavy smoker.

If we allow $Y$ to depend on the survival data, it may be possible to allow censoring to depend on $X$ but this dependence must be balanced with the coarsening mechanism in a rather unintuitive way. Thus generally, if the censoring depend on the covariate we cannot expect it to be ignorable. Obviously, if some of the covariates are always completely observed, then censoring may depend on these covariates and still be ignorable. We get similar results for coarsenings in general sample spaces by replacing $P\{Y = y|X = x\}$ by $k(s|x)$.

## 5 Estimation with coarsely observed covariates

In the following subsections we will indicate how some of the existing methods for handling survival data with missing covariates may be extended to handling survival data with coarsened covariates.

### 5.1 Likelihood based estimation

We will first discuss maximum likelihood estimation. Consider a general transformation model:

$$P\{T > t|X\} = 1 - F_\gamma(\log\Lambda(t) + \beta X) \overset{\text{def}}{=} \bar{F}_\gamma(\log\Lambda(t) + \beta X)$$

where $F_\gamma$ is a known distribution function (on $\mathbb{R}$) except possibly for a finite dimensional parameter $\gamma$; with $F_\gamma(t) = \exp(-e^{-t})$ we obtain the Cox regression model. Assuming random censoring, the interesting part of the likelihood of the survival data given $X$ is

$$L_{T\wedge C,1_{\{T\leq C\}}|X}(\Lambda,\gamma,\beta) = \begin{cases} \bar{F}_\gamma(\log\Lambda(T\wedge C) + \beta X) & \text{if } 1_{\{T\leq C\}} = 0 \\ \bar{F}'_\gamma(\log\Lambda(T\wedge C) + \beta X)d\Lambda(T\wedge C) & \text{if } 1_{\{T\leq C\}} = 1 \end{cases}$$

To calculate the likelihood based on the survival data and the coarse observation of $X$, we need to choose a reference measure. The simplest choice is to use a reference measure which makes the survival data and the covariate independent. Thus we let

$$\bar{L}_{T\wedge C,1_{\{T\leq C\}}|X}(\Lambda,\gamma,\beta) = \frac{L_{T\wedge C,1_{\{T\leq C\}}|X}(\Lambda,\gamma,\beta)}{L_{T\wedge C,1_{\{T\leq C\}}|X}(\Lambda_0,\gamma_0,0)}$$

for some suitable choice of $\Lambda_0$ and $\gamma_0$ in the model. When $X$ is coarsened at random and independent of $C$, the likelihood of the observed data becomes

$$\begin{aligned} &L_{T\wedge C,1_{\{T\leq C\}},Y}(\Lambda,\gamma,\beta,f) \\ &= E_0\left[\bar{L}_{T\wedge C,1_{\{T\leq C\}}|X}(\Lambda,\gamma,\beta)f(X)\middle| T\wedge C, 1_{\{T\leq C\}}, Y = y\right] \end{aligned} \tag{5}$$

where the conditional expectation is taken with respect to the chosen reference measure, and $f$ is the density of the marginal distribution of $X$ with respect to this reference measure. As the marginal distribution of $X$ is unknown, $f$ is an unknown parameter either in a finite dimensional space (a parametric family) or an infinitely dimensional space (a semi- or non-parametric model). As in Theorem 1 we see that the likelihood (5) may be written

$$L_{T\wedge C,1_{\{T\leq C\}},Y}(\Lambda,\gamma,\beta,f) = \begin{cases} \dfrac{\int_y \bar{L}_{T\wedge C,1_{\{T\leq C\}}|x}(\Lambda,\gamma,\beta)f(x)dP_0(x)}{P_0\{X\in y\}} & \text{if } P_0\{X\in y\} > 0 \\ \bar{L}_{T\wedge C,1_{\{T\leq C\}}|X}(\Lambda,\gamma,\beta)f(X) & \text{if } Y = \{X\} \end{cases}$$

When maximising the likelihood function (5) we replace $\Lambda$ by a step function with steps at observed deaths.

If $C$ is not independent of $X$, the censoring mechanism must be included in (5) as discussed in Section 3.2.

An alternative to this full maximum likelihood approach is the conditional profile likelihood approach suggested for survival data with missing covariates by Chen and Little (2001). The idea here is to reparameterise:

$$(\gamma,\beta,\Lambda,f) \longrightarrow (\gamma,\beta,R,f)$$

where

$$R(t) = \tau_{\gamma,\beta,f}(\Lambda)(t) = E_0\left[\bar{F}_\gamma(\log\Lambda(t)+\beta X)f(X)\right]$$

is the marginal survival function of $T$. As above the expectation is with respect to the reference measure. Assuming again that $C$ is independent of $X$, censoring is ignorable and $R$ may be estimated from the observed survival data by the usual Kaplan-Meier estimator. Let $\hat{\Lambda}_{\gamma,\beta,f}$ be the result of applying $\tau^{-1}_{\gamma,\beta,f}$ to this estimator. Then the remaining parameters may be estimated from the conditional profile likelihood

$$\frac{L_{T\wedge C,1_{\{T\leq C\}},Y}(\Lambda,\gamma,\beta,f)}{L_{T\wedge C,1_{\{T\leq C\}},E}(\Lambda,\gamma,\beta,f)}$$

Simulations reported by Chen and Little (2001) in the missing covariate case indicate that the loss of efficiency compared to full maximum likelihood estimation is negligible. Note however that if the censoring is not independent of $X$, then specifying a model for the censoring will not help: We need censoring to be ignorable in the marginal model of the survival data.

In both cases the EM algorithm may be useful for the actual maximisation as may Monte Carlo methods. As in subsection  we may estimate $f$ from $Y$ if $Y$ is independent of $(T,C)$ given $X$ and plug it into the likelihood or the profile likelihood.

Both approaches has some clear disadvantages. Firstly, it requires independence of $C$ and $X$, which in applications may be unreasonable. Alternatively, in the full maximum likelihood approach the censoring mechanism must be specified and estimated too, but this will not help us in the conditional profile likelihood approach. Secondly, it requires the marginal distribution of $X$ which is a disadvantage unless $X$ is discrete. Usually we are not interested in this part of the model and would therefore prefer to leave it unspecified. Furthermore, as $X$ is coarsely observed, specifying and checking a model for the marginal distribution of $X$ may be difficult. Finally for the conditional profile likelihood approach we need to be able to invert $\tau_{\gamma,\beta,f}$. The advantage of these methods is that we would expect a high degree of efficiency of both methods.

With a parametric model for the survival data (i.e. with $\Lambda$ either known or known up to a finite-dimensional parameter) the full maximum likelihood approach can still be applied. The conditional profile likelihood approach, however, will typically not be useful as the corresponding marginal survival function, $R$, will now be restricted by the parametrisation and therefore difficult to estimate directly.

### Notes

The EM algorithm for Cox's proportional hazards model with missing covariates has been discussed by a number of authors. Martinussen (1999) and Chen and Little (1999) consider the full likelihood function as done in this paper, whereas Lipsitz and Ibrahim (1998) and Herring and Ibrahim (2001) apply the EM algorithm to the partial likelihood function. The latter two papers also consider the use of Monte Carlo methods in connection to the EM algorithm. See Zhou and Pepe (1995) and Zhou and Wang (2000) for a non-parametric approach. The presentation given here is mainly based on Chen and Little's (1999, 2001) work.

### 5.2 Weighted estimating equations

Another approach to inference in survival analysis is to use martingale estimating functions, i.e. functions like

$$M_s(\theta) = \int_0^s W_s(X,\theta)d\big(N - \Lambda(X;\theta)\big)(s), \quad s \geq 0 \tag{6}$$

where $N$ is the counting process generated by the data, $\Lambda(X;\theta)$ is the integrated hazard, and $W_s(X;\theta)$ is a predictable process (see e.g. Gill (1984) for an introduction). Many popular regression models can be handled in this way. Chen and Newell (2001) consider models with hazards given by

$$\lambda(t|x) = \alpha\big(t \cdot \exp(\gamma X)\big) \cdot \exp(\beta X).$$

Cox regression ($\gamma = 0$) and accelerated failure time ($\beta = 0$) models are obtained as special cases. With $\alpha, \beta$ and $\gamma$ unknown, $\theta = (\alpha,\beta,\gamma)$.

If $Y$ is independent of the survival data, $(T \wedge C, 1_{\{T \leq C\}})$, given $X$, then a complete case analysis works. This corresponds to using the estimating function

$$\Delta M_s(\theta), \quad s \geq 0,$$

where $\Delta = 1$ if $X$ is completely observed, 0 otherwise. Another option is to weight this estimating equation by the "inverse probability" of $X$ being completely observed:

$$\frac{\Delta}{q_{\{X\}}} M_s(\theta), \quad s \geq 0, \quad \text{where } q_{\{X\}} = P\big\{Y = \{X\}\big|X\big\} = E[\Delta|X] \tag{7}$$

The weighted estimating function (7) is unbiased and should therefore yield consistent estimators of the parameters of interest. Generally $q_{\{X\}}$ will be unknown and must be estimated from the data. It is a part of the coarsening mechanism and may be estimated from the conditional likelihood of $S$ (or $Y$) given $X$, which is given by $q_y$ in the discrete case or generally by $k(s|x)$ for any $x \in y$.

**Example 7** *To show the effect of CAR on the estimation of $q_{\{X\}}$ we will consider two examples.*

- *Suppose that $X$ is right censored with censoring variable $G$. Then $q_{\{x\}} = P\{X > G|X = x\} = 1 - \int_0^x h(g)dg$ where $h$ is the conditional density of $G$ given $X = x$ which does not depend on $x$ when we consider $g < x$. One should note that $G$ is not necessarily censored at random; to have both $X$ and $G$ censored at random would require independence of $X$ and $G$. Note also that $h$ is not the marginal density of $G$, unless $X$ and $G$ are independent. In fact $h$ may not even be a density function; it is possible that $\int_0^\infty h(g)dg < 1$. To estimate $q_{\{X\}}$ observe that the conditional likelihood of $S$ given $X = x$ can be written*

$$h(g)^{1-\Delta}\left(1-\int_0^x h(g)dg\right)^{\Delta} = (-dq_{\{g\}})^{1-\Delta}q_{\{x\}}{}^{\Delta}$$

  *In a non-parametric setting, it would be natural to estimate $q_{\{x\}}$ by a decreasing step function with jumps at the observed values of $X \wedge G$.*

- *If $Y$ is a heaping then $q_{\{x\}}$ may be estimated by the fraction of unheaped observations with $\lfloor X/c \rfloor = \lfloor x/c \rfloor$, as*

$$q_{\{x\}} = 1 - P\{Y = y|X = x\} = 1 - P\left\{Y = y \,\middle|\, X^* = c\left\lfloor \frac{x}{c} \right\rfloor\right\}$$

  *for $y = \{c\lfloor\frac{x}{c}\rfloor, c\lfloor\frac{x}{c}\rfloor + 1, \ldots, c\lfloor\frac{x}{c}\rfloor + c - 1\}$.* □

There is no guarantee that this weighting will lead to improved estimators. As such we are still only using complete cases to estimate the parameters of interest even if all the data is used to estimate $q_{\{X\}}$. Dividing by $q_{\{X\}}$ will typically improve the precision of the estimating function but also increase its variance. There appears to be no known sufficient condition to decide if weighting improves the estimator or not. However, estimating the weights, $q_{\{x\}}$, may actually improve the estimator of $\theta$. In fact, the asymptotic variance of $\theta$ will not increase but may well decrease as more parameters are included in the specification of $q_{\{x\}}$. Indeed, letting $q_{\{x\}}$ depend also on the survival data $T \wedge C$ and $1_{\{T \leq C\}}$ may improve the estimation of $\theta$ as the following example shows. We should however keep in mind that the complete case estimator may perform better still.

*Table 3*: *Inverse probability weighting. Efficiency is calculated with respect to the MLE from Table 2.*

| Method | Mean | StDev | Efficiency |
|---|---|---|---|
| Complete cases | 1.012 | 0.1323 | 64% |
| True weights | 1.014 | 0.1376 | 59% |
| Estimated weights I | 1.014 | 0.1378 | 59% |
| Estimated weights II | 1.030 | 0.1186 | 79% |

**Example 8** *Consider again the censored exponential survival times. In Table 3 we compare the complete case estimator of $\theta$ to various estimators of $\theta$ obtained from weighted estimating equations with weights either known ("True weights") or estimated using a model only depending on X (the "true" model; "Estimated weights I") as well as using a model with dependence on X and the survival data $(T \wedge C, 1_{\{T \leq C\}})$ ("Estimated weights II"). We see that in this example complete cases do as well as –if not better than– weighted estimating functions with weights known or estimated using the true model. However using the larger model there is a considerable gain of efficiency.* □

A further advantage of this inverse probability weighting approach is that if the coarsening mechanism depend on $T$ and/or $C$, we can incorporate this by allowing $q_{\{X\}}$ to depend on the survival data $(T \wedge C, 1_{\{T \leq C\}})$:

$$\begin{aligned} q_{\{X\}} &= q(X, T \wedge C, 1_{\{T \leq C\}}) = E\left[\Delta | X, T \wedge C, 1_{\{T \leq C\}}\right] \\ &= P\left\{Y = \{X\} | X, T \wedge C, 1_{\{T \leq C\}}\right\} \end{aligned}$$

Using these weights, the weighted estimating function (7) is still unbiased and should therefore yield consistent estimators.

Still it is not quite satisfactory that the estimation is based on complete cases only even if some improvement due to the estimation of $q_{\{X\}}$ may be expected. Some improvement may be obtained by finding the optimal estimating function (7), where the optimisation is performed over the predictable function $W$. The optimal $W$ will typically depend on the coarsening mechanism and may therefore be unobtainable in practice. Even for the original estimating function (6) the optimal $W$ may be difficult to obtain; see Chen and Newell (2001).

A further improvement on (7) is to add terms of mean 0 to the estimating functions:

$$\frac{\Delta}{q_{\{X\}}} M_s(\theta) + (1 - \Delta)\phi_s(\theta) - \frac{\Delta}{q_{\{X\}}} E\left[(1 - \Delta)\phi_s(\theta) | X, T \wedge C, 1_{\{T \leq C\}}\right] \qquad (8)$$

for some function $\phi_s(\theta) = \phi_s(Y, T \wedge C, 1_{\{T \leq C\}}; \theta)$. By construction the added term has expectation 0 regardless of $\theta$ so that the estimating function (8) will be an unbiased estimating function with the same precision as the simpler weighted estimating function

(7) but lower variance if the added term has a small variance but a large negative correlation with $\frac{\Delta}{q_{\{X\}}} M_s(\theta)$ (see Nielsen (1998) for details). Nielsen (1998) discuss optimal choice of $\phi$ for semi-parametric regression models with coarsely observed regressors; it seems likely that his results may be generalised to the problem and the estimating functions considered here. If so, optimal choices of $\phi$ and $W$ exist (leading to efficient estimates), but they depend on the coarsening mechanism as well as the distribution of $X$ given $(Y, T \wedge C, 1_{\{T \leq C\}})$; thus in practice the optimal choices of $\phi$ and $W$ will hardly be possible to obtain.

It may still be a good idea to add a term, though. One suggestion would be to simply replace the coarsened $X$ in the original estimating function by a suitably chosen value $X^*$ in the coarsening $Y$, i.e. use

$$\phi_s(\theta) = \int_0^s W_s(X^*, \theta) d(N - \Lambda(X^*; \theta))(s)$$

For instance, if $X$ is censored we could use $X^* = X \wedge G$. Still, it should be noted that $E[(1-\Delta)\phi_s(\theta)|X, T \wedge C, 1_{\{T \leq C\}}]$ in most cases will be extremely hard to find.

**Example 9** *We apply this idea to the censored exponentials of the previous examples using $X^* = 2$, 2.5 and 3. The results are reported in Table 4; the first row uses estimated weights depending on X only, the second row weights depending on X and the survival data $(T \wedge C, 1_{\{t \leq C\}})$. We see that adding a term leads to a considerable improvement over the complete case estimator (see Table 3); in fact the efficiency is very close to the efficiency of the maximum likelihood estimator. Furthermore, the choice of $X^*$ does not seem to matter very much. Also the benefit of using a large model for $q_{\{X\}}$ appears to be almost negligible when a term is added to the estimating function.* □

***Table 4**: Estimation with added terms: First row uses weights estimated from the correct model, the second row weights estimated from an extended model. Efficiency is calculated with respect to the MLE from Table 2.*

| | Estimated weights I | | | Estimated weights II | | |
|---|---|---|---|---|---|---|
| Method | Mean | StDev | Efficiency | Mean | StDev | Efficiency |
| $X^* = 2$ | 1.009 | 0.1067 | 98% | 1.009 | 0.1065 | 99% |
| $X^* = 2.5$ | 1.008 | 0.1066 | 98% | 1.008 | 0.1066 | 98% |
| $X^* = 3$ | 1.008 | 0.1084 | 95% | 1.008 | 0.1070 | 98% |

The obvious disadvantage of this approach is that it requires modelling of the coarsening mechanism, at least to the level of modelling the probability, $q_{\{X\}}$, of $X$ being completely observed. Also, as indicated by the simulations it may be as inefficient as the complete case analysis unless additional terms, which depend on the coarsening mechanism, are added to the simple estimating function. Furthermore, we need $q_{\{X\}} > 0$ for all values of $X$ ruling out application to e.g. a covariate that is unobserved due to a fixed detection limit. The advantage of this approach is that it actually allows us to estimate the parameters of interest even if the censoring depends on the covariate $X$ without modelling the censoring mechanism or the marginal distribution of $X$.

### *Notes*

Inverse probability weighting for Cox regression with missing covariates is considered by Pugh, Robins, Lipsitz and Harrington (1993); see also Robins, Rotnitzky and Zhao (1994). Nielsen (1998) considers inverse probability weighting for regression models with coarsely observed covariates.

### *5.3 Bias-variance trade-off*

In many cases a complete case analysis will yield consistent but inefficient estimates. The two approaches discussed in the previous subsections improve the efficiency of the estimators but do this at the cost of much additional work. Moreover they both need specification and estimation of nuisance parts, either the marginal distribution of the covariate or the coarsening mechanism. Another option would be to allow a certain amount of bias in the estimators if the decrease in variance is sufficiently large. Thus in some cases it may be worth considering simply to replace the coarsened value of $X$ by $X^*$ suitably chosen in the observed coarsening $Y$. Unlike the case of missing covariates, the coarsening $Y$ may give a very precise idea about the unobserved value of $X$. Of course, we should expect this imputation approach to lead to biased estimators but also in a reduction of variance compared to the complete case analysis since we are now using all cases. Furthermore, it will be a lot simpler than the methods discussed in the previous subsections. In small samples the reduction in variance may be enough to reduce the mean squared error. However, as the sample size increases the variance will decrease but the bias will not disappear. Hence in large samples the bias will dominate the mean squared error making this approach unacceptable. We illustrate the potential benefits by a simulation example.

**Example 10** *Again we consider the censored exponentials. If we impute either* 2 *or* 3 *when we observe* $Y = \{2, 3\}$, *we get an bias but also a reduction of variance. When we impute* 3, *the reduction is sufficient to make the mean squared error smaller for the biased estimator than for the complete case estimator; see Table 5. If we impute 2.5 there is no bias and the mean squared error is similar to the mean squared error of the MLE reported in Subsection 3.2.*

*If the sample size increases to 1000 then we get worse results. We get roughly the same bias as before but as the variance is smaller, the mean squared errors of the biased estimators (imputing 2 or 3) are now larger than the mean squared error of the complete* *case estimator. Imputing 2.5 is still a good idea, though. This is due to 2.5 being the conditional mean of $X$ given $Y = \{2,3\}$; results by Schafer and Schenker (2000) suggest that the resulting estimator is consistent.* □

***Table 5**: Imputation.*

| | $n = 200$ | | | $n = 1000$ | | |
|---|---|---|---|---|---|---|
| Method | Mean | StDev | MSE | Mean | StDev | MSE |
| Complete cases | 1.012 | 0.1323 | 0.0176 | 1.000 | 0.0591 | 0.0035 |
| Impute $X^* = 2$ | 1.084 | 0.1128 | 0.0198 | 1.075 | 0.0501 | 0.0081 |
| Impute $X^* = 2.5$ | 0.999 | 0.1041 | 0.0108 | 0.991 | 0.0463 | 0.0022 |
| Impute $X^* = 3$ | 0.927 | 0.0975 | 0.0148 | 0.920 | 0.0434 | 0.0083 |

Of course this example is "nice" as fairly little information is lost in the coarsening. How this approach will work more generally is difficult to predict but given its simplicity, it should be considered as an option in small data sets with "small" coarsenings –i.e. coarsenings where $Y$ is a small set–, where a good idea of the true value of $X$ is available and modelling of nuisance parts may be problematic.

### Notes

Imputation has a long tradition as a tool for handling missing or incomplete data; see e.g. Little and Rubin (1987) for an overview. The imputations suggested in this section are naïve and as a consequence they introduce bias. It is possible to construct imputations which will lead to consistent estimators but this will of course make the method more computationally complicated. One possibility is to impute conditional means; this is considered for Cox's proportional hazards model with missing covariates by Paik and Tsai (1997). Another is to generate random imputations for instance by resampling complete cases as done by Paik (1997).

## 6 Conclusions

In this paper we have considered inference for survival data with incompletely observed covariates. We have discussed how ignorable incompleteness may be modelled using random coarsenings and looked at various methods of estimation in these models.

Throughout the paper we have illustrated the estimation methods by a simple simulation example: Exponential regression with independent censoring and a simple coarsening mechanism affecting on average one third of the observations. In this simple example, we have seen that even though a complete case analysis leads to consistent estimators, the loss of information is considerable, and there is a lot to be gained by incorporating cases with incomplete covariates.

All the methods discussed have their own advantages and disadvantages. Most involve some degree of modelling of nuisance parts, either the coarsening mechanism or the marginal distribution of the covariates. Some are very inefficient or yield inconsistent estimators. Which method to use depends on which –if any– nuisance part is easier/safer to specify balanced with the need for efficiency and consistency.

We have also seen how incompleteness in the covariates affect the ignorability of the censoring: If censoring depends on incompletely observed covariates, then it is not (generally) ignorable. This has consequences for most of the methods of estimation we discuss: If the censoring is not ignorable, it needs to be modelled and estimated in order for us to estimate the parameters of interest. The only exception to this rule is the inverse probability weighted estimating equations discussed in Section 5.2. In its simplest form, however, this method may be as inefficient as a complete case analysis, and the conditional expectation needed for the possibly more efficient version (8) will be very difficult if not impossible to calculate in practice.

## 7 Details on the simulations

All simulations in this paper are done using the statistical programming language `R` (Ihaka, R. and Gentleman, R. (1996), www.r-project.org), version 1.5, running on a 1133 MHz Intel Pentium III computer under Suse Linux. The simulations for $n = 200$ were all done in a single function (CPU-time: 13 minutes, 28.65 seconds). More user-friendly functions can be found on `www.stat.ku.dk/~feodor/publications/survival.R`. Approximate CPU-times for the results in Tables 1-4 are respectively $4'58.76''$, $6'33.24''$, $5'43.71''$, $4'58.76''$, and $2'42.47''$ using the user-friendly but less efficient programs and simulating new datasets for each table.

## 8 References

Billingsley, P. (1979). *Probability and measure.* Wiley: New York.

Chen, H. Y. and Little, R. J. (1999). Proportional hazards regression with missing covariates. *Journal of the American Statistical Asociation*, 94, 896-908.

Chen, H. Y. and Little, R. J. (2001). A profile conditional likelihood approach for the semiparametric transformation regression model with missing covariates. *Lifetime Data Analysis*, 7, 207-224.

Chen, Y. C. and Newell, N. P. (2001). On a general class of semiparametric hazards regression models. *Biometrika*, 88, 687-702.

Dempster, A. P., Laird, N. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1-38.

Gill, R. D. (1984). Understanding Cox's regression model: a martingale approach. *Journal of the American Statistical Association*, 79, 441-447.

Gill, R. D., van der Laan, M. and Robins, J. (1997). Coarsening at random: characterisations, conjectures and counter-examples *in* D.-Y. Lin (ed.), *Proc. First Seattle Conference on Biostatistics*. Springer New York, pp. 255-294.

Heitjan, D. F. (1993). Ignorability and coarse data: Some biomedical examples. *Biometrics*, 49, 1099-1109.

Heitjan, D. F. and Rubin, D. B. (1991). Ignorability and coarse data. *Annals of Statistics*, 19, 2244-2253.

Herring, A. H. and Ibrahim, J. G. (2001). Likelihood-based methods for missing covariates in the Cox proportional hazards model. *Journal of the American Statistical Society*, 96, 292-302.

Ihaka, R. and Gentleman, R. (1996). R: A language for data analysis and graphics. *Journal of Computational and Graphical Statistics*, 5(3), 299-314.

Jacobsen, M. and Keiding, N. (1995). Coarsening at random in general sample spaces and random censoring in continuous time. *Annals of Statistics*, 23, 774-786.

Lipsitz, S. R. and Ibrahim, J. G. (1998). Estimating equations with incomplete categorical covariates in the Cox model. *Biometrics*, 54, 1002-1013.

Little, R. J. A. and Rubin, D. B. (1987). *Statistical Analysis with Missing Data*. Wiley: New York.

Martinussen, T. (1999). Cox regression with incomplete covariate measurements using the EM-algorithm. *Scandinavian Journal of Statistics*, 26, 479-492.

Nielsen, S. F. (1998). Semi-parametric regression with coarsely observed regressors. *Preprint 3*. Department of Theoretical Statistics. `http://www.stat.ku.dk/~feodor/publications/`

Nielsen, S. F. (2000). Relative coarsening at random. *Statistica Neerlandica*, 54, 79-99.

Paik, M. C. (1997). Multiple imputation for the Cox proportional hazards model with missing covariates. *Lifetime Data Analysis*, 3, 289-298.

Paik, M. C. and Tsai, W.-Y. (1997). On using Cox proportional hazards model wih missing covariates. *Biometrika*, 84, 579-595.

Pugh, M., Robins, J., Lipsitz, S. and Harrington, D. (1993). Inference in the Cox proportional hazards model with missing covariates. *Technical report*. Department of Biostatistics, Harvard School of Public Health.

Robins, J. M., Rotnitzky, A. and Zhao, L. P. (1994). Estimation of regression coeeficients when some regressors are not always observed. *Journal of the American Statistical Asociation*, 89, 846-866.

Rubin, D. B. (1976). Inference and missing data. *Biometrika*, 63, 581-590.

Schafer, J. L. and Schenker, N. (2000). Inference with imputed conditional means. *Journal of the American Statistical Association*, 95, 144-154.

Zhou, H. and Pepe, M. S. (1995). Auxillary covariate data in failure time regression. *Biometrika*, 82, 139-149.

Zhou, H. and Wang, C.-Y. (2000). Failure time regression with continuous covariates measured with error. *Journal of the Royal Statistical Society, Series B*, 62, 657-665.

## Resum

En aquest treball considerem anàlisis de supervivència amb informació incompleta sobre les covariàncies. Proposem un model per a les covariàncies com una agrupació aleatòria (random coarsening) de la covariància complet, i donem una panoràmica de la teoria de l'agrupació aleatòria (random coarsening). Diverses formes d'estimar els paràmetres del model per a les dades de supervivència donades les covariàncies es discuteixen i es comparen.

*MSC:* 62N01, 62N02

*Paraules clau:* Algorisme EM; dades incompletes; eficiència; funció d'estimació de martingala; ponderació inversa de la probabilitat

SORT 27 (1) January-June 2003, 65-78

# Aspects of the analysis of multivariate failure time data

R. L. Prentice[a] and John D. Kalbfleisch[b]

[a] *Fred Hutchinson Cancer Research Center*

[b] *University of Michigan*

## Abstract

Multivariate failure time data arise in various forms including recurrent event data when individuals are followed to observe the sequence of occurrences of a certain type of event; correlated failure time when an individual is followed for the occurrence of two or more types of events for which the individual is simultaneously at risk, or when distinct individuals have dependent event times; or more complicated multistate processes when individuals may move among a number of discrete states over the course of a follow-up study and the states and associated sojourn times are recorded. Here we provide a critical review of statistical models and data analysis methods for the analysis of recurrent event data and correlated failure time data. This review suggests a valuable role for partially marginalized intensity models for the analysis of recurrent event data, and points to the usefulness of marginal hazard rate models and nonparametric estimates of pairwise dependencies for the analysis of correlated failure times. Areas in need of further methodology development are indicated.

*MSC:* 62N01, 62N02, 62N03,62H10, 92B15

*Keywords:* Correlated failure times; independent censoring; marginal models; survivor function estimation; recurrent events

## 1 Introduction

While univariate failure time methods, including Kaplan-Meier curves, censored data rank tests, and Cox regression methods are well developed, methods for the analysis of

*Address for correspondence: R. L. Prentice. Division of Public Health Sciences. Fred Hutchinson Cancer Research Center. Seattle, WA 98109, U.S.A. rprentic@fhcrc.org, jdkalbfl@umich.edu

Received: September 2002

Accepted: January 2003

multivariate failure times are less unified and their comparative properties have not been extensively studied. Here, we review the state of development of statistical models and methods for the analysis of recurrent event time data and of correlated (or clustered) failure time data. Our aims are to identify known comparative properties of available methods, and to highlight areas for needed research.

There is a long history of point process modelling and estimation for recurrent event data, with emphasis on Poisson and renewal processes (e.g., Cox and Lewis, 1966; Snyder, 1975; Cox and Isham, 1980; Andersen *et al.*, 1993). Cox (1973) discusses these types of models, modulated by regression variables, while other authors (Gail *et al.*, 1980; Prentice et al, 1981) consider more general classes of regression models which allow the intensity rate at a given time to depend on the individual's prior failure history through stratification or regression modeling. Andersen and Gill (1982) give a thorough account of the asymptotic distribution theory for Cox-type modulated Poisson processes using martingale methods. Additional work (Lawless, 1987; Aalen and Huseby, 1991) has added random effects toward extending the applicability of Poisson and renewal process models.

Much of the recent work on recurrent event data analysis has emphasized mean models. These models express the failure intensity at a given follow-up time as a function of regression variables, but do not condition on the individual's preceding failure history (Nelson, 1988, 1995; Lawless and Nadeau, 1995). These models have the attractive feature of providing a simple specification for the expected number of failures as a function of follow-up time. Lin and colleagues provide asymptotic distribution theory for the fitting of Cox-type mean models (Lin *et al.*, 2000), and accelerated failure time models (Lin *et al.*, 1998). However, the independent censoring assumption that attends these regression models may be inappropriately strong. Wang *et al.* (2001) introduce a multiplicative random effect into Cox-type mean models for recurrent events, toward relaxing the independent censoring assumption. Related work focuses on the distribution of gap times between successive events under mean models (Wang and Chang, 1999; Lin *et al.*, 1999).

It is natural to seek a nonparametric estimator of the multivariate survivor function for the analysis of correlated failure time data. Similar to the role played by the Kaplan-Meier estimator for univariate failure time data, such an estimator could form the basis for the display of failure time data, for comparisons among samples, and for regression generalizations. Such an estimator could also allow one to assess the potential of data on auxiliary failure time variables to strengthen the marginal analysis of a failure time variate of interest by exploiting dependent censorship, the so-called auxiliary data problem. Unfortunately the multivariate survivor function estimation problem has yet to be completely solved. There are many possible strongly consistent nonparametric estimators of the multivariate survivor function, but an estimator that is computationally convenient with attractive moderate and large sample efficiency properties has yet to be developed. For example, there are computationally convenient estimators (e.g., Dabrowska, 1988; Prentice and Cai, 1992) of good moderate sample performance, but these estimators are in general not nonparametrically efficient and, in particular, since they use Kaplan-Meier margins, they do not address the auxiliary data problem. On the other hand, van der Laan (1996) has developed a nonparametric maximum likelihood approach to this problem that has the possibility of nonparametric efficiency, but it involves some data reduction, and moderate sample efficiency may be less than that of the simpler estimators. However, available survivor function estimators are either unnecessary for, or adequate for, the study of the relationship between marginal hazard rates and regression variables (e.g., Wei *et al.*, 1989), and for the nonparametric assessment of pairwise dependencies among correlated failure time variables (e.g., Fan *et al.*, 2000).

Subsequent sections amplify the above comments in a manner that relates closely to Chapters 9 and 10 of the second edition of our book on failure time data analysis (Kalbfleisch and Prentice, 2002). Additional general references on multivariate failure time data analysis methods include Hougaard (2000) and Chapters 9 and 10 of Andersen *et al.* (1993). These works place substantial emphasis on random effects or frailty models. In conjunction with Chapter 8 of Kalbfleisch and Prentice (2002) these sources also provide a recent account of the literature on competing risk and more general multistate models for failure time data.

## 2 Recurrent event modelling

Consider a point process of event times $T_1, T_2, \ldots$ on an individual in a study population, and suppose the process is right censored by a censoring time $C$. Often there will be a baseline covariate $x = (x_1, \ldots, x_p)'$ for the individual or, more generally, an evolving covariate process having history $X(t) = \{x(u), 0 \leq u < t\}$ prior to follow-up time $t$. Let $\tilde{N}(t)$ denote the number of failures on an individual by follow-up time $t$; that is, in the time interval $(0,t]$. Also let $N(t)$ denote the observed number of failures on the individual in $(0,t]$. Note that $N(t)$ may be less than $\tilde{N}(t)$ because of the censoring. Data analytic questions of interest may involve the relationship of recurrent event rates to treatment choices, or repair activities, or other aspects of the preceding covariate history. In other instances, questions may involve the relationship of recurrent event rates to the preceding event history. In some applications principal interest may focus on overall event rates in the study population and on the 'population-averaged' relationship of such rates to covariates.

The overall (cumulative) intensity process $\Lambda$ is defined by

$$d\Lambda(t) = E\{d\tilde{N}(t) | \tilde{N}(u), 0 \leq u < t, X(t)\}. \qquad (1)$$

Note that the intensity rate (1) is allowed to depend on both the preceding covariate history and the preceding event history for the individual. In comparison a marginal intensity process, also denoted by $\Lambda$, is defined by

$$d\Lambda(t) = E\{d\tilde{N}(t)|X(t)\}, \tag{2}$$

so that the marginal intensity rate at time $t$ depends on the preceding covariate history, but not the preceding counting process history for the individual. One can also entertain various partially marginalized intensity processes $\Lambda$, as defined by

$$d\Lambda(t) = E[d\tilde{N}(t)|q\{\tilde{N}(u), 0 \leq u < t\}, X(t)] \tag{3}$$

which conditions on some aspects $q\{N(u), 0 \leq u < t\}$ of the preceding counting history, as well as the preceding covariate history. For example $q(\cdot)$ could be defined as $\tilde{N}(t^-)$ which conditions on the number of preceding events on the individual (Pepe and Cai, 1993), as $[\tilde{N}(t^-), 1\{N(t^-) \neq N(t^- - 1)\}]$ which conditions on the number of preceding events along with an indicator of whether the individual has experienced an event in the preceding unit of time. Note that the intensities (1) and (2) are also special cases of (3). Note also that (3) differs from the (continuous time) intensity models $\Lambda_s^*$ of Wei *et al.* (1989) which can be defined for $s = 0, 1, 2, ...$ by

$$d\Lambda_s^*(t) = P\{d\tilde{N}(t) = 1, \tilde{N}(t^-) = s|X(t)\}. \tag{4}$$

The model (4) is somewhat unappealing in the recurrent event setting in that a study subject is considered at risk for a second event at time $t$ without having experienced a first event prior to time $t$. The models (4) are, however, natural and useful for the analysis of correlated failure time, as is elaborated below.

Consider the partially marginalized intensity rate (3) and the ability to (asymptotically) identify $\Lambda$ under right censorship. Such identifiability requires an independent censorship assumption that can be written

$$\begin{aligned} &E[dN(t)|q\{N(u); 0 \leq u < t\}, X(t), \{Y(u), 0 \leq u < t\}] \\ &= Y(t)\Lambda(t) \end{aligned} \tag{5}$$

where the 'at risk' process $Y$ is given by $Y(t) = 1$ if $C \geq t$ and 0 if $C < t$. For independent censorship to hold the censoring rate at follow-up time $t$ can depend on $X(t)$ and $q\{N(u), 0 \leq u < t\}$, but not on other aspects of $\{N(u); 0 \leq u < t\}$. Hence identifiability of the marginal intensity rate (2) requires that censorship not depend in any way on the preceding counting process history, while the overall intensity (1) can be identified under arbitrary dependencies of the censoring on the preceding counting process history.

The same types of regression models can be entertained for recurrent event intensities as for univariate hazard functions. For example one may specify for (3) a relative risk or Cox (1972)-type regression model

$$d\Lambda(t) = d\Lambda_0(t)\exp\{Z(t)'\beta\} \tag{6}$$

where $Z(t) = \{Z_1(t), \ldots, Z_m(t)\}'$ is formed from $q\{N(u); 0 \leq u < t\}$ and $X(t)$, giving a Markov model that is modulated by covariates. A more flexible stratified model would specify

$$d\Lambda(t) = d\Lambda_{os}(t)\exp\{Z(t)'\beta\} \tag{7}$$

where the time-dependent stratification $s = s(t)$ is also formed from $q\{N(u); 0 \leq u < t\}$ and $X(t)$; for example, $s(t) = N(t^-)$. Another class of stratified Cox-type models is given by

$$d\Lambda(t) = d\Lambda_{os}(v)\exp\{Z(t)'\beta\} \tag{8}$$

where $v = t - T_{N(t^-)}$ is the backward recurrence or gap time; that is the time having elapsed since the immediately preceding event (with the convention that $T_0 = 0$). This gives a renewal process that is modulated by covariates.

One could also entertain log-linear or accelerated failure time (AFT) models for (3); for example,

$$d\Lambda(t) = d\Lambda_0\left\{\int_0^t \exp\{Z(u)'\beta\}du\right\}. \tag{9}$$

### 3 Estimation in relative risk models for recurrent events

Now consider estimation of the regression parameter $\beta$ in the Cox-type relative risk model (6). From (3) and (5)

$$dM_i(t) = dN_i(t) - Y_i(t)\exp\{Z_i(t)'\beta\}d\Lambda_0(t)$$

informally has expectation zero under independent censorship. Hence

$$U(\beta) = \sum_{i=1}^{n}\int_0^\infty \{Z_i(u) - \varepsilon(\beta,u)\}\,dM_i(u)$$

has expectation zero where $i$ indexes the sample of $n$ study subjects and

$$\varepsilon(\beta,u) = \sum_{i=1}^{n} Y_i(u)Z_i(u)\exp\{Z_i(u)'\beta\}\Big/\sum_{i=1}^{n} Y_i(u)\exp\{Z_i(u)'\beta\}.$$

Straightforward algebra shows that $M_i$ can be replaced by $N_i$ in $U(\beta)$ so that $U(\beta) = 0$ is an unbiased estimating equation giving rise to the estimate $\hat{\beta}$. Under independent and identically distributed assumptions on $\{N_i, Y_i, Z_i\}$, $i = 1, \ldots, n$ and regularity conditions, Lin *et al.* (2000) use empirical process theory to show that

$$n^{-1/2}U(\beta) \xrightarrow{d} N(0,\Sigma).$$

The variance of the limiting normal distribution is consistently estimated by

$$\hat{\Sigma} = n^{-1}\sum_{i=1}^{n}\hat{U}_i\hat{U}_i',$$

where $\hat{U}_i = \int_0^\infty \{Z_i(u) - \varepsilon(\hat{\beta},u)\} d\hat{M}_i(u)$, $\hat{\Lambda}_0(t) = \int_0^t \sum_{i=1}^n dN_i(u) / \sum_{i=1}^n Y_i(u) \exp\{Z_i(u)'\hat{\beta}\}$ and $\hat{M}_i$ is equal to $M_i$ with $\hat{\beta}$ and $\hat{\Lambda}_0$ in place of $\beta$ and $\Lambda_0$.

It follows that

$$n^{1/2}(\hat{\beta}-\beta) \xrightarrow{d} N\{0, I(\beta)^{-1}\Sigma I(\beta)^{-1}\}$$

and $I(\beta)$ is consistently estimated by $-n^{-1}dU(\hat{\beta})/d\hat{\beta}'$.

These results generalize to the stratified Cox models (7) and (8). Also the same empirical process approach, and a rather similar development, lead to corresponding asymptotic distribution theory for estimation under the accelerated failure time model (9) (Lin *et al.*, 1998).

It is important to note that the covariate history $X(t)$ included in (3) need not be increasing across time, for the estimation procedures just outlined to apply. Thus, for example, models (1)-(3) and the empirical process asymptotic arguments can be used to study the dependence of failure rate on the recent history of an evolving covariate without conditioning on the entire preceding covariate history. This is subject, of course, to the appropriateness of the corresponding independent censorship assumption.

## 4 Bladder tumor illustration

Byar (1980) discusses a randomized trial conducted by the Veteran's Administration Cooperative Urological Group of patients having superficial bladder tumors. One question of interest involved the comparison of tumor recurrence rates following randomization of 48 patients assigned to placebo to that for 38 patients assigned to the drug thiotepa. Trial follow-up continued for 31 months on average with 87 recurrences (recorded in months) among placebo patients as compared to 45 recurrences among thiotepa patients. Individual patients experienced from zero to nine recurrences during follow-up. See Andrews and Hertzberg (1985, pp.254-259) or Kalbfleisch and Prentice (2002, p.292) for a listing of these data. Baseline covariates included the number of bladder tumors for a patient prior to randomization (truncated at 8), and the diameter of the largest such tumor in millimeters.

Table 1 shows related regression analyses of these bladder tumor recurrence rates with an emphasis on the effect of thiotepa treatment. The first analysis (Lin *et al.*, 2000) applies a Cox model (6) to the rates (2). In addition to its interpretation in terms of the failure rates (2), the expected number of recurrences in $(0,t]$ for an individual is proportional to $\exp\{Z(t)'\beta\}$ since $Z(t) = z$ is time independent. This gives a useful mean model interpretation to the regression parameters.

The second analysis (Lin *et al.*, 1998) applies the accelerated failure time model (9) to the rates (2). The Cox model analysis indicates an estimated relative risk of $\exp(-0.524) = 0.59$ for the thiotepa as compared to placebo recurrence rate with a corresponding significance level of 0.05. The AFT model provides a very similar point estimate $(\exp(-0.542) = 0.58)$ for a rescaling of the rate at which a patient traverses the time axis under thiotepa versus placebo. The similarity of all three regression coefficients in these two analyses arises from the fact that $\Lambda_0(t)$ in (9) is approximately proportional to $t$ in this application.

These mean model analyses require the censoring rate at follow-up time $t$ to be independent of the patient's prior recurrence time history. Simple Cox model analyses of censoring rates that include treatment, number of initial tumors and initial tumor size as covariates do not indicate any important dependence of the censoring rate on the number of prior recurrences for a patient, but did show a nearly threefold, highly significant increase in the censoring rate if the patient had a recurrence during the preceding month.

***Table 1**: Regression parameter estimate of the rate of recurrence of superficial bladder tumors under various models.*

| Regression Model | Treatment (0-placebo; 1-thiotepa) | Number Initial Tumors | Initial Tumor Size | Gap Time ($v$) | Recurrence Within Past Month |
|---|---|---|---|---|---|
| Cox model (6) | −0.524 | 0.201 | −0.041 | | |
| for (2) | (0.262)* | (0.064) | (0.076) | | |
| AFT model (9) | −0.542 | 0.204 | −0.038 | | |
| for (2) | (0.312) | (0.066) | (0.084) | | |
| Cox model (7) | −0.346 | 0.122 | −0.017 | −0.082 | −1.387 |
| with $s = N(t^-)$ for (3) | (0.185) | (0.047) | (0.061) | (0.027) | (0.579) |

* Estimated standard errors in parentheses.

Hence a model for the partially marginalized recurrence rate (3) may be needed for a valid analysis of these data with conditioning on $q\{N(u), 0 \leq u < t\}$ that includes at least an indicator of whether a recent recurrence was recorded. The final analysis of Table 1 uses a Cox model (7) that stratifies at follow-up time $t$ on the number of prior tumors $N(t^-)$ and includes in the regression function an indicator of whether a recurrence occurred within the past month, as well as the gap time $(v)$ since the immediately preceding recurrence. All three of these aspects of the preceding counting process history relate strongly to the recurrence rate at a given follow-up time. For example, patients recording a recurrence in the preceding month had an estimated recurrence rate of about one quarter the rate of those without such recurrence $(\exp(-1.387) = .25)$. This lower rate may correspond in part to the withdrawal of patients having a comparatively poor prognosis from further trial participation, arguing for an appropriate control of the preceding counting process history in assessing treatment effects. In fact, the relative recurrence risk associated with thiotepa in this analysis is estimated as $\exp(-0.346) = 0.71$, somewhat closer to the null compared to the other analyses, though some moderate evidence of benefit for thiotepa remains with a standardized test statistic of value $-0.346/0.185 = -1.87$ and corresponding significance level of about 0.06.

### 5 Correlated failure time data analysis

Consider now failure times $\tilde{T}_1, \ldots, \tilde{T}_m$ that may be correlated. For example, these variates may represent times to (ages at) disease occurrence in a family study in genetic epidemiology, or times to the occurrence of $m$ distinct diseases for an individual in a clinical trial or cohort study. Denote by $x = (x_1, \ldots, x_p)'$ baseline covariates corresponding to $(\tilde{T}_1, \ldots, \tilde{T}_m)$. Additionally, there may be evolving covariates $X_j(t_j) = \{x_j(u); 0 \leq u < t_j\}$ corresponding to $\tilde{T}_j, j = 1, \ldots, m$. Topics of interest in the analysis of correlated failure time data include the relationship of marginal hazard rates $d\Lambda_j(t)$ on the corresponding preceding covariate history $X_j(t)$, which for notational convenience can be defined to include the baseline covariate vector $x$; and study of the dependencies among failure times, or failure rates, given covariates.

One can define a hazard rate corresponding to any subset of $\{\tilde{T}_1, \ldots, \tilde{T}_m\}$. For example, an $s$th order hazard rate at $(t_1, \ldots, t_s)$ can be defined, in an obvious notation, by

$$\begin{aligned}&\Lambda_{1\ldots s}\{dt_1, \ldots, dt_s; X_j(t_j), j = 1, \ldots, s\}\\&= P\left\{\tilde{T}_j \in [t_j, t_j + dt_j), j = 1, \ldots, s | \tilde{T}_j \geq t_j, X_j(t_j), j = 1, \ldots, s\right\}.\end{aligned}$$

Suppose that $\tilde{T}_j$ is subject to right censoring by $C_j, j = 1, \ldots, m$, so that one observes $T_j = \tilde{T}_j \wedge C_j$, and $\delta_j = \mathbf{1}(T_j = \tilde{T}_j), j = 1, \ldots, m$. In general a rather strong independent censorship condition is needed to allow the identifiability of hazard rates of all orders. For example, for identifiability of $\Lambda_{1\ldots s}$ one needs to assume

$$\begin{aligned}&P\left\{T_j \in [t_j, t_j + dt_j), \delta_j = 1, j = 1, \ldots, s | Y_j(u); 0 \leq u < t_j, X_j(t_j), j = 1, \ldots, s\right\}\\&= \prod_{j=1}^{s} Y_j(t_j) \Lambda_{1\ldots s}\{dt_1, \ldots, dt_s; X_j(t_j), j = 1, \ldots, s\}, \qquad (10)\end{aligned}$$

with a corresponding assumption for the hazard rates corresponding to other subsets of $\tilde{T}_1, \ldots, \tilde{T}_m$. Such conditions will be fulfilled, for example, with fixed covariates $x$, if $(\tilde{T}_1, \ldots, \tilde{T}_m)$ is independent of $(C_1, \ldots, C_m)$ given $x$. The applicability of an independent censoring assumption must be carefully considered if $\tilde{T}_1, \ldots, \tilde{T}_m$ correspond to the times to disease events on individual study subjects, as potential censoring times for one type of disease may depend on the occurrence times for another type of disease.

Often the questions of interest focus on regression effects on marginal hazard rates which may, for example, be addressed using Cox-type models of the form

$$\Lambda_j\{dt_j; X_j(t_j)\} = \Lambda_{0j}(dt_j) \exp\{Z_j(t_j)'\beta\}, j = 1, \ldots, m. \qquad (11)$$

Under an independent censoring assumption of the type (10) for the marginal rates $\Lambda_1, \ldots, \Lambda_m$ one can construct an unbiased estimating function for $\beta$ as

$$\begin{aligned} U(\beta) &= \sum_{i=1}^{n}\int_0^\infty \sum_{j=1}^{m}\{Z_{ji}(u)-\varepsilon_j(\beta,u)\}\,dM_{ji}(u) \\ &= \sum_{i=1}^{n}\int_0^\infty \sum_{j=1}^{m}\{Z_{ji}(u)-\varepsilon_j(\beta,u)\}\,dN_{ji}(u) \qquad (12)\end{aligned}$$

based on a sample $(T_{1i},\dots,T_{mi}),(\delta_{1i},\dots,\delta_{mi}),i=1,\dots,n$ with

$$\varepsilon_j(\beta,u)=\sum_{i=1}^{n}Y_{ji}(u)Z_{ji}(u)\exp\{Z_{ji}(u)'\beta\}\Big/\sum_{i=1}^{n}Y_{ji}(u)\exp\{Z_{ji}(u)'\beta\}.$$

Under iid conditions on counting, at risk and censoring processes for the $n$ observations, empirical process methods imply (Wei, Lin and Weissfeld, 1989) that

$$n^{1/2}(\hat\beta-\beta)\xrightarrow{d}N\{0,I(\beta)^{-1}\Sigma I(\beta)^{-1}\}$$

where $\hat\beta$ solves $U(\beta)=0$. It further can be seen that $I(\beta)$ is consistently estimated by $-n^{-1}\partial U(\hat\beta)/\partial\hat\beta'$ and $\Sigma$ is consistently estimated by

$$\hat\Sigma=n^{-1}\sum_{i=1}^{n}\hat U_{\cdot i}\hat U_{\cdot i}',$$

where $\hat U_{\cdot i}=\int_0^\infty\sum_{j=1}^{m}\{Z_{ji}(u)-\varepsilon_j(\hat\beta,u)\}\hat M_{ji}(du)$,

$$\hat M_{ji}(du)=N_{ji}(du)-Y_{ji}(u)\exp\{Z_{ji}(u)'\hat\beta\}\hat\Lambda_{0j}(du),$$

and

$$\hat\Lambda_{oj}(du)=\sum_{i=1}^{n}N_{ji}(du)\Big/\sum_{i=1}^{n}Y_{ji}(u)\exp\{Z_{ji}(u)'\hat\beta\}.$$

The estimating function (12) effectively makes a working independence assumption among the correlated failure times. Some modest efficiency improvement is possible by introducing a weight function into (12) (Cai and Prentice, 1995), a topic that relates closely to the auxiliary data problem mentioned above. These methods have been adapted to models (11) that specify a common baseline hazard rate $\Lambda_{oj}\equiv\Lambda_0$ in (11) (Lee *et al.*, 1992; Cai and Prentice, 1997), and AFT models have also been considered for marginal hazard regression modeling (Lin and Wei, 1992).

Now consider the nonparametric estimation of pairwise dependencies from censored correlated failure time data. Pairwise dependency measures can be generated by an appropriate integration of a local dependency measure over a follow-up region of interest. Ignoring covariates and denoting the joint survivor function for $(\tilde T_1,\tilde T_2)$ by $F(t_1,t_2)=P(\tilde T_1>t_1,\tilde T_2>t_2)$, two potential local dependency measures (Oakes, 1989) at a point $(s_1,s_2)$ are the cross ratio

$$\begin{aligned} c(s_1,s_2) &= F(ds_1,ds_2)F(s_1^-,s_2^-)/\{F(s_1^-,ds_2)F(ds_1,s_2^-)\} \\ &= \lambda_1(s_1|T_2=s_2)/\lambda_1(s_1|T_2\ge s_2) \\ &= \lambda_2(s_2|T_1=s_1)/\lambda_2(s_2|T_1\ge s_1),\end{aligned}$$

and a local concordance measure

$$\tilde{c}(s_1,s_2) = E\{\text{sign}(\tilde{T}_{11}-\tilde{T}_{12})(\tilde{T}_{21}-\tilde{T}_{22})|\tilde{T}_{11}\wedge\tilde{T}_{12}=s_1,\tilde{T}_{21}\wedge\tilde{T}_{22}=s_2\}$$

where $(\tilde{T}_{11},\tilde{T}_{21})$ and $(\tilde{T}_{12},\tilde{T}_{22})$ are independent observations from $F$. These local dependency measures give rise, respectively, to nonparametric dependency measures of ready interpretation over a follow-up region $(0,t_1]\times(0,t_2]$ as follows (Fan *et al.*, 2000): An average reciprocal cross ratio measure can be defined by

$$C(t_1,t_2) = \int_0^{t_1}\int_0^{t_2} c(s_1,s_2)^{-1}F(ds_1,ds_2)\Big/\int_0^{t_1}\int_0^{t_2} F(ds_1,ds_2),$$

while an average concordance measure is given by

$$\begin{aligned}\mathfrak{T}(t_1,t_2) &= E\{\text{sign}(\tilde{T}_{11}-\tilde{T}_{12})(\tilde{T}_{21}-\tilde{T}_{22})|\tilde{T}_{11}\wedge\tilde{T}_{12}\le t_1,\tilde{T}_{21}\wedge\tilde{T}_{22}\le t_2\}\\ &= \frac{\int_0^{t_1}\int_0^{t_2}F(s_1^-,s_2^-)F(ds_1,ds_2)-\int_0^{t_1}\int_0^{t_2}F(s_1^-,ds_2)F(ds_1,s_2^-)}{\int_0^{t_1}\int_0^{t_2}F(s_1^-,s_2^-)F(ds_1,ds_2)+\int_0^{t_1}\int_0^{t_2}F(s_1^-,ds_2)F(ds_1,s_2^-)}.\end{aligned}$$

Corresponding nonparametric estimators $\hat{C}(t_1,t_2)$ and $\hat{\mathfrak{T}}(t_1,t_2)$ arise by inserting a nonparametric strongly consistent estimator for $F$. Such estimators can be shown to be strongly consistent and to converge weakly to a Gaussian process, and bootstrap procedures are applicable for variance estimation.

The pairwise dependency estimators just described rely on a nonparametric estimator of the bivariate survivor function. Also the efficiency of the marginal regression parameter estimation may possibly be improved if an efficient nonparametric procedure were available to estimate marginal survivor and hazard functions. Such an estimator would need to exploit dependencies between the correlated failure times in order to make better use of censored observations. However, the problem of efficient nonparametric estimation of a bivariate survivor function has proven to be quite difficult, and a fully satisfactory estimation procedure has yet to be developed.

All proposed nonparametric estimators of $F$ place mass within the risk region, defined by points $(t_1,t_2)$ such that $\#\{T_1\ge t_1,T_2\ge t_2\}>0$, only on the grid formed by uncensored $T_1$ and $T_2$ values. Let $\hat{\Lambda}(t_1,t_2)=\hat{\Lambda}_{12}(t_1,t_2)=\hat{F}(\Delta t_1,\Delta t_2)/\hat{F}(t_1^-,t_2^-)$ denote a bivariate hazard rate estimator at $(t_1,t_2)$. Then given estimators $\hat{F}_1(t_1)=\hat{F}_1(t_1,0)$ and $\hat{F}_2(t_2)=\hat{F}_2(0,t_2)$, for example Kaplan-Meier estimators, of the marginal survivor functions one can recursively and uniquely generate a survivor function estimator using

$$\hat{F}(t_1,t_2)=\hat{F}(t_1^-,t_2)+\hat{F}(t_1,t_2^-)-\hat{F}(t_1^-,t_2^-)\{1-\hat{\Lambda}(\Delta t_1,\Delta t_2)\}.$$

The Bickel survivor function estimator (e.g., Dabrowska, 1988) uses a simple empirical hazard rate estimator

$$\hat{\Lambda}_E(\Delta t_1,\Delta t_2)=\#\{T_1=t_1,T_2=t_2,\delta_1=1,\delta_2=1\}/\#\{T_1\ge t_1,T_2\ge t_2\}.$$

Estimators of better efficiency assign mass at $(t_1,t_2)$ in a manner that acknowledges the amount of marginal mass remaining along $T_1 = t_1$ and $T_2 = t_2$ at or beyond $(t_1,t_2)$. Specifically, if one defines

$$\begin{aligned}\hat{L}_1(\Delta t_1,t_2^-) &= -\hat{F}(\Delta t_1,t_2^-)/\hat{F}(t_1^-,t_2^-)\\ \text{and}\ \hat{\Lambda}_1(\Delta t_1,t_2^-) &= \#\{T_1 = t_1, T_2 \geq t_2, \delta_1 = 1\}\Big/\#\{T_1 \geq t_1, T_2 \geq t_2\},\end{aligned}$$

with a corresponding specification for $\hat{L}_2$ and $\hat{\Lambda}_2$, then the Prentice-Cai (1992) hazard rate estimator can be written

$$\hat{\Lambda}_E(\Delta t_1,\Delta t_2)+\hat{L}_1(\Delta t_1,0)\{\hat{L}_2(t_1^-,\Delta t_2)-\hat{\Lambda}_2(t_1^-,\Delta t_2)\}+\hat{L}_2(0,\Delta t_2)\{\hat{L}_1(\Delta t_1,t_2^-)-\hat{\Lambda}_1(\Delta t_1,t_2^-)\}$$

and the Dabrowska (1988) hazard rate estimator is given by

$$\hat{L}_1(\Delta t_1,t_2^-)\hat{L}_2(t_1^-,\Delta t_2) + \frac{\{1-\hat{L}_1(\Delta t_1,t_2^-)\}\{1-\hat{L}_2(t_1^-,\Delta t_2)\}}{\{1-\hat{\Lambda}_1(\Delta t_1,t_2^-)\}\{1-\hat{\Lambda}_2(t_1^-,\Delta t_2)\}}$$
$$\{\hat{\Lambda}_E(\Delta t_1,\Delta t_2)-\hat{\Lambda}_1(\Delta t_1,t_2^-)\hat{\Lambda}_2(t_1^-,\Delta t_2)\}$$

These estimators tend to have excellent moderate sample performance although they are generally not nonparametric efficient due, at least in part, to their use of Kaplan-Meier estimates of marginal survivor function.

Nonparametric maximum likelihood estimation of $F$ suffers from serious uniqueness problems. Van der Laan (1996) provided a method for repairing the NPMLE over a region $(0,\tau_1) \times (0,\tau_2)$. His method begins by truncating the $T_1$ data at $\tau_1$ and the $T_2$ data at $\tau_2$. Fixed partitions of $(0,\tau_1]$ and $(0,\tau_2]$ are then defined and potential censoring times (assumed to be available) are replaced by potential censoring times at the immediately preceding partition point. Nonparametric maximum likelihood estimation then proceeds using the E-M algorithm by distributing singly censored observations in a manner that conditions on the partition strip in which they reside. Van der Laan develops the impressive result that nonparametric efficient estimation is possible if the partition bandwidths decrease to zero at a slow rate as sample size increases. Unfortunately the moderate sample performance of the repaired NPMLE is often found to be poorer than that of the Dabrowska and Prentice-Cai estimators in spite of the iterative calculation and the need to have potential censoring times available. Hence this survivor function estimation problem evidently needs further development.

## 6 Additional comments

Multivariate failure time methods have not yet achieved the state of development of corresponding univariate methods. However, flexible models and estimation procedures are available for the analysis of recurrent events. Methods based on frailty models (e.g., Hougaard, 2000) also have application to aspects of this problem, and frailties can provide an approach for relaxing an independent censorship assumption alternative to the analysis of partially marginalized rates discussed here (e.g., Wang *et al.*, 2001). Inverse censoring probability weighting potentially provides a means of retaining the desirable interpretation of the mean model (2) while avoiding an unduly strong independent censorship assumption. A simple version of this approach (e.g., Robins *et al.*, 1994) would estimate $\beta$ in (6) for a mean model (2) using an estimating function

$$U(\beta) = \sum_{i=1}^{n} \int_{0}^{\infty} \hat{\pi}_i(u)^{-1} \{Z_i(u) - \varepsilon(\beta, u)\} dN_i(u),$$

where $\hat{\pi}(u)$ is an estimate of $P\{C_i < u|X(u)\}$ and $X(u)$ is comprised of covariates that are external to the recurrent event process. Further analysis of the relative merits of this approach to the partially marginalized hazard rate modeling approach would be of interest.

Correlated failure time methods are available that are adequate for most practical purposes. The development of a convenient efficient nonparametric multivariate survivor function estimator could, however, unify such methods and strengthen them for a variety of purposes. In particular, methods for using data on auxiliary variables, including high dimensional variables that may arise in genomic and proteomic problems in molecular genetics could provide a valuable advance for the analysis of such heavily censored endpoints on disease occurrence and mortality in epidemiologic and disease prevention contexts.

## Acknowledgment

This work was partially supported by grant CA-53996 from the U.S. National Institutes of Health.

## References

Aalen, O.O. and Husebye, E. (1991). Statistical analysis of repeated events forming renewal processes. *Statistics in Medicine* 10, 1227-1240.

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*, New York: Springer-Verlag.

Andersen, P.K. and Gill, R.D. (1982). Cox's regression model for counting processes: a large sample study. *Annals of Statistics*, 10, 1100-1120.

Andrews, D.F. and Herzberg, A.M. (1985). *Data: A Collection of Problems from Many Fields for the Student and Research Worker*, New York: Springer-Verlag.

Byar, D.P. (1980). The Veteran's Administration study of chemoprophylaxis of recurrent stage I bladder tumors: comparisons of placebo, pyridoxine, and topical thiotepa. In: *Bladder Tumors and Other*

*Topics in Urological Oncology*. M. Pavone-Macaluso, P.H. Smith and F. Edsmyn, eds., pp.363-370, New York: Plenum.

Cai, J. and Prentice, R.L. (1995). Estimating equations for hazard ratio parameters based on correlated failure time data. *Biometrika*, 82, 151-164.

Cai, J. and Prentice, R.L. (1997). Regression estimation using multivariate failure time data and a common baseline hazard function model. *Lifetime Data Analysis*, 3, 197-213.

Cox, D.R. (1972). Regression models and life tables (with discussion). *Journal of the Royal Statistical Society, Series B*, 34, 187-220.

Cox, D.R. (1973). The statistical analysis of dependencies in point processes. In: *Symposium on Point Processes*. P.A.W. Lewis, ed., pp. 55-66, New York: Wiley.

Cox, D.R. and Isham, V. (1980). *Point Processes*, London: Chapman and Hall.

Cox, D.R. and Lewis, P.A. (1966). *The Statistical Analysis of a Series of Events*, London: Methuen.

Dabrowska, D.M. (1988). Kaplan-Meier estimate on the plane. *Annals of Statistics*, 16, 1475-1489.

Fan, J., Hsu, L. and Prentice, R.L. (2000). Dependence estimation over a finite bivariate failure time region. *Lifetime Data Analysis*, 6, 343-355.

Gail, M.H., Santner, T.J. and Brown, C.C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics*, 36, 255-266.

Hougaard, P. (2000). *Analysis of Multivariate Survival Data*. Springer-Verlag, New York.

Kalbfleisch, J.D. and Prentice, R.L. (2002). *The Statistical Analysis of Failure Time Data, Second Edition*. New York: Wiley.

Lawless, J.F. (1987). Regression methods for Poisson process data. *Journal of the American Statistical Association*, 82, 808-815.

Lawless, J.F. and Nadeau, C. (1995). Some simple and robust methods for the analysis of recurrent events. *Technometrics*, 37, 158-168.

Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J.P. Klein and P.K. Goel (eds.), Kluwer Academic Publishers, 237-247.

Lin, D.Y., Sun, W. and Ying, Z. (1999). Nonparametric estimation of the gap time distribution for serial events with censored data. *Biometrika*, 86, 59-70.

Lin, J.S. and Wei, L.J. (1992). Linear regression for multivariate failure time observations. *Journal of the American Statistical Association*, 87, 1091-1097.

Lin, D.Y., Wei, L.J., Yang, I. and Ying, Z. (2000). Semiparametric regression for the mean and rate functions of recurrent events. *Journal of the Royal Statistical Society, B*, 62, 711-730.

Lin, D.Y., Wei, L.J. and Ying, Z. (1998). Accelerated failure time models for counting processes. *Biometrika*, 85, 605-618.

Nelson, W.B. (1988). Graphical analysis of system repair data. *Journal of Quality Technology*, 20, 24-35.

Nelson, W.B. (1995). Confidence limits for recurrence data-applied to cost or number of product repairs. *Technometrics*, 37, 147-157.

Oakes, D. (1989). Bivariate survival models induced by frailties. *Journal of the American Statistical Association*, 84, 487-493.

Pepe, M.S. and Cai, J. (1993). Some graphical displays and marginal regression analyses for recurrent failure times and time dependent covariates. *Journal of the American Statistical Association*, 88, 811-820.

Prentice, R.L. and Cai, J. (1992). Covariance and survivor function estimation using censored multivariate failure time data. *Biometrika*, 79, 495-512.

Prentice, R.L., Williams, B.J. and Peterson, A.V. (1981). On the regression analysis of multivariate time data. *Biometrika*, 68, 373-379.

Robins, J.M., Rotnitsky, A. and Zhao, L.P. (1994). Estimation of regression coefficients when some regressors are not always observed. *Journal of the American Statistical Association*, 89, 846-866.

Snyder, D.L. (1975). *Random Point Processes*, New York: Wiley.

Van der Laan, M.J. (1996). Efficient estimation in the bivariate censoring model and repairing NPMLE. *Annals of Statistics*, 24, 596-627.

Wang, M.-C. and Chang, S.-H. (1999). Nonparametric estimation of a recurrent survival function. *Journal of the American Statistical Association*, 94, 146-153.

Wang, M.-C., Qin, J. and Chiang, C.-T. (2001). Analyzing recurrent event data with informative censoring. *Journal of the American Statistical Association*, 96, 1057-1065.

Wei, L.J., Lin, D.Y. and Weissfeld, L. (1989). Regression analysis of multivariate incomplete failure time data by modeling marginal distributions. *Journal of the American Statistical Association*, 84, 1065-1073.

## Resum

Les dades multivariants de temps de supervivència sorgeixen en situacions diverses. Entre d'altres inclouen a) dades d'esdeveniments recurrents: obtingudes quan s'observa la seqüència d'ocurrències d'un cert tipus d'esdeveniment; b) temps de fallades correlacionats: quan s'estudia l'ocurrència de dos o més tipus d'esdeveniments per individus que estan simultàniament a risc; c) dades obtingudes d'individus diferents que tenen temps fins a un esdeveniment depenents; d) processos multi-estat més complicats en els quals els individus es mouen entre un número discret d'estats, durant el transcurs d'un estudi de seguiment, i en els quals es registren els diferents estats així com el temps transcorregut en ells. En aquest article presentem una revisió crítica dels models i dels mètodes estadístics per a l'anàlisi de dades d'esdeveniments recurrents i de temps de fallada correlacionats. Aquesta revisió indica el rol important que els models d'intensitats parcialment marginalitzats poden jugar en les anàlisis de dades recurrents i remarca la utilitat dels models de funcions de risc marginals i dels estimadors noparamètrics de les dependències dos a dos per les anàlisis de dades correlacionades. S'indiquen àrees on és necessari més desenvolupament metodològic.

*MSC:* 62N01, 62N02, 62N03,62H10, 92B15

*Paraules clau:* Censurament independent; esdeveniments recurrents; estimació de la funció supervivència; temps de fallada correlacionats

SORT 27 (1) January-June 2003, 79-94

# Indirect inference for survival data

Bruce W. Turnbull* and Wenxin Jiang

*Cornell University*

## Abstract

In this paper we describe the so-called "indirect" method of inference, originally developed from the econometric literature, and apply it to survival analyses of two data sets with repeated events. This method is often more convenient computationally than maximum likelihood estimation when handling such model complexities as random effects and measurement error, for example; and it can also serve as a basis for robust inference with less stringent assumptions on the data generating mechanism. The first data set concerns recurrence times of mammary tumors in rats and is modeled using a Poisson process model with covariates and frailties. The second data set involves times of recurrences of skin tumors in individual patients in a clinical trial. The methodology is applied in both parametric and semi-parametric regression analyses to accommodate random effects and covariate measurement error.

*MSC:* 62N01, 62N02, 62P10

*Keywords:* Estimating equations, frailty, hazard rate regression, indirect inference, measurement error, naive estimators, overdispersion, quasi-likelihood, random effects, robustness

## 1 Introduction

Methods of *indirect inference* (Gourieroux, Monfort and Renault, 1993) have been developed and used in the field of econometrics where they have proved valuable for parameter estimation in highly complex models. This paper recasts the basic technique in a likelihood-flavoured approach and illustrates some applications in biostatistics, in particular for survival and repeated events data.

We begin by illustrating the steps involved in the indirect method in the following simple pedagogic example.

* *Address for correspondence:* B. W. Turnbull. School of Operations Research and Department of Statistical Science, 227 Rhodes Hall, Cornell University, Ithaca, NY 14853-3801, USA. turnbull@orie.cornell.edu

Received: October 2002
Accepted: January 2003

**Example 1:** *exponential survival with censoring.* Consider lifetimes $\{T_1,\ldots,T_n\}$, which are independent and identically distributed (i.i.d.) according to an exponential distribution with mean $\theta$. The data are subject to Type I single censoring after fixed time $c$. Thus the observed data are $\{Y_1,\ldots,Y_n\}$, where $Y_i = \min(T_i, c)$, $(i = 1,\ldots,n)$. We consider indirect inference based on the intermediate statistic $\hat{s} = \overline{Y}$. This choice can be considered either as the basis for a method of moments estimator or as the MLE (maximum likelihood estimator) for a misspecified model M$'$ in which the presence of censoring has been ignored. The naive estimator $\overline{Y}$ in fact consistently estimates not $\theta$ but the "naive" or "auxiliary" parameter

$$s(\theta) = \theta\,[1-\exp(-c/\theta)], \tag{1}$$

the expectation of $\overline{Y}$. The equation (1) is an example of what may be termed as a "bridge relation" (Jiang and Turnbull, 2001) or a "binding relation" (Gourieroux, *et al.*, 1993). We can see the obvious effect of the misspecification, namely that $\hat{s}$ underestimates $\theta$. However a consistent estimator $\hat{\theta}$ of $\theta$ as $n \longrightarrow \infty$ can be obtained by solving (1) for $\theta$ with $s(\theta)$ replaced by $\hat{s} = \overline{Y}$. That is, $\hat{\theta} = s^{-1}(\hat{s})$. (Note that $s(\cdot)$ is strictly increasing on $\Re^+$ and thus invertible). We note also that $\hat{\theta}$ is not the MLE of $\theta$ which is $n\overline{Y}/[\sum_{i=1}^n I(Y_i < c)]$.)

More generally, a consistent estimator can be constructed based on an intermediate statistic $\hat{s}$ that does not need to have the interpretation of a 'naive' estimator. For example, above we could have chosen perhaps $\hat{s} = \overline{Y^2} = n^{-1}\sum_{i=1}^n Y_i^2$ so that $\hat{\theta} = s^{-1}(\hat{s})$ where $s^{-1}$ is the inverse function of $s(\theta) \equiv E(\overline{Y^2}|\theta)$. In fact,the dimension of $\hat{s}$ can be greater than that of $\theta$ — e.g., we could take $\hat{s} = (\overline{Y}, \overline{Y^2})^T$ in the above example. Now a consistent estimator $\hat{\theta}$ of $\theta$ can be found by using weighted least squares,

$$\hat{\theta} = \arg\min_{\theta}\{\hat{s} - s(\theta)\}^T A\{\hat{s} - s(\theta)\},$$

where an optimal choice of $A$ is the inverse of the estimated variance matrix of $\hat{s}$, as we will discuss later. This is the principal idea of indirect inference– statistical inference of $\theta$ based on an indirect data "summary" $\hat{s}$. The choice of $\hat{s}$ is not unique, but in most applications there will natural one to use as we shall see.

We will term $\hat{\theta}$ as the "indirect MLE", since it can be viewed as the MLE using an approximate likelihood based on the indirect data summary $\hat{s}$. We will also see how to obtain the standard error for $\hat{\theta}$.

## 2 Indirect inference

In general, the indirect MLE has properties similar to those of the usual MLE: consistency, asymptotic normality, and certain efficiency properties. In addition, chi-squared goodness-of-fit tests can be based on the indirect likelihood.

Advantages of the indirect method include ease of computation; robustness; and informativeness on the effect of model misspecification. We will summarize the framework of indirect inference below.

### 2.1 The basic approach

Suppose we have a data set consisting of $n$ independent units. The essential ingredients of the indirect approach, when reformulated in a likelihood-flavoured treatment, are as follows.

- There is a hypothesized true model M for data generation, with distribution $P^{(\theta)}$ which depends on an unknown parameter $\theta$ of interest which is of dimension $p$.
- One first computes an *intermediate* or *auxiliary* statistic $\hat{s}$ of dimension $q \geq p$, which is asymptotically normal with mean $s(\theta)$, say, under model M.
- An *indirect likelihood* $L(\theta|\hat{s})$ is then constructed based on the normal approximation, so that, apart from an additive constant,

$$-2\log L(\theta|\hat{s}) = \{\hat{s} - s(\theta)\}^T v^{-1}\{\hat{s} - s(\theta)\} \; = \; H(\theta), \text{ say,} \qquad (2)$$

  where $v$ is a consistent estimate of the asymptotic variance $\widehat{var}(\hat{s})$. A typical choice might be the 'robust' or 'sandwich formula', when $\hat{s}$ solves an estimating equation (see e.g. Carroll, Ruppert and Stefanski 1995, Section A.3).
- This indirect likelihood is then maximized to generate an indirect maximum likelihood estimate (*indirect MLE*) or *adjusted* estimate $\hat{\theta}(\hat{s})$ for $\theta$. In the case when the dimension ($q$) of the intermediate statistic equals that ($p$) of the parameter $\theta$ and $s(\theta)$ is invertible, it can be seen from (2) that maximization of the indirect likelihood is equivalent to solving the "bridge" or "binding" equation $s(\theta) = \hat{s}$ for $\theta$, because then (2) can be made zero.

In the "indirect" analysis of the pedagogic example of Section 1, M is the i.i.d. exponential model with censoring: $Y_i = \min(T_i, c)$ and $P^{(\theta)}(T_i \leq t) = 1 - e^{-t/\theta}$, for $t \in [0, \infty)$, $i = 1, \ldots, n$. In the initial approach, the intermediate statistic was $\hat{s} = n^{-1}\sum_{i=1}^{n} Y_i$, which is asymptotically normal as $n \longrightarrow \infty$ by the central limit theorem, with asymptotic mean $s(\theta) = \theta\,[1 - \exp(-c/\theta)]$. The indirect likelihood $L(\theta|\hat{s})$ is given by

$$-2\log L(\theta|\hat{s}) = \{n^{-1}\sum_{i=1}^{n} Y_i - s(\theta)\}^T v^{-1}\{n^{-1}\sum_{i=1}^{n} Y_i - s(\theta)\}$$

where one can substitute the robust estimate $\hat{v} = \{n(n-1)\}^{-1}\sum_{i=1}^{n}(Y_i - \bar{Y})^2$ for the asymptotic variance $v = var(\hat{s})$. Finally the adjusted estimate (or indirect MLE) is $\hat{\theta} = s^{-1}(\hat{s})$.

In summary, in this indirect approach, the data are first summarized by the intermediate statistic $\hat{s}$. Its asymptotic mean $s$ is referred to as the *auxiliary parameter*. The auxiliary parameter is related to the original parameter by a relation $s = s(\theta)$, termed the *bridge relation* or *binding function*.

The starting point is the choice of an intermediate statistic $\hat{s}$. This can be chosen as some set of sample moments, or the solution of some estimating equations, or the MLE based on some convenient model M$'$, say, termed the *auxiliary* (or *naive*) model. If the last, then the model M$'$ is a simpler but misspecified or partially misspecified model. As stated previously, the choice of an intermediate statistic $\hat{s}$ is not necessarily unique; however in any given situation there is often a natural one to use.

### 2.2 Intermediate statistics arising from estimating equations

Most intermediate statistics can be defined implicitly as a solution, $s = \hat{s}$, of a ($q$-dimensional) estimating equation of the form $G(\mathbf{W}, s) = 0$, say. [Clearly this includes any statistic $\hat{s} = \hat{s}(\mathbf{W})$ that has an explicit expression as a special case, by taking $G = s - \hat{s}(\mathbf{W})$.] The estimating equation could be the normal equation from a least-squares analysis, or the score equation based on some likelihood function.

In such situations there is a parallel formulation of indirect inference in 'implicit form'. For instance, one can state the 'bridge relation' $s(\theta)$ implicitly as $F(\theta, s) = 0$ where $F(\theta, s) \equiv E_{\mathbf{W}|\theta} G(\mathbf{W}, s)$, which is the limiting version of the estimating equation $G(\mathbf{W}, \hat{s}) = 0$. Correspondingly, in the definition of the indirect likelihood $L$, $H$ can be (asymptotically) equivalently defined by $H(\theta, \hat{s}) = F(\theta, \hat{s})^T v^{-1} F(\theta, \hat{s})$. Here $v$ is (a sample estimate of) the avar of $F(\theta, \hat{s})$, which can be evaluated by the delta method (e.g. Bickel and Doksum (2001), Sec. 5.3.2), and found to be the same as $var(G)$ evaluated at $s = s(\theta)$ (the auxiliary parameter). Then we define the *adjusted estimator* (or the *indirect MLE*) $\hat{\theta}$ to be the maximizer of $L$, or the minimizer of $H$.

### 2.3 Properties of indirect MLE

In general, the indirect MLE has a set of properties analogous to those of the usual MLE. These include, under appropriate regularity conditions:

(i) (Indirect Score Function). The asymptotic mean and variance of the indirect likelihood score function satisfy the usual relations $E(\nabla_\theta \log L) = 0$ and $var(\nabla_\theta \log L) + E(\nabla_\theta^2 \log L) = 0$.

(ii) (Asymptotic Normality). The adjusted estimator $\hat{\theta}$ is asymptotically normal (AN) with mean $\theta$, and with asymptotic variance (avar) estimated by $-(\nabla_\theta^2 \log L)^{-1}$ or $2(\nabla_\theta^2 H)^{-1}$ where consistent estimates are substituted for parameter values.

(iii) (Tests). Likelihood-ratio statistics based on the indirect likelihood for testing simple and composite null hypotheses have the usual asymptotic $\chi^2$ distributions.

(iv) (Efficient use of indirect data). The adjusted estimator has smallest avar among all consistent AN estimators $f(\hat{s})$ of $\theta$, which are constructed from the naive estimator $\hat{s}$ by continuously differentiable mappings $f$.

These results can be found in the references cited in Section 2.5, and are summarized in Jiang and Turnbull (2001, Proposition 1).

When different intermediate statistics are used, the asymptotic efficiency can be different. In general the indirect MLE is not as efficient as the MLE based on the true model M; although there are situations that can be identified where the efficiency will be high, as in the example of Section 3.1.

In a special case when the dimension of the intermediate statistic ($q$) equals that ($p$) of the parameter $\theta$, and $s(\cdot)$ is a diffeomorphism on the parameter space $\Theta$ of $\theta$, maximization of $L$ is equivalent to the bias correction $\hat{\theta} = s^{-1}(\hat{s})$ (from solving $F(\theta, \hat{s}) = 0$), which is AN and consistent for $\theta$. See, e.g., Kuk (1995), Turnbull *et al.* (1997) and Jiang *et al.* (1999) for biostatistical applications.

When $q < p$, there are more unknown true parameters than 'naive parameters'. In this case the bridge relation is many-to-one and does not in general permit the construction of adjusted estimates. It is mainly of interest for investigating the effects of misspecification when the naive estimators are constructed under misspecified models. However, in such situations it may be possible to construct consistent estimates for a subset of true parameters, which may be of interest. In other situations, some components of the higher-dimensional true parameter are known or can be estimated from other outside data sources. This enables the other components to be consistently estimated by inverting the bridge relation. Examples of this kind arising from errors-in-variables regression models are given in Sections 3.2 and 3.3.

### 2.4 *Why consider the indirect method?*

This indirect approach offers the following advantages:

1. *Ease of computation.* The indirect method is typically computationally simpler and more convenient. For example, when $\hat{s}$ is based on some simplified model M$'$, it can often be computed with available standard computer software.
2. *Informativeness on the effect of model misspecification.* When $\hat{s}$ is a 'naive estimate' obtained from a naive model M$'$ neglecting certain model complexities, the approach is very informative on the effect of model misspecification — the bridge relation $s = s(\theta)$ provides a dynamic correspondence between M$'$ and M. For example, in errors-in-variable regression, such a relation is sometimes termed an

'attenuation relation' (see e.g., Carroll, Ruppert and Stefanski 1995, Chapter 2), and tells how regression coefficients can be underestimated when neglecting the measurement error in a predictor.

3. *Robustness.* The validity of the inference based on an intermediate statistic essentially relies on the correct specification of its asymptotic mean. This is typically a less demanding assumption than the correct specification of a full probability model, which would be generally needed for a direct likelihood inference to be valid. Therefore inferences based on the adjusted estimate $\hat{\theta}$ can remain valid despite some departure of the data generation mechanism from the hypothesized true model M.

### 2.5 Bibliography and notes

The above very brief exposition of the indirect method of inference represents a summary of results that have appeared in the econometric and statistical literature in varying forms and generality and tailored for various applications. Examples include: the generalized method of moments (GMM: Hansen 1982); the method of linear forms and minimum $\chi^2$ (Ferguson 1958); the regular best asymptotic normal estimates that are functions of sample averages (Chiang 1956, Theorem 3); simulated method of moments and indirect inference [McFadden (1989), Pakes and Pollard (1989), Gourieroux *et al.* (1993), Gallant and Tauchen (1996, 1999) Gallant and Long (1997)]. Newey and McFadden (1994, Chapters 6 and 8) discuss two-stage parametric and nonparametric estimation in the GMM context, where some 'nuisance' parameter, possibly infinite dimensional, is estimated from a preliminary consistent method.

Applications of GMM in the settings of generalized estimating equations from biostatistics are discussed in Qu, Lindsay and Li (2000). McCullagh and Nelder (1989, p. 341), as referred to by Qin and Lawless (1994, p. 315), consider optimal linear combination of estimating equations, as is traditionally done in GMM literature. Qin and Lawless (1994) also provide an alternative but asymptotically equivalent way of combining estimating equations using empirical likelihood.

The theory of estimators obtained from misspecified likelihoods goes back at least as far as Cox (1962), Berk (1966) and Huber (1967) and is summarized in the comprehensive monograph by White (1994). The use of $\hat{s}$ (based on an auxiliary model M$'$) in indirect inference about $\theta$ (under model M) appears recently in the field of econometrics to treat complex time series and dynamic models, see, e.g., Gourieroux *et al.* (1993) and Gallant and Tauchen (1996, 1999); as well as in the field of biostatistics to treat regression models with random effects and measurement error, see e.g., Kuk (1995), Turnbull, Jiang and Clark (1997), and Jiang *et al.* (1999). This bibliography is far from exhaustive. A thorough review and synthesis of the methods of indirect inference are given in Jiang and Turnbull (2001).

## 3 Three applications with survival data

In this section we discuss three applications with recurrent event data which use models of increasing order of complexity:

3.1 A Poisson process regression model with random effects ("frailties" or "unexplained heterogeneity");
3.2 A Poisson process regression model with random effects and covariate measurement error;
3.3 A semi-parametric intensity rate regression model with random effects and measurement error.

The first example uses mammary tumor recurrence times from a rodent carcinogenicity experiment. The remaining two examples use data on skin cancer recurrences in the Nutritional Prevention of Cancer (NPC) trial — a long-term randomized clinical trial for cancer prevention (Clark *et al.* 1996).

### *3.1 Animal carcinogenicity data: multiple times to tumor*

Gail, Santner and Brown (1980, Table 1) present data on multiple mammary tumor incidence times from an experiment conducted by Thompson *et al.* (1978). Forty-eight female rats which remained tumor-free after sixty days of pre-treatment of a prevention drug (retinyl acetate) were randomized with equal probability into two groups. In Group 1 they continued to receive treatment ($Z = 1$), in Group 2 they received placebo ($Z = 0$). All rats were followed for an additional 122 days and the time of any newly diagnosed mammary tumor was recorded. The numbers of tumors diagnosed in individual rats ranged from 0 to 13. The objective of the study was to estimate the effect of the preventive treatment (Z) on tumor recurrence.

Suppose we consider a model in which the tumors occur over time in a given subject (rat) according to a Poisson process with a constant intensity rate which depends on treatment Z, a fixed effect, and on subject, a random effect. If we define $Y$ to be the number of tumors diagnosed in a particular rat during the 122 day followup time, the model M specifies that, given $Z$ and $\epsilon$, $Y$ is Poisson distributed with mean $\exp(\alpha + Z\beta + \epsilon)$. Here the assigned treatment $Z$ is observed, but $\epsilon$ represents an unobserved random effect modeled as normally distributed with zero mean and constant variance $\sigma^2$, independent of Z. This random effect or "unexplained heterogeneity" could be considered to be caused by omitted covariates. We observe $n = 48$ i.i.d. pairs $W_i = (Y_i, Z_i), i = 1, \dots, n$. The likelihood for the observed data involves integration over $\epsilon$ and is difficult to compute. (However it is possible - see below.) Instead, we start by taking the indirect approach with an auxiliary statistic $\hat{s} = (\hat{a}, \hat{b}, \hat{t}^2)^T$, where

$(\hat{a},\hat{b})$ are the regression coefficient estimates maximizing a naive log-likelihood $R = \sum_1^n\{Y_i(a+Z_ib) - e^{a+Z_ib}\}$, and $\hat{t}^2 = n^{-1}\sum_{i=1}^n Y_i^2$ is the second sample moment. Here the auxiliary parameter is $s = \text{plim}(\hat{a},\hat{b},\hat{t}^2)^T$, whereas the true parameter to be estimated is $\theta = (\alpha,\beta,\sigma^2)^T$. The use of the naive log-likelihood $R$ corresponds to a simplified model M$'$ in which the presence of the random effect $\epsilon$ is neglected. The second sample moment is included in the intermediate statistic to provide information for estimation of the variance parameter. Therefore $\hat{s}$ is solved from the estimating equation $G(\mathbf{W},s) = 0$, where (formally) $G = (n^{-1}\partial_aR, n^{-1}\partial_bR, \hat{t}^2 - t^2)^T$, i.e.

$$G = n^{-1}\sum_{i=1}^n (Y_i - e^{a+Z_ib},\ Z_i(Y_i - e^{a+Z_ib}),\ Y_i^2 - t^2)^T \quad = n^{-1}\sum_{i=1}^n g_i, \text{ say.}$$

The solution $\hat{s} = (\hat{a},\hat{b},\hat{t}^2)^T$ can be computed easily. For the rat carcinogenicity data we obtain the auxiliary estimates $\hat{a} = 1.7984$; $\hat{b} = -0.8230$; $\hat{t}^2 = 31.875$. The asymptotic variance $var(\hat{s})$ can be estimated by the sandwich formula (see e.g. Carroll, Ruppert and Stefanski 1995, Section A.3)

$$v = (\nabla_sG)^{-1}\widehat{var}(G)(\nabla_sG)^{-T}|_{s=\hat{s}}$$

where $\widehat{var}(G) = n^{-2}\sum_{i=1}^n g_ig_i^T|_{s=\hat{s}}$, $\nabla_sG$ is a $3\times 3$ matrix with elements $(\nabla_sG)_{jk} = \partial_{s_k}G_j$, $j,k = 1,2,3$, and $A^{-T} = (A^{-1})^T$ for a generic matrix $A$.

The indirect likelihood $L(\theta|\hat{s})$, up to an additive constant, satisfies

$$-2\log L(\theta|\hat{s}) = \{\hat{s} - s(\theta)\}^T v^{-1}\{\hat{s} - s(\theta)\},$$

where $s(\theta)$ is the asymptotic mean or large sample almost sure limit of $\hat{s}$. Since $\hat{s}$ solves the estimating equation $G = 0$, its limit is the solution of the limiting estimating equation $F(\theta,s) = E_{\mathbf{W}|\theta}G(\mathbf{W},s) = 0$, which can be explicitly solved to obtain $s = s(\theta)$. This yields the bridge equation:

$$\begin{aligned} s &= \text{plim}(\hat{a},\hat{b},\hat{t}^2)^T \\ &= s(\theta) = \left(\alpha + \sigma^2/2,\ \beta,\ \frac{1}{2}(1+e^{\beta})e^{\alpha+\frac{1}{2}\sigma^2} + \frac{1}{2}(1+e^{2\beta})e^{2(\alpha+\sigma^2)}\right)^T. \end{aligned}$$

Because $\dim(s) = \dim(\theta) = 3$ and $s(\theta)$ is a smooth invertible mapping, the indirect MLE $\hat{\theta} = \arg\max_\theta L(\theta|\hat{s})$ can be obtained by solving $\hat{s} = s(\theta)$, which gives the adjusted estimates $\hat{\theta} = (\hat{\alpha},\hat{\beta},\hat{\sigma}^2) = s^{-1}(\hat{s})$. Thus $\hat{\beta} = \hat{b}$, and $\hat{\alpha} = \hat{a} - \hat{\sigma}^2/2$ where $\hat{\sigma}^2 = \log\left\{\frac{2\hat{t}^2 - e^{\hat{a}}(1+e^{\hat{b}})}{e^{2\hat{a}}(1+e^{2\hat{b}})}\right\}$. For the rat data, this leads to adjusted estimates $\hat{\alpha} = 1.6808(0.1589)$; $\hat{\beta} = -0.8230(0.1968)$; $\hat{\sigma} = 0.4850(0.1274)$.

The estimated standard errors shown in parentheses are obtained using the delta method formula: $\widehat{var}(\hat{\theta}) = (\nabla_\theta s)^{-1}v(\nabla_\theta s)^{-T}|_{\theta=\hat{\theta}}$, and then taking the square roots of the 3 diagonal elements of this matrix. It is noted that this delta method expression is equivalent to deriving the variance by

$$\{-\nabla_\theta^2 \log L(\theta|\hat{s})\}^{-1}|_{\theta=\hat{\theta}}$$

based on the 'indirect likelihood Fisher information', where $\nabla_\theta^2$ represents the Hessian. This follows because the $jk$th element of the Hessian is, for $j,k = 1,2,3$,

$$\begin{aligned}\{-\nabla_\theta^2 \log L(\theta|\hat{s})\}_{jk}|_{\theta=\hat{\theta}} &= -\partial_{\theta_j}\partial_{\theta_k} \log L(\theta|\hat{s})|_{\theta=\hat{\theta}} \\ &= (\partial_{\theta_j} s^T)v^{-1}(\partial_{\theta_k} s)|_{\theta=\hat{\theta}} - (\partial_{\theta_j}\partial_{\theta_k} s^T)v^{-1}(\hat{s}-s)|_{\theta=\hat{\theta}} \\ &= (\partial_{\theta_j} s^T)v^{-1}(\partial_{\theta_k} s)|_{\theta=\hat{\theta}} + 0 \\ &= \{\widehat{var}(\hat{\theta})^{-1}\}_{jk}.\end{aligned}$$

If we wish to obtain the MLE of $\theta = (\alpha,\beta,\sigma^2)$ based on model M, then it can be found by a somewhat tedious iterative numerical maximization of the true likelihood which involves numerical integration over the distribution of $\epsilon$. These estimates are: $\hat{\alpha}_{ML} = 1.6717$ (0.1560); $\hat{\beta}_{ML} = -0.8125$ (0.2078); $\hat{\sigma}_{ML} = 0.5034$ (0.0859). For the MLEs, the estimated standard errors are based on the inverse of the Fisher information matrix, evaluated at the corresponding estimate values.

The estimated standard errors suggest that the efficiency of indirect estimation of the treatment effect parameter $\beta$ is high here in this example. Related results (Cox, 1983; Jiang *et al.*, 1999) show that such high efficiency is achievable if the follow-up times are about the same across different subjects (which is true here), or if the overdispersion is small. Also it should be noted that the adjusted estimator $\hat{\beta}$ is robust, in the sense that it remains consistent, essentially as long as the mean function $E(Y|Z,\epsilon)$ is correctly specified and $\epsilon$ and $Z$ are independent. (Its standard error estimate from the sandwich formula is also model-independent and robust.) In particular, the consistency property does not depend on the specification of a complete probability model, namely that $Y$ is Poisson and $\epsilon$ is normal. Thus the indirect estimator enjoys a robustness advantage over the MLE.

The indirect approach, although formulated from the different perspective of using naive model plus method of moments, is intimately related to the work of Breslow (1990) based on quasi-likelihood and method of moments. Breslow used a different linear combination of $Y_i$'s based on quasi-likelihood (Wedderburn, 1974; McCullagh and Nelder, 1989), which enjoy general efficiency properties among linear estimating equations. However, (i) our approach can be interpreted as basing inference on the simple moments $n^{-1}\sum Y_i$, $n^{-1}\sum Z_i Y_i$ and $n^{-1}\sum Y_i^2$ (which can be easily seen from the estimating equation $G = 0$), and (ii) our approach shows clearly, by the use of bridge relations, the sensitivity and robustness of parameter estimates to the omission of over-dispersion in modeling. Also note that here we used a log-normal distribution to model the random effects and the variance parameter also enters the mean model (unconditional on $\epsilon$), whereas Breslow (1990) focused on the examples such as ones with gamma multiplicative random effects in which the mean model does not change.

For the only comparable parameter $\beta$ (the treatment effect), the Breslow method (from his equations (1), (2) and (7)) gives exactly the same answer as our adjusted analysis: $\hat{\beta}_{\text{Breslow}} = -0.8230(0.1968)$. This is because, for this special two-group design, both methods essentially use the log(frequency ratio) to estimate the treatment effect.

### 3.2 Skin cancer recurrence data from the NPC trial: parametric modeling

Clark *et al.* (1996) have described the results of the "Nutritional Prevention of Cancer" (NPC) trial. This trial, begun in 1983, studied the long-term safety and efficacy of a daily 200$\mu$g nutritional supplement of selenium (Se) for the prevention of cancer. It was a double-blind, placebo-controlled randomized clinical trial with $n = 1312$ patients accrued and followed for up to about ten years. Here we shall consider a particular primary endpoint — namely squamous cell carcinoma (SCC) of the skin. The results for this endpoint are of particular interest because Clark *et al.* (1996) found a negative (but not statistically significant, P = 0.15) effect of selenium (Se) supplementation. This was opposite to previous expectations, and contrasted sharply with findings of highly significant positive benefits of the selenium supplementation in preventing a number of other types of cancers. However in their analysis, Clark *et al.* used only data on the time to *first* occurrence of SCC in each subject and employed a Cox model that ignored patient heterogeneity (i.e. that assumed a common baseline hazard) and ignored that some explanatory covariates were measured with error.

We consider the recurrences of SCC over time, measured from date of randomization, for patients $i = 1,...,n$ as $n$ i.i.d. discrete point processes $\{Y_i(t)\}$. Here $Y_i(t)$ is the observed number of recurrences for patient $i$ on day $t$ (usually zero or one). Time $t$ is measured in days on a discrete time scale $t = 1,\ldots,K$, where $K = 4618$ days, the longest followup time. The indicator variable $H_i(t)$ is one if patient $i$ is still on study ("at risk") on day $t$ and zero otherwise. For illustration purposes, we will consider only two explanatory variables, namely treatment assignment indicator $a$ and baseline Se level $x$. The latter is an important predictor, measured prior to randomization in each patient, but is contaminated with measurement error so that the observed value is recorded as $z$ not $x$. In the parametric approach, we postulate an independent Poisson process model with constant baseline mean event rate as the underlying data generating mechanism: for $i = 1,2,\ldots,n; t = 1,...,K$, $Y_i(t)$ are independent Poisson random variables with mean $E[Y_i(t)] = H_i(t)\psi_i\lambda\exp(a_i\gamma + x_i\beta)$.

Here the $\{\psi_i\}$ represent subject-specific random effects or "frailties", which modulate the constant baseline mean rate $\lambda$. In this framework, the sufficient statistics is $Y_i \equiv \sum_{t=1}^{K} H_i(t)Y_i(t)$ which follows Poisson distribution with mean $\tau_i\psi_i\lambda\exp(a_i\gamma + x_i\beta)$, with $\tau_i = \sum_{t=1}^{K} H_i(t)$ being the length of follow-up for patient $i$.

When a conjugate distribution Gamma(mean 1, variance $\nu$) for $\psi_i$ is used, the integration over the unobservable $\psi_i$ can be carried out analytically, so that unconditional on $\psi_i$:

> $Y_i$ follows a negative binomial distribution with mean $\mu_i$ and variance $\mu_i + \nu\mu_i^2$,

where $\mu_i = \tau_i\lambda\exp(a_i\gamma + x_i\beta)$. We refer to this as our base model "M(para)".

In Turnbull *et al.* (1997), the intended $x_i$ is the long-term average of the baseline Se level (in log-scale), which is subject to measurement error and temporal fluctuation. An error-contaminated version $z_i = x_i + u_i$ is observed, where $x_i$ and $u_i$ are assumed to be independent normal with zero means (after centering) and respective variances $\sigma_x^2$ and $\sigma_u^2$. A naive analysis ignoring measurement error would involve a negative binomial regression of $Y_i$ on $(a_i, z_i)$, instead of on $(a_i, x_i)$. The auxiliary model is then:

> M′(para): $Y_i$ is negative binomial with mean $q_i$ and variance $q_i + \nu q_i^2$,

where $q_i = \tau_i m\exp(a_i g + x_i b)$ and $s = (g, b, m, \nu)$ is the naive / auxiliary parameter corresponding to the parameter $\theta = (\gamma, \beta, \lambda, \nu)$ used in M(para).

***Table 1**: Statistical analyses for several models of NPC trial SCC data.*

| Model | Treatment | | Baseline Se | |
|---|---|---|---|---|
| | estimate | (s.e.) | estimate | (s.e.) |
| 1) Parametric: Constant Intensity | | | | |
| a) Naive (Model M′ (para)) | $\hat{g}$=0.122 | (0.059) | $\hat{b}$=−0.725 | (0.145) |
| b) Adjusted (Model M (para)) | $\hat{\gamma}$=0.122 | (0.125) | $\hat{\beta}$=−2.181 | (0.963) |
| 2) Semi-parametric | | | | |
| a) Naive (Model M′ (semi-par)) | $\hat{g}$=0.117 | (0.059) | $\hat{b}$=−0.690 | (0.146) |
| b) Adjusted (Model M (semi-par)) | $\hat{\gamma}$=0.117 | (0.125) | $\hat{\beta}$=−2.076 | (0.963) |

Such a naive analysis based on M′(para) was run and the resulting estimator $\hat{s}$ for $s$ forms our intermediate statistic. Computer packages for negative binomial regression can be used for this task, e.g. the procedure *nbreg* in STATA 5.0 (StataCorp 1997). The bridge relation $s(\theta)$ as a consistent limit of $\hat{s}$ when the true parameter is $\theta$ was shown (Turnbull *et al.*, 1997) to include an implicit equation for solving for $\nu$, as well as the following explicit formulae:

$$g = \gamma, \quad b = \pi\beta \quad \text{and} \quad m = \lambda\exp(0.5\beta^2\sigma^2_{x|z}),$$

where $\pi = \sigma_x^2/(\sigma_x^2 + \sigma_u^2)$ is the attenuation coefficient, and $\sigma^2_{x|z} = \pi\sigma_u^2$, which were obtained from an internal validation study (Turnbull *et al.*, 1997).

This bridge relation is then inverted to obtain a consistent adjusted estimator $\hat{\theta}$ for the true parameter $\theta = (\gamma, \beta, \lambda, \nu)$. Robust sandwich variance estimates were used to obtain standard errors. Details of the calculations are given by Turnbull *et al.*, (1997). Inference on the regression parameters of interest, $(\gamma,\beta)$, are summarized in lines 1a and 1b of Table 1 and compared with the results from the semi-parametric approach described next.

### 3.3 Skin cancer recurrence data from the NPC trial: semi-parametric modelling

Jiang *et al.* (1999) consider a semi-parametric approach to analyze the NPC study, for the purpose of removing the following assumptions used in the parametric approach: (i) Constant mean rate $\lambda$; (ii) Poisson distribution assumption on $Y_i(t)$ conditional on the random effects; (iii) Gamma distribution assumption on the frailties $\psi_i$.

Specifically, now we assume a model M(semi-par) for the observed mean response:

$$E[Y_i(t)] = H_i(t)\psi_i\lambda(t)\exp(a_i\gamma + x_i\beta), \text{ for all } i = 1,2,\ldots,n,\ t = 1,\ldots,K. \qquad (3)$$

Without loss of generality we may take $E[\psi_i] = 1$. Note that only the mean responses are modeled (not the higher moments) and the Poisson assumption is removed. Instead of the constant baseline mean rate $\lambda$, we use a nonparametric baseline mean rate $\lambda(t)$. There is no distributional assumption on frailties $\{\psi_i\}$ either. The semi-parametric approach is therefore considerably more flexible.

Here the parameter of interest is $\theta = (\gamma,\beta,\lambda(\cdot))$, where $\lambda(\cdot) = (\lambda(1),\ldots,\lambda(K))$. This is clearly a complex model, particularly because the frailties $\{\psi_i\}$ are unobserved, and only the surrogate $z_i$ is observed in place of $x_i$. Jiang *et al.* (1999) proposed an indirect inference approach based on the auxiliary model M′(semi-par) given by nonhomogeneous Poisson process model with multiplicative intensity $m(t)\exp(a_i g + z_i b)$. Note M′(semi-par) is simpler; it ignores the presence of frailties and measurement error. This leads to consideration of the intermediate statistic $\hat{s} = (\hat{g},\hat{b},\hat{m}(\cdot))$. Here $(\hat{g},\hat{b})^T$ is the Cox (1972) partial likelihood estimate and $\hat{m}(t)$ is a discrete intensity estimate for $\lambda(t)$ that corresponds to the Nelson-Aalen estimate of the cumulative intensity (see Andersen *et al.* 1993, Sec.VII.2.1). Standard computer software can be employed to compute these estimates — e.g. in Splus Release 6 (Insightful Corp. 2001). The auxiliary or 'naive' estimator $\hat{s}$ is computed ignoring both the random effect (by taking $\psi_i$ to be its mean 1) and the measurement error (by taking $x_i$ to be $z_i$). The dimensionality of $\hat{s}$ and $\theta$ are equal and so the $\hat{\theta}$ can be obtained from the bridge relation $s = s(\theta)$. Under the same Gaussian additive model for the measurement error as described in the last section, they go on to find the auxiliary or 'naive' parameter $s = (g,b,m(\cdot))$, the asymptotic mean of $\hat{s}$, leading to the bridge relations:

$$g = \gamma, \quad b = \pi\beta, \quad m(t) = \lambda(t)\exp(0.5\beta^2\sigma^2_{x|z}).$$

This bridge relation is then inverted to obtain a consistent adjusted estimator $\hat{\theta}$ for the true parameter $\theta = (\gamma,\beta,\lambda(\cdot))$. Robust sandwich variance estimates were used to obtain standard errors. Details of the calculations are given by Jiang *et al.* (1999). The results are summarized in lines 2a and 2b of Table 1.

Note that there is a qualitative difference between the estimates of treatment effect: in the general model M(semi-par), the treatment is no longer statistically significant. The results based on model M(semi-par) (line 1b) are robust against misspecifications of models on the response $\{Y_i(t)\}$– only a very general model for the mean need be postulated (cf. Lawless and Nadeau 1995). Assumptions on higher moments, such as those that might be imposed by the Poisson distribution, are not needed for valid inference.

When we compare the results of the previous parametric analysis described in Section 3.2 as displayed in lines 1a and 1b of Table 1, we find that the results are similar. This suggests that the much simpler constant intensity function model may well be adequate here.

## 4 Acknowledgements

The authors were partially supported by a grant from the U.S. National Institutes of Health and a grant from the National Science Foundation.

## 5 References

Andersen, P.K., Borgan, O., Gill, R.D. and Keiding, N. (1993). *Statistical Models Based on Counting Processes*. New York: Springer-Verlag.

Berk, R.H. (1966). Limiting behavior of posterior distributions when the model is incorrect. *Annals of Mathematical Statististics*, 37, 51-58.

Bickel, P.J. and Doksum, K.A. (2001). *Mathematical Statistics*, 2nd Ed. Upper Saddle River, New Jersey: Prentice Hall.

Breslow, N. (1990). Tests of hypotheses in overdispersed Poisson regression and other quasi-likelihood models. *Journal of American Statistical Association*, 85, 565-571.

Carroll, R.J., Ruppert, D. and Stefanski, L.A. (1995). *Measurement Error in Nonlinear Models*. London: Chapman and Hall.

Chiang, C.L. (1956). On regular best asymptotically normal estimates. *Annals of Mathematical Statististics*, 27, 336-351.

Clark, L.C. , Combs, G.F., Turnbull, B.W., Slate, E.H., Chalker, D.K., Chow, J., Davis, L.S., Glover, R.A., Graham, G.F., Gross, E.G., Krongrad, A., Lesher, J.L., Park, H.K., Sanders, B.B., Smith, C.L., Taylor, J.R. and the Nutritional Prevention of Cancer Study Group. (1996). Effects of Selenium Supplementation for Cancer Prevention in Patients with Carcinoma of the Skin: A Randomized Clinical Trial. *Journal of American Medical Association*, 276 (24), 1957-1963. (Editorial: p1984-5.)

Cox, D.R. (1962). Further results on tests of separate families of hypotheses. *J. R. Statist. Soc.* B, 24, 406-424.

Cox, D.R. (1972). Regression models and life-tables (with discussion). *J.R. Statist. Soc.* B, 34, 187-207.

Cox D.R. (1983). Some remarks on overdispersion. *Biometrika*, 70, 269-274.

Ferguson, T.S. (1958). A method of generating best asymptotic normal estimates with application to the estimation of bacterial densities. *Annals of Mathematical Statististics*, 29, 1046-1062.

Gail, M.H., Santner, T.J. and Brown, C.C. (1980). An analysis of comparative carcinogenesis experiments based on multiple times to tumor. *Biometrics*, 36, 255-266.

Gallant, A.R. and Long, J.R. (1997). Estimating stochastic differential equations efficiently by minimum chi-squared. *Biometrika*, 84, 125-141.

Gallant, A.R. and Tauchen, G. (1996). Which moments to match? *Econometric Theory*, 12, 657-681.

Gallant, A.R. and Tauchen, G. (1999). The relative efficiency of method of moments estimators. *Journal of Econometrics*, 92, 149-172.

Gourieroux, C., Monfort, A. and Renault, E. (1993). Indirect inference. *Journal of Applied Econometrics* 8S, 85-118.

Hansen, L.P. (1982). Large sample properties of generalised method of moments estimators. *Econometrica*, 50, 1029-1054.

Huber, P.J. (1967). The behavior of maximum likelihood estimates under nonstandard conditions. In *Procceedings of the Fifth Berkeley Symposium on Probability and Statistics I*, pp. 221-233. Berkeley: Univ. of California Press.

Insightful Corporation (2001). *S-PLUS 6*. Seattle, Washington.

Jiang, W. and Turnbull, B.W. (2001). The indirect method — robust inference based on intermediate statistics. *Technical Report, Department of Statistics, Northwestern University*, Evanston, IL, USA.

Jiang, W., Turnbull, B.W. and Clark, L.C. (1999). Semiparametric Regression Models for Repeated Events with Random Effects and Measurement Error. *Journal of American Statistical Association*, 94 111-124.

Kuk, A.Y.C. (1995). Asymptotically unbiased estimation in generalised linear models with random effects. *Journal of Royal Statistical Association* B, 57, 395-407.

Lawless J.F. and Nadeau, C. (1995) Some simple robust methods for the analysis of recurrent events. *Technometrics*, 37, 158-168.

McCullagh, P. and Nelder, J.A. (1989). *Generalized Linear Models*. New York: Chapman and Hall.

McFadden, D. (1989). A method of simulated moments for estimation of discrete response models without numerical integration. *Econometrica*, 57, 995-1026.

Newey, W.K. and McFadden, D. (1994). Large sample estimation and hypothesis testing. in *Handbook of Econometrics*, edited by Engle, R.F. and McFadden, D.L., Elsevier, New York.

Pakes, A. and Pollard, D. (1989). Simulation and the asymptotics of optimization estimators. *Econometrica*, 57, 1027-57.

Qin, J. and Lawless, J. (1994). Empirical likelihood and general estimating equations. *Annals of Statistics*, 22, 300-325.

Qu, A., Lindsay, B.G. and Li, B. (2000). Improving generalised estimating equations using quadratic inference functions. *Biometrika*, 87, 823-836.

StataCorp (1997). Stata Statistical Software, Release 5.0. Stata Corporation, College Station, Texas.

Thall, P.F. and Vail, S.C. (1990). Some covariance models for longitudinal count data with overdispersion. *Biometrics*, 46, 657-71.

Thompson, H.F., Grubbs, C.J., Moon, R.C. and Sporn, M.B. (1978). Continual requirement of retinoid for maintenance of mammary cancer inhibition. *Proceedings of the Annual Meeting of the American Association for Cancer Research*, 19, 74.

Turnbull, B.W., Jiang, W. and Clark, L.C. (1997). Regression models for recurrent event data: parametric random effects models with measurement error. *Statistics in Medecine*, 16, 853-64.

Wedderburn, R.W.M. (1974). Quasi-likelihood, generalized linear models and the Gauss-Newton method. *Biometrika*, 61, 439-447.

White, H. (1994). *Estimation, Inference and Specification Analysis*. Cambridge: Cambridge University Press.

## Resum

Es descriu en aquest treball l'anomenat mètode indirecte d'inferència. Aquest mètode es va desenvolupar inicialment en la literatura econòmica i nosaltres l'apliquem a l'anàlisi de la supervivència de dos conjunts de dades amb esdeveniments repetits. Aquest mètode acostuma a ser més convenient computacionalment que el mètode de màxima versemblança quan el model inclou, per exemple, complexitats tals com efectes aleatoris i errors de mesura, i també pot servir com a base per a inferències robustes sota hipòtesis menys estrictes sobre el mecanisme que ha generat les dades. El primer conjunt de dades conté temps de recurrència de tumors mamaris en rates i es modela fent servir un procés de Poisson amb covariàncies i fragilitats (frailties). El segon conjunt de dades involucra temps de recurrència de tumors de pell en individus d'un assaig clínic. S'aplica la metodologia a anàlisis de regressió, tant paramètrics com semiparamètrics, que acomoden efectes aleatoris i errors de mesura en les covariàncies.

*MSC:* 62N01, 62N02, 62P10

*Paraules clau:* Efectes aleatoris; equacions d' estimació; error de mesura; estimadors naive; frailty; inferència indirecta; quasi-versemblança; regressió de la taxa de risc; sobredispersió; robustesa

SORT 27 (1) January-June 2003, 95-112

# Optimization of touristic distribution networks using genetic algorithms

Josep R. Medina[a] and Víctor Yepes[b] *

[a] *Universidad Politécnica de Valencia*
[b] *Agència Valenciana del Turisme*

## Abstract

The eight basic elements to design genetic algorithms (GA) are described and applied to solve a low demand distribution problem of passengers for a hub airport in Alicante and 30 touristic destinations in Northern Africa and Western Europe. The flexibility of GA and the possibility of creating mutually beneficial feed-back processes with human intelligence to solve complex problems as well as the difficulties in detecting erroneous codes embedded in the software are described. A new three-parent edge mapped recombination operator is used to solve the capacitated vehicle routing problem required for estimating associated costs with touristic distribution networks of low demand. GA proved to be very flexible especially in changing business environments and to solve decision-making problems involving ambiguous and sometimes contradictory constraints.

*MSC:* 90B20

*Keywords:* Distribution networks, vehicle routing problem, tourism demand, air transportation, genetic algorithms, edge mapped recombination operator

## 1 Introduction

Travel and tourism represent a total market which is of interest world-wide (Middleton, 1988). Regarding tourism, the transport of travellers to their destinations through the appropriate distribution channels constitutes one of the essential elements of competitiveness for the sector. Operational factors, infrastructure, equipment and

* *Address for correspondence:* J. R. Medina. Department of Transportation. Universidad Politécnica de Valencia. Camino de Vera s/n, 46022 Valencia, SPAIN. Tel.: 34-96-387-7375. Fax: 34-96-387-7379. jrmedina@tra.upv.es victor.yepes@turisme.m400.gva.es

Received: September 2002
Accepted: February 2003

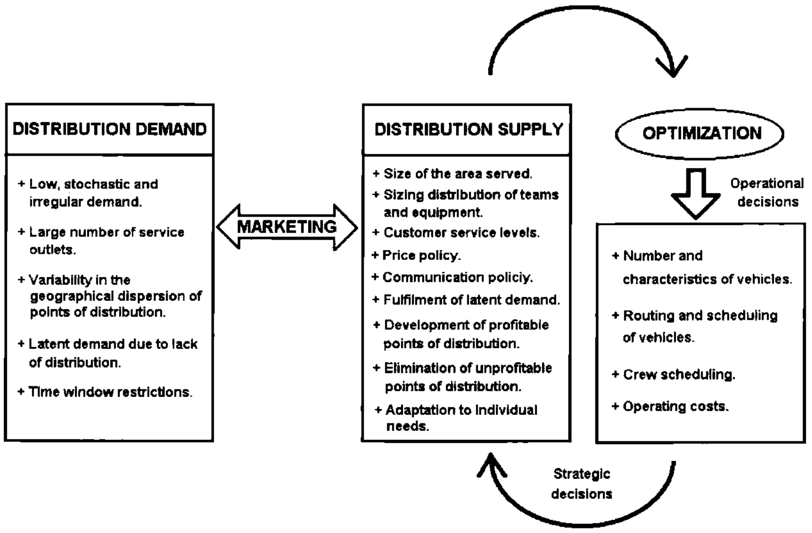
DISTRIBUTION DEMAND
+ Low, stochastic and irregular demand.
+ Large number of service outlets.
+ Variability in the geographical dispersion of points of distribution.
+ Latent demand due to lack of distribution.
+ Time window restrictions.
MARKETING
DISTRIBUTION SUPPLY
+ Size of the area served.
+ Sizing distribution of teams and equipment.
+ Customer service levels.
+ Price policy.
+ Communication policiy.
+ Fulfilment of latent demand.
+ Development of profitable points of distribution.
+ Elimination of unprofitable points of distribution.
+ Adaptation to individual needs.
OPTIMIZATION
Operational decisions
+ Number and characteristics of vehicles.
+ Routing and scheduling of vehicles.
+ Crew scheduling.
+ Operating costs.
Strategic decisions

***Figure 1**: Marketing and managing low demand distribution networks.*

government regulations are elements that affect the cost, speed and convenience with which a traveller may reach his destination. Therefore, operating costs are a primary input to pricing decisions. The decision-making problems associated with the management of touristic distribution networks of low demand require the use of efficient methods for routing optimization. Vehicle routing optimization at the operational level has a significant impact on both tactical and strategic levels, affecting cost estimation, optimum fleet size and optimum publicity policy. Figure 1 shows the variables to be taken into consideration for low demand distribution networks.

Many touristic services are sold in packages including round trip transportation from a fixed origin or hub airport to the corresponding touristic destinations; prices are fixed months before the transportation demand is known and last minute cancellations and new clients frequently change the demand estimations. When the number of passengers to be transported to each destination is low compared to the capacity of the optimum vehicle at the corresponding travel distance, the benefits or losses of the tour-operator are critically dependent on the optimization of the distribution system. Typically, the management of distribution networks implies the search for solutions which must fulfil a variety of objectives and constraints with a minimum use of resources; however, the number of possible solutions grows more than exponentially with the number of destinations and fleet size, and the optimal solution is not workable with exact optimization techniques because of the required computing time. The variety of objectives, resources and constraints that characterise actual management and transportation problems makes these problems inappropriate for conventional optimization techniques, which only search for the optimal solution in a deterministic way. In these circumstances, approximate methods of resolution as Genetic Algorithms

(GA), which emulate an efficient optimization strategy developed by Nature, do not attempt to find the optimal solution, but rather a reasonably good feasible solution depending on the computational effort. In the field of combinatory optimization and transportation problems, the term of metaheuristic is commonly used for these methods, while in other areas these methods are known as intelligent systems (Goonatilake *et al.*, 1995).

The management of low demand distribution networks, in which efficient vehicle routing is essential, generates decision-making problems with huge spaces of solutions. When the exact optimization techniques are not viable, heuristics, meta-heuristics, and probabilistic methods are reasonable alternatives. These methods do not guarantee finding the absolute optimal solution, but they do provide a limited search for feasible solutions, taking advantage of the particular characteristics of the problem under study, or taking advantage of the subjective human perception of what the characteristics of a good solution to the problem should be. Intelligent systems are appropriate for very complex optimization problems with multiple variables under a variety of constraints, including ambiguous and also contradictory objectives, such as those found in real-world problems. Goonatilake *et al.* (1996) described a number of methods for a variety of business and financial applications, from neural networks guiding direct mail campaigns to GA for credit assessment. Intelligent systems may also be used in data mining, transforming information into knowledge (Fayyad *et al.*, 1996). A case in point is the use of a feed-forward neural network pruned by simulated annealing for modelling runup and overtopping of rubblemound breakwaters (Medina, 1999). Yepes (2002) systematizes the set of methods of heuristic optimization and establishes the state of the technique regarding the used procedures in the solving of the Vehicle Routing Problem with Time Windows (VRPTW) and its extents.

Genetic Algorithms (GA) are meta-heuristic methods that usually serve as very flexible and robust tools for complex combinatorial optimization problems (Díaz *et al.*, 1996). In this paper, the general design procedure of a GA for optimization is described in reference to an example of GA originally developed to solve the classical Travelling Salesman Problem (TSP), but modified to solve the Capacitated Vehicle Routing Problem (CVRP). The practical lesson here is that a GA code can be adapted to solve increasingly complex problems with a minimum additional effort. In a dynamic business environment in which laws, rules, relationships and variables are constraints which change from day to day, it is imperative to have flexible and robust decision-making support tools which can easily change as quickly as the business environment does. There are a number of tourism studies in which these techniques are employed (see Canestrelli and Costa, 1991; Kottke, 1988; van der Knijff and Oosterhaven, 1990; Teodorovic, Kalic and Pavkovic, 1994; Brusco *et al.*, 1995, Hurley *et al.*, 1998; Medina and Yepes, 2000).

At the strategic level, an efficient tool to optimise vehicle routing problems is necessary to better estimate the travel and distribution costs associated with specific destinations and periods of low touristic demand. Offering a new touristic destination or eliminating an old one are decisions that require estimating the total economic impact. This in turn means taking into consideration the overall passenger distribution network. Strategic decisions require robust optimization methods which provide solutions related to the lowest possible transportation costs and the most reliable economic estimates.

At the tactical level, fleet size and characteristics of the vehicles must be decided on the basis of estimations of passenger transportation demand, distribution optimization and estimation of costs. Once the strategic decisions have been taken, affecting the structure of the distribution network, the optimization of fleet characteristics will favour to obtain the best results at the operational level.

At the operational level, routing and scheduling of vehicles must be adapted to a constantly changing non-deterministic passenger transportation demand. In this paper, an application at the operational level is provided considering a passenger distribution network with 30 touristic destinations in the Western Mediterranean area and Alicante (Spain) as the hub airport. Touristic demand is not deterministic and the CVRP is optimised using GA; the goal of these GA is to minimise the total transportation cost in a given scenario of low demand with a given stochastic structure. Therefore, GA serve not only as decision-making support tools at the operational level but at the tactic and strategic levels as well. Darwinian evolution emulated by GA is so intrinsically flexible and robust that it is hard to imagine a decision-making problem in which GA were not applicable; however, the practical problem here is to find efficient GA designs for each specific application.

## 2 Implementation of genetic algorithms

During the past decade, applications of a number of optimization techniques commonly used in artificial intelligence have been published in a variety of technical and scientific journals (Ansari and Hou, 1997). Evolutionary Programs (EP), Genetic Algorithms (GA), Neural Networks, Tabu Search, Simulated Annealing and Fuzzy Systems are some of the new techniques which have proved to be very effective in data mining and knowledge discovery (Fayyad *et al.*, 1996). They are also effective for solving complex optimization problems in a number of technical fields. GA emulate the natural evolution processes of complex living creatures. Populations of solutions form successive generations, in which survival and reproduction are probabilistically controlled by the best fitting criterion. The specific characteristic of GA is the efficient parallel exploration of the space of solutions based on the exploitation of good solutions induced by the crossover operators and the exploration of new solutions forced by the mutation operators. GA may run with single or several isolated populations or "islands" with a migration system; the size of the population may also differ from one application to another. The diversity not only increases with the size of the populations and the number of islands, but also with the computational cost. Additionally, GA are optimization techniques which may easily be adapted to run in co-operation with human intelligence because humans may provide specific sub-optimal solutions and GA have the ability to assimilate any solution during the evolution process. This characteristic may be crucial in finding feasible implementations of GA designed to replace decision-making tasks carried out by humans.

Typically, GA first define a specific letter or number code to represent any solution of the problem. Once the codification of the problems is defined, an initial population of solutions is randomly generated or created using any subjective or objective procedure. Then a cyclic process of selection, crossover and mutation changes the population in successive generations towards the optimum. If the diversity is too high, the evolution process may degenerate; on the contrary, if the population diversity is too low, the evolution process may stagnate in a local minimum and only a small proportion of the space of solutions are later explored during the computer time available.

There are several ways to implement GA for solving complex optimization problems (Davis, 1996), but the efficiency in a specific application depends exclusively on the GA design. This paper describes in detail an eight-step procedure, proposed by Medina (1998), which indicates the corresponding natural analogy and provides key ideas regarding the specific application to the CVRP example described in this paper.

### *1) Genetic architecture*

Chromosomes made of strings of amino acids define the genotypes of living creatures. GA use chromosomes of strings of letters or numbers to define the characteristics of the solutions, named "individuals" in GA jargon. An explicit or implicit relationship between letters or numbers in the chromosome and the corresponding actual solution must be fixed in advance. Each possible solution must be related at least to one specific chromosome, which has to be attainable by crossover and mutation. Any possible chromosome obtainable by crossover and mutation must have the possibility of being decoded, and only one specific individual must be associated with a given chromosome. By analogy with natural evolution, the letters and numbers of the chromosome are named "genes", and the alternative values of each gene are named "alleles". A precise description of the genetic architecture and the possible alleles to each gene is required to run efficient GA; the effectiveness of the crossover operation is very sensitive to the selected genetic codification. In the study described in this paper, the airports are ordered and numbered, the chromosome of an individual is a string of numbers representing the successive airports to be visited, and the hub represents a new aeroplane to carry the passengers of the succeeding route. The length of the chromosome is equal to the number of airports plus the fleet size, and the alleles of each gene are the numbers of the airports. Figure 2 shows an illustrative example of a chromosome indicating a solution with two planes, one hub identified by 0 and nine airports identified by the corresponding numbers; the route of the first plane is {hub-1-2-3-5-4-hub} and the route of the second plane is {hub-6-7-9-8-hub}.

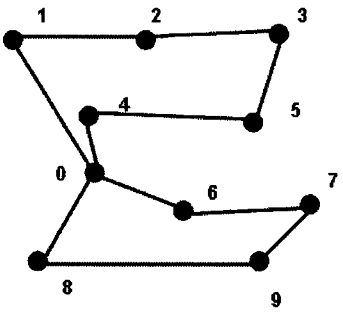

*Figure 2: Solution identified by chromosome {0-1-2-3-5-4-0-6-7-9-8-0}.*

### 2) *Population size and population distribution*

Natural populations of a given species are distributed in the space forming independent groups isolated from the rest of the species or groups connected by periodic migration movements. In the case of this research, the GA software is run with a personal computer with a single processor; evolutions with isolated populations with a number of individuals of five to ten times the number of destinations has given the best results. After generating the first independent evolutions in ten isolated islands, a sample of each one of the ten isolated final populations was taken to run the last evolution.

### 3) *Initial population*

The final result of the optimization process using GA is usually not sensitive to the fitness of the individuals in the initial population. Pure random alleles can be selected for each gene of each individual in the initial population; however, heuristics may help to improve the fitness of the initial population and reduce the time for convergence to the optimal solution. In the study described in this paper, radial order from the hub, minimum distance criteria between destinations and 2-opt branch exchange were used to form the individuals of the initial populations. Additionally, solutions created directly from human intelligence might be introduced in the intermediate isolated final populations to participate in the last evolution.

### 4) *Evaluation and cost function*

The natural selection tends to eliminate deterministically or probabilistically the individuals, which are not adapted to the environment or are less fit for the objective.

Given a specific individual and its environment, there is a cost associated to the individual. The environment (cost function) may or may not be stationary; in the case of the TSP, the cost associated to a solution is the total travel distance, the optimum being the one with the lowest cost. In the case of CVRP, the vehicles have a limited capacity and the demand of transportation may change on a daily basis. In a number of real-world problems, the environment is stochastic and non-stationary; in these cases, the solutions of the past may be used as members of the initial population of future evolutions in a different or a changing environment. Evaluation is critical for explaining the flexibility of GA because all the variables and constraints have to be reduced to the single variable "cost". Not only can objective variables such as "distance" or "time" be used to define the cost; but qualitative, ambiguous as well as subjective opinions may be used in the cost function to provide a better guide to the optimization process. For instance, an accident in a given airport or political turmoil can make it subjectively unpleasant for passengers to use routes which imply the use of specific airports during certain periods of time. In addition to the objective cost induced by "distance" and "time", it may be reasonable to include a subjective virtual cost associated with the use of specifically hard-to-deal airports. Experience can also help optimization because it may define virtual costs associated with subjective characteristics of the solution that will tend to favour solutions with the prescribed valuable characteristics. In the case which is analysed in this paper, the cost function takes into consideration the objective economic costs associated with travel distance and time of each vehicle of the fleet, but also includes virtual costs associated to violation of constrains as maximum capacity, maximum duration of the route, maximum number of flights for each crew in a route, etc. Legal and social constraints as well as quality of service must be transformed in virtual costs to properly guide the optimization process to a satisfactory solution.

### *5) Selection*

Emulating natural selection, the individuals with lower costs in a given population have higher survival and reproduction rates. The worst individuals have to be deterministically or probabilistically eliminated; the probability of elimination may be proportional to the cost, proportional to the order, etc. The GA employed for this study uses a probability of survival inversely proportional to the order of the individual in the population, having the best individual two times the probability of survival than the second best, three times the third, etc. With these probabilistic criteria of survival, the individuals are randomly selected from a given generation to produce the offspring of the following generation. Additionally, the method included some degree of elitism because a very small probability of selection is always given to the absolute champion, although the absolute best individual found during the evolution does not belong to the current generation.

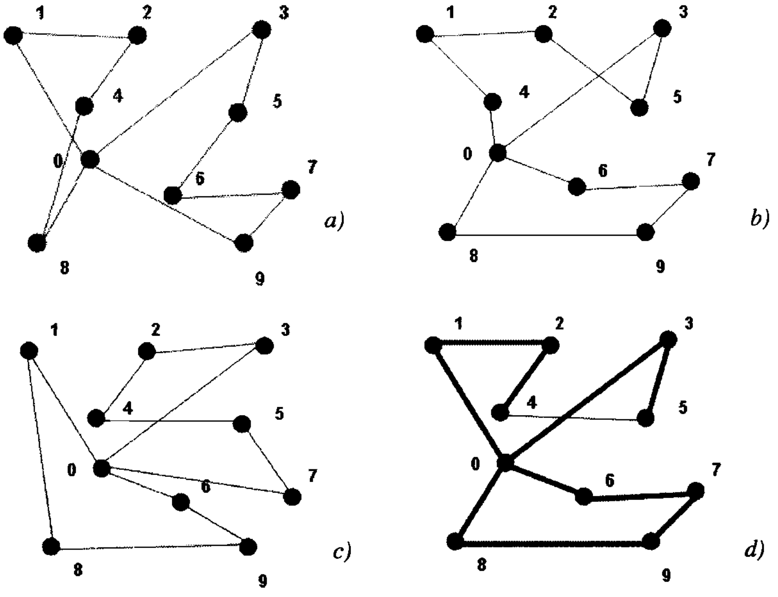

***Figure 3**: Three parents edge mapped recombination operator generates offspring (d) based on parents (a), (b) and (c).*

## *6) Crossover*

The crossover operation must exploit the good solutions transferring and spreading the desired characteristics of the best individuals from one generation to the next. The crossover operators read the genetic information of the selected individuals (parents) of a generation to produce viable individuals for the following generation (offspring). There are usually two parents who must have different genes. The crossover operator must not only ensure that the offspring are different from their parents, but that they also have some of their parents' characteristics. One of the crossover operators used in this study is the classical one-point crossover operator which reads the genetic code of two selected parents, cuts the genetic chain in the same point, interchanges the chain tails, and changes genes to make viable offspring. A second crossover operator of the family edge mapped recombination operator described by Whitley *et al.* (1996) is also used in this study; this new operator is a generalisation of the edge mapped recombination operator considering three or more parents to produce one single offspring. The best performance has been obtained only with three parents; the offspring is formed by taking the routes that are present in the two or the three parents and completing feasible routings using the minimum distance criterion to form the offspring. Figure 3 shows the offspring obtained from three parents using the three parent edge mapped recombination operator.

The offspring have the arcs that are presents in two or three of the three parents; if three arcs are connected in a specific node, one of the three is randomly eliminated. Finally, the full offspring chromosome is obtained connecting the rest of the nodes randomly. Arc {0-3} belongs to offspring (d) represented in Figure 3 because it is part of the routes of the three parents; arcs {7-9}, {3-5}, and others belong to offspring (d) because they are part of the routes of the parents (a) and (b); arcs {0-1} and {2-4} belong to offspring (d) because they are part of the routes of the parents (a) and (c); arc {8-9} belongs to offspring (d) because it is part of the routes of the parents (b) and (c). No node in the offspring has three arcs and only the connection of nodes 4 and 5 is necessary to complete the chromosome of the offspring.

### *7) Mutation*

The mutation has to explore the space of solutions providing the desired diversity in the population. The mutation operators change the alleles in a genetic code randomly, usually generating poor solutions that are eliminated in the successive selection processes. However, sometimes the mutation is beneficial to the solution and the new characteristic is rapidly spread in the population by selection and crossover. The GA used in this paper have seen different mutation operators designed to solve the specific local problems of routing: (1) exchange of one pair of random genes in the chromosome, (2) exchange of two pairs of random genes in the chromosome, (3) exchange of two random strings of genes of random length in the chromosome, (4) cut the chromosome and exchange first and second part, (5) change the order of a random string of genes of random length, (6) exchange two consecutive strings of genes of random length, and (7) select a string of genes of random length and put it in the first position of the chromosome shifting the corresponding genes.

### *8) Probability of crossover and mutation*

According to Mitchell (1996), efficient crossover rates fall in the range of 75% to 95%, while bit mutation rates must be in the range of 0.1% to 1% to be effective. However, she indicated that for most applications crossover and mutation rates should not be constant, but should change during the evolution process to maintain a balance between exploration and exploitation of solutions. Julstrom (1995) proposed a dynamic assignment of probabilities for the different operators depending on the improvement obtained during the evolution process. In this study, the crossover operators have initial rates of 5% and 15% respectively, and the initial mutation rates are in the range of 5% to 20%. However, if no improvement is detected after a fixed number of generations, the probability rates assigned to each operator change following a first order auto-regressive process which is activated when stagnation is detected.

## 3 Application of genetic algorithms to routing problems

The GA described in this paper has been implemented in Visual BASIC to be used in a personal computer. In its present form, a homogeneous fleet is considered with the following input variables: (1) latitude and longitude of the hub and the destination airports, (2) mean velocity and maximum capacity of the aeroplane, (3) take-off, landing and taxi time plus mean airport delays, (4) costs associated with aeroplanes, crews, passengers and travel distance, and (5) the number of passengers to be transported to and picked up from each destination to the hub airport. Before using the GA software to solve transportation problems in different scenarios, it is first necessary to run several typical cases in order to check the different parameters controlling the evolution, such as the probabilities of crossover and mutation, the population size, the number of islands and the number of generations. The diversity in the population during evolution has to be studied in order to avoid both premature stagnation and degenerative processes. Once the parameters of the genetic program are fixed, GA are ready to solve efficiently the transportation problem associated with the simulated scenarios representing the unknown future, calculating costs of different alternatives.

A specific GA implementation requires making dozens of decisions about structure and parameters of the algorithm. There are millions of alternative GA implementations of a given problem and the efficiency of a specific GA implementation is highly nonlinearly dependent on the selected parameters; therefore, human intelligence is required to define reasonable structural parameters, crossover and mutation rates, etc. The results may differ with different GA parameters (i.e. crossover and mutation rates, etc.) but the intrinsic robustness of GA makes it easy to find a reasonably good implementation after some exploratory work. Nevertheless, it is not possible to be sure that a specific GA implementation of a given problem cannot be significantly improved by changing the GA parameters. For this reason, the GA algorithm employed in this research uses a dynamic assignment of crossover and mutation rates.

For a given problem, the possible comparisons of the quality of solutions using GA and other intelligent systems (i.e. Simulated Annealing, Neural Networks, Tabu Search, etc.) are inconclusive (see Yepes, 2002). Only specific computer codes can be compared and the specific implementations of intelligent systems depend critically on the author's competence in designing the algorithms and detecting erroneous code. For a given problem, the comparisons with simple heuristics are also inconclusive. The popular heuristics (i.e. nearest neighbour procedure and 2-opt branch exchange) are efficient for solving simple problems (i.e. TSP) with simple cost function variables (i.e. minimum total distance). For more realistic problems (i.e. CVRP) with complex cost functions variables (i.e. total distance, total time, number of aeroplanes, service interruption cost, maximum route time, etc.), the popular heuristics are not efficient.

GA may easily interact with human intelligence in solving the problems because humans may include solutions during the evolutionary process. The evolution may then "learn" from the human skills for optimization, and once a different solution is obtained, humans may also learn from how the machine successfully changed the human solution to find a better solution. This mutually beneficial response may be crucial for the final success in using GA as a decision-maker support system of a company.

Although GA are indeed extremely flexible and robust methods for solving complex optimization problems, it is necessary to warn the readers of the difficulty in checking the software code. Because of the intrinsic robust performance of GA, it is relatively easy to write an erroneous code, which may become embedded in the main code, being almost impossible to detect. The difficulty in detecting errors lies in the fact that neither the optimum solution nor a reasonable pace of progression of the evolutionary process is known in advance. If the basic evolutionary code is well written, it may give apparently good solutions although a certain erroneous code reduced significantly the effectiveness of the algorithm. The software could be checked using known optimal solutions of small size problems, but the result could never prove the absence of erroneous code, because the robustness of GA frequently allows one to find the optimal solution of small size problems. When an erroneous code is detected and corrected, the GA is more efficient and able to solve larger problems to the optimal point. Therefore, an undetected erroneous code should always be considered as a probability included in any GA code; it can always be generated by accident and sometimes it may also be generated intentionally when a team works together to write the code.

## 4 Application of genetic algorithms to optimize a low demand air transportation problem

For illustrative purposes, an application of GA is given below to analyze and optimize a low demand air transportation problem. The problem can be described by four factors: (a) the low demand distribution problem, (b) the model of transportation demand, (c) the fleet and transportation constraints and (d) the optimization and results.

### *a) The low demand distribution problem*

In this study, we use the CVRP version to solve the operational problem which affects a tour-operator based in Alicante (Spain) specialised in non-massive touristic destinations in the Western Mediterranean area. When prices must be fixed in advance and demand is low and non-deterministic, it is of critical importance to use a flexible and robust optimization technique to minimise the transportation cost and to estimate mean costs in different scenarios. An estimation of the demand is known in advance, but a significant random component due to last-minute cancellations is always present.

*b) The model of transportation demand*

In this study, a stationary lognormal pdf (probability density function) and a double-variable first-order autoregressive stochastic structure are used to model the demand of transportation from the hub to each destination and from each destination to the hub. This model resembles the wave climate simulator proposed by Medina *et al.* (1991) for simulating time series of significant waves and periods in the Pacific Ocean. If demands follow the lognormal pdf, the normalised time series may be defined as:

$$x(i,n) = \frac{(\log[q_1(i,n)]) - Q_0(i)}{SQ_0(i)} \tag{1a}$$

$$y(i,n) = \frac{(\log[q_2(i,n)]) - Q_0(i)}{SQ_0(i)} \tag{1b}$$

in which $q_1(i,n)$ is the transportation demand during the day $n$ from the hub airport to destination $i$,and $q_2(i,n)$ is the transportation demand from destination $i$to hub. $Q_0(i)$ and $SQ_0(i)$ are the parameters of the lognormal pdf. In this study, $Q_0(i)$=2.5 and $SQ_0(i)$=0.4 for all destinations; the average transportation demand is about 13 passengers with a coefficient of variation of 43%. The stochastic structure is given by

$$x(i,n) = A\,x(i,n-1) + \sqrt{(1-A^2)}\;w(i,n); \quad n = 1,2,3,\ldots \tag{2}$$

$$y(i,n) = B\,x(i,n-\delta) + \sqrt{(1-B^2)}\;v(i,n); \quad n = 1,2,3,\ldots \tag{3}$$

in which $x(i,n)$and y$(i,n)$ are the normalised time series; $w(i,n)$and v$(i,n)$ are independent white noises; A and B are the correlation parameters; and $\delta$ is the time lag parameter. In this paper, A=0.50, B=0.95 and $\delta$ =7.

*c) The fleet and transportation constraints*

The aeroplanes have a maximum capacity of 50 seats, a range of 2,250 km, and a travel speed of 240 knots. Each aeroplane departs from the hub airport with the passengers corresponding to the destinations in its prescribed route and has to deliver them to the destination, taking all the passengers in the route that have to return to the hub airport the same day. The number of planes to address the transportation demand and the total distance and flight time has to be minimum, without exceeding the capacity of the aeroplane and a maximum routing time of 10 hours. In the example given in this paper, the hub airport is Alicante (Spain) and the 30 destinations are: Ajaccio, Brest, Burdeos, Cannes, Cardiff, Dublin, Tangier, Casablanca, Tunis, Malta, Cagliari, Genoa, Palermo, Venezia, Lisbon, Oporto, Girona, Vitoria, Santander, Oviedo, Santiago de Compostela, Pamplona, Zaragoza, Granada, Jerez, San Sebastian, Reus, Mallorca, Menorca and Ibiza. The region of Alicante (Spain) has more than 53,000 hotel rooms and a theme park for mor than 2 million visitors (AVT, 2001).

### *d) The optimization and results*

The GA optimization of passenger distribution at the operational level affects the decision-making process at both the tactical and strategic level. On the one hand, the effectiveness of GA affects the fleet size and prices to be charged; on the other hand, the sensitivity analysis for increasing or reducing demand of specific destinations may guide the publicity policy and the suppression or addition of new destinations (see Figure 1). The demand simulator described by Equations 1 to 3 provides the scenario in which the GA operate providing the best transportation solution each day. The service levels are considered in the routing optimization process by the maximum routing time and the economic extra-cost due to service interruption (over-booking). Figure 4 shows a typical graphic output for the CVRP defined above. Unit prices, aeroplane characteristics and travel conditions may be easily modified to affect only the cost function. Additionally, a graphic interface allows for the introduction of man-made solutions in the evolution process of GA. The cost function used in this example includes the following factors (rates): aeroplane travel distance (3 euro/mile), number of vehicles (3,000 euro/aeroplane), number of crews (1,800 euro/crew), optimal and maximum route time (6 and 9 hours), optimal and maximum load (45 and 50 passengers), mean aeroplane speed (240 knots), and service breakage (6,000 euro/time or load failure). In addition to the actual costs used to calculate the results given in Table 1, the cost function has virtual costs associated to non-optimal time and load solutions that guide the optimization problem to optimal solutions.

***Table 1**: Estimated daily mean value and coefficient of variation of the most significant transportation load and cost variables.*

| Variable | Daily mean | Coefficient of variation |
|---|---|---|
| Number of passengers from and to each destination | 13 | 43% |
| Total number of passengers transported | 800 | 6.1% |
| Number of passengers per kilometre to the hub | 605,000 | 6.4% |
| Occupation of seats in aeroplanes (%) | 65 | 4.2% |
| Cost of passenger transported from or to hub (EUR) | 118 | 2.7% |

Although the low demand is not known in advance, the fleet and the crews are known in advance. In the example, 4 aeroplanes and 15 crews are considered to satisfy the demand. Because transportation supply is stable, but transportation demand has a significant random component (i.e. last-minute cancellations), transportation load,

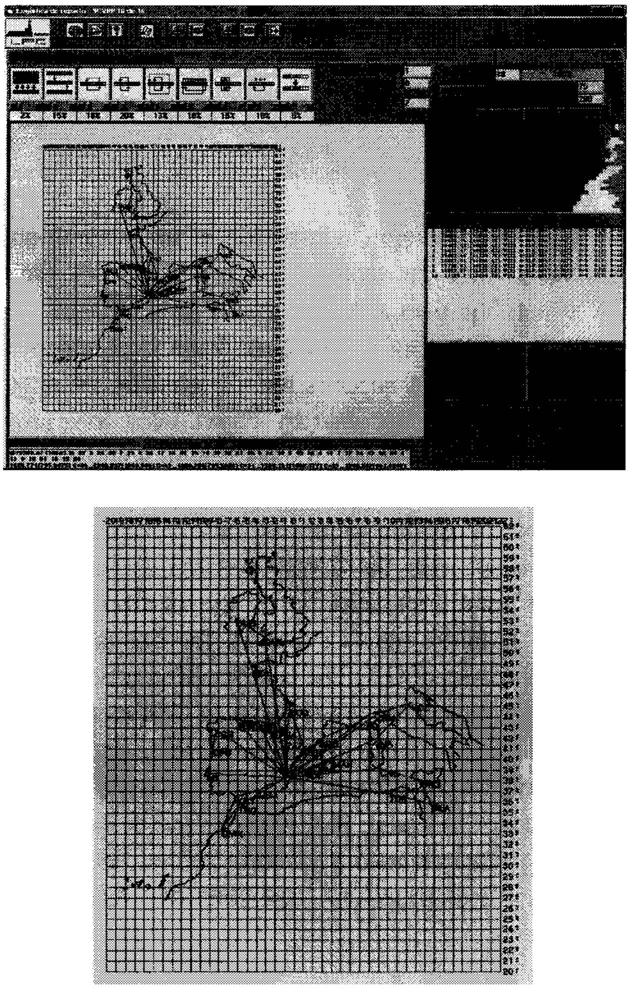

***Figure 4**: Typical graphic output for the CVRP.*

optimal routes and profitability change daily. A simulation of thirty days with the parameters given above was generated and the corresponding GA optimised solutions were obtained. In this simulation, the mean travel time was 3.0 hours for passengers and 5.8 hours for crews. The daily mean values and coefficient of variations of the most significant transportation load and cost variables are given in Table 1. A sharp reduction of variability from the isolated and disperse transportation demand (CV=43%) to the cost of transported passenger (CV= 2.7%) is obtained everyday by the method of routing optimization using GA. The mean occupation of seats in the aeroplanes was 65%. The heuristics of nearest neighbour procedure and 2-opt branch exchange is almost as efficient as GA minimising total distances, but it is highly inefficient optimising more realistic multi-objective problems such as the CVRP described in this paper. However, efficient heuristics for TSP may be used to create the initial population of the GA, reducing the computational effort to find the optimal solution of the transportation problem.

## 5 Conclusions

GA are inspired in Darwinian natural evolution and provide very flexible and robust optimization techniques which are being applied in a wide range of scientific and technical fields. GA may be designed to work in co-operation with human intelligence in solving optimization problems, generating a mutually beneficial feedback process that might be essential in supporting or replacing human decision-making systems. The flexibility of GA allows for operations in changing business environments including the consideration of subjective, ambiguous and sometimes contradictory constraints. The intrinsic robustness of the optimization methods based on GA makes them very attractive for a variety of applications; however, robustness is also responsible when erroneous codes are accidentally embedded in the GA software which is usually very difficult to detect.

Although it is hard to imagine a decision-making problem in which GA were not applicable, efficient results require adequate GA designed for each specific application. Eight basic elements to design GA for general purposes are described in detail, pointing out the parallelism with natural evolution while a specific application to a given touristic distribution problem is analysed. The genetic architecture, the cost function and the crossover operators are the key elements for a successful implementation of GA for most specific applications. The GA employed in this paper use a new three-parent edge mapped recombination operator which was found to be very efficient. A genetic architecture and cost function easily changed the problem typology from the classic TSP to CVRP. The flexibility and capability to adapt to changing environments are indeed the strong points of GA.

GA are applied to solve a passenger distribution CVRP at the operational level with the hub airport in Alicante (Spain) and 30 touristic destinations in the Western Mediterranean area. The low touristic demand is modelled as a stationary lognormal with a double-variable first-order autoregressive stochastic structure. GA optimise the distribution of passengers day by day generating a precise description of the transportation scenarios and the costs associated with them. In the simulations, the coefficient of variations of both the cost of transported passenger and the cost of transported passengers per kilometre to hub were significantly lower than the coefficient of variations of transported passengers and transported passenger per kilometre to hub. The optimization method used as cost estimator affects the decision-making process both at tactical and strategic levels.

## 6 Acknowledgements

The authors would like to acknowledge the financial support provided by the *Direccön General de Enseñanza Superior e Investigación Científica* (Spanish Ministry of Education) under grant MAR98-0339, and the information provided by the *Agència Valenciana del Turisme* (*Generalitat Valenciana*). Juan J. Navarro wrote the software code and Francisco Amor gave valuable comments on the marketing aspects. The authors would also like to thank Debra Westall for her assistance in the revision of the final version of this paper.

## 7 References

Agència Valenciana del Turisme (2001). *El turismo en la Comunidad Valenciana* (in Spanish). Agència Valenciana del Turisme. Valencia (Spain).

Ansari, N. and Hou, E. (1997). *Computational Intelligence for Optimization*, Kluwer Academic Publishers. Boston (USA).

Brusco, M.J., Jacobs, L.W., Bongiorno, R.J., Lyons, D.V. and Tang, B.X. (1995). Improving personnel scheduling at airline stations. *Operations Research*, 43, 741-751.

Canestrelli, E. and Costa, P. (1991). Tourist carrying capacity: a fuzzy approach. *Annals of Tourism Research*, 18, 295-311.

Davis, L. (1996). *Handbook of Genetic Algorithms*. International Thompson Computer Press. Boston (USA).

Diaz, A., Glover, F., Ghaziri H.M., González, J.L., Laguna, M., Moscato, P. and Tseng, F.T. (1996). *Optimización Heurística y Redes Neuronales en Dirección de Operaciones e Ingeniería*, (in Spanish) Editorial Paraninfo, S.A., Madrid (Spain).

Fayyad, U.M., Piatetski-Shapiro, G., Smyth, P. and Uthurusami, R. (1996). *Advances in Knowledge Discovery and Data Mining*. MIT Press.

Goonatilake, S. and Treleaven, P. (1996). *Intelligent Systems for Finance and Business*. John Wiley.

Hurley, S., Moutinho, L. and Witt, S.F. (1998). Genetic algorithms for touristic marketing. *Annals of Tourism Research*, 25, 2, 498-514.

Julstrom, B.A. (1995). What have you done for me lately? Adapting operator probabilities in a steady-state genetic algorithm. *Proc. 6th International Conference on Genetic Algorithms*, Morgan Kauffmann Pub., San Mateo, California, pp. 81-87.

Kottke, M. (1988). Estimating economic impacts of tourism. *Annals of Tourism Research*, 15, 122-133.

Medina, J.R. (1998). Algoritmos genéticos para la optimizaci6n de redes de distribución, (in Spanish). *Actas del X Congreso Panamericano de Ingeniería de Tránsito y Transporte*, Santander 1998, Ministerio de Fomento (Spain), pp. 339-347.

Medina, J.R. (1999). Neural network modelling of runup and overtopping. *Coastal Structures'99* (Vol. 1), 421-429. A.A. Balkema.

Medina, J.R., Giménez, M.H. and Hudspeth, R.T. (1991). A wave climate simulator. *Proc. XXIV IAHR Congres*, IAHR, (B) 521-528.

Medina, J.R. and Yepes, V. (2000). Optimización de redes de distribución con algoritmos genéticos, (in Spanish). *Actas del IV Congreso de Ingeniería del Transporte*, 1, 205-213. Valencia (Spain).

Middleton, V.T.C. (1996). *Marketing in Travel and Tourism*. Heinemann Professional Publishing. Oxford (U.K.).

Mitchell, M. (1996). *An Introduction to Genetic Algorithms*. MIT Press.

Teodorovic, D., Kalic, M. and Pavkovic, G. (1994). The potential for using fuzzy set-theory in airline network design. *Transportation Research Part B-Methodological*, 28, 103-121.

Van der Knijff, E.C. and Oosterhaven, J. (1990). Optimizing tourist policy: a linear programming approach. *Regional Studies*, 24, 55-64.

Whitley, D., Starkweather, T. and Shaner, D. (1996). The travelling salesman and sequence scheduling: quality solutions using genetic edge recombination. In *Handbook of Genetic Algorithms* (L. Davis Ed.), International Thompson Computing Press, pp. 350-372.

Yepes, V. (2002). *Economic heuristic optimization applied to VRPTW type transportation networks* (in Spanish). Doctoral Dissertation. Higher Technical School of Civil Engineering. Polytechnic University of Valencia (Spain), 352 pp.

## Resum

Es descriuen els vuit elements bàsics per al disseny d'algoritmes genètics (AG), i s'apliquen a la resolució d'un problema de distribució de passatgers de baixa demanda amb un exemple centrat a l'aeroport d'Alacant i 30 destinacions turístiques del Nord d'Àfrica i l'Oest d'Europa. Es descriu tant la flexibilitat dels AG com la possibilitat de crear processos de retroalimentació amb la intel·ligència humana mútuament beneficiosos per a la resolució de problemes complexos, així com la dificultat d'identificar codis erronis en la programació. S'usa un nou operador de recombinació d'arcs de tres pares per resoldre el problema de rutes i vehicles amb restricció de capacitat, necessari per estimar els costos associats a les xarxes de distribució turística de baixa demanda. Els AG han demostrat una gran flexibilitat, especialment en entorns d'empresa canviants i en la solució de problemes de presa de decisions que involucren restriccions que són ambigües i, de vegades, contradictòries.

*MSC:* 90B20

*Paraules clau:* Xarxes de distribució, problema de rutes i vehicles, demanda turística, transport aeri, algoritmes genètics, operador de recombinació d'arcs

SORT 27 (1) January-June 2003, 113-136

# An empirical evaluation of small area estimators

Àlex Costa[a], Albert Satorra[b] and Eva Ventura[b] *

[a] *Institut d'Estadística de Catalunya (Idescat)*

[b] *Universitat Pompeu Fabra*

## Abstract

This paper compares five small area estimators. We use Monte Carlo simulation in the context of both artificial and real populations. In addition to the direct and indirect estimators, we consider the optimal composite estimator with population weights, and two composite estimators with estimated weights: one that assumes homogeneity of within area variance and squared bias and one that uses area-specific estimates of variance and squared bias. In the study with real population, we found that among the feasible estimators, the best choice is the one that uses area-specific estimates of variance and squared bias.

*MSC:* 62J07, 62J10, 62H12

*Keywords:* Regional statistics, small areas, root mean square error, direct, indirect and composite estimators

## 1 Introduction

Official statistics is faced with the need to generate estimates for small administrative units, while working with relatively small samples and within stringent budgetary limit. This conflict has been accentuated in recent years: on the one hand, politics is becoming more and more local, necessitating better local information; on the other hand, the public service nature of official statistics makes it more and more clear that producing quality work in this sphere involves not only optimising some theoretical parameters but also applying appropriate methodological strategies to achieve a positive cost/benefit relationship for society. Within this context, the vital nature and relevance that small

*Address for correspondence:* A. Costa. Via Laietana, 58. 08003 Barcelona, Spain. acosta@idescat.es (A. Costa), albert.satorra@econ.upf.es (A. Satorra) and eva.ventura@econ.upf.es (E. Ventura)

Received: December 2002
Accepted: January 2003

area statistics have had in the 1990s is understandable, as is the interest generated by official regional statistics.

There is a varied methodology on developing small area estimators. The reader can consult Platek *et al.* (1987), Isaki (1990), Ghosh and Rao (1994), and Singh, Gambino and Mantel (1994) to gain an overview of them. An initial classification divides the different existing methods into two categories: traditional and model-based. Traditional models include direct and indirect estimators and their combinations. Traditional direct estimators use only data from the small area being examined. Usually they are unbiased, but their exhibit a high degree of variation. Traditional indirect and model-based estimators are more precise since they also use observations from related or neighbouring areas. Indirect estimators are obtained through unbiased large area estimators. Based on them, it is possible to derive estimators for smaller areas under the assumption that they exhibit the same structure (with regard to the phenomenon being studied) as the initial large area. If this condition is not met, biased estimators could result. Traditional composite estimators are linear combinations of direct and indirect estimators. Model-based estimators can be interpreted as composite estimators, but unlike the traditional estimators, the weighting factors depend on the structure of the estimator's covariances. More information on this topic can be obtained from Cressie (1995), Datta *et al.* (1999), Farrell, MacGibbon and Tomberlin (1997), Ghosh and Rao (1994), Pfeffermann and Barnard (1991), Raghunathan (1993), Singh, Mantel and Thomas (1994), Singh, Stukel and Pfeffermann (1998), and Thomas, Longford and Rolph (1994).

In a previous study (see Costa, Satorra and Ventura, 2002), we began to examine these improved estimators, starting with a scenario in which we have two estimators, neither of which is entirely satisfactory:

- a direct estimator, obtained through the sample data pertaining to the small area; unbiased but in general not very precise.
- an indirect (synthetic) estimator, obtained through auxiliary information from other areas, periods or statistical sources; with smaller variance but generally biased.

Statistical theory of small area estimation proposes a way of combining both estimators in a linear fashion so that the resulting estimator represents a compromise between the absence of bias and minimal variance. The resulting composite estimator is the linear combination of the direct and indirect estimator that minimises mean squared error (MSE).

In our previous study, in which the autonomous regions in Spain are the small areas, the following results were obtained: 1) When the small area is centred and quite large (such as Catalonia), the composite estimator is as efficient as the indirect (or synthetic) one, in that it has a very low bias; 2) The composite estimator works well in general, especially in medium-sized and large areas. In our previous work, we wished to study in more detail the behaviour of composite estimates, but the information with which we were working, the National Statistical Institute (INE)'s Survey of the Active Population

(EPA), was a complex survey that made analysis difficult. For this reason, we decided to learn more about composite estimators in a simpler context in which we could carry out a Monte Carlo experiment. This is the objective of the present article.

Specifically, we decided that we needed to estimate the optimal weighting factors, not an easy task given that the variances and covariances of the estimators must themselves be estimated, as must their bias. For this reason we concentrated on a comparative analysis of the direct and indirect estimators with three composite estimators: one that has optimal weighting factors (theoretical), and two that use estimated weighting factors. One of the estimators based on estimated weighting factors uses the hypothesis of homogeneity of bias and variance for all the areas (this is the so-called *classic* composite estimator). The other estimates the area-specific biases and variances (this is the so-called *alternative* composite estimator). The characteristics of these estimators are studied in relation to the distribution of the mean squared error (MSE) and in a scenario with varied sample sizes.

## 2 The small area estimators

Consider the random variables $\hat{\theta}_j \sim N\left(\theta,\sigma_j^2\right)$ , $j=1,2,\ldots,J$ and $\hat{\theta}_* \sim N\left(\theta_*,\sigma_*^2\right)$, and $\gamma_j$ the covariance between $\hat{\theta}_j$ and $\hat{\theta}_*$. Our objective is to estimate $\theta_j$ using $\hat{\theta}_j$ and $\hat{\theta}_*$. It is well known that the best linear composite estimator of $\theta_j$ (in the sense of minimising the MSE) is

$$\tilde{\theta}_j = \pi_j \hat{\theta}_* + (1-\pi_j)\hat{\theta}_j$$

with

$$\pi_j = \frac{\sigma_j^2 - \gamma_j}{(\theta_j-\theta_*)^2 + \sigma_j^2 + \sigma_*^2 - 2\gamma_j}$$

For simplicity, assume that the covariance $\gamma_j = 0$ and the $\sigma_*^2$ is negligible. We also assume that $\theta_j \sim N\left(\theta_*,b_j^2\right)$. The value of $\pi_j$ that minimises the MSE is

$$\pi_j = \frac{\sigma_j^2}{\left(b_j^2+\sigma_j^2\right)} \qquad (1)$$

In practice, the values of the variance and bias are unknown (they are population-based parameters), and they must thus be estimated if we wish to approach the optimal value of $\pi_j$in o(1).

The quantity of interest for an area can be estimated "naively", using the sample mean of observations in the small area (direct estimator), or the mean of the observations of the entire population sample (indirect estimator). The direct estimator uses only the information on the area $j$ being examined, while the indirect estimator is based on the sample information gathered in all the areas. It is obvious that the direct estimator is unbiased for the mean of the area. Nevertheless, it has a high variance (given that if the area is small only few of the observations fall in this area). In contrast, the indirect estimator, based on the sample from the entire population, will have a low variance (given the large sample size), but it will suffer from bias when estimating the characteristics of a certain area, which will almost certainly differ from the common characteristics of the entire population.

An estimator that combines the qualities of the direct and indirect estimators with optimal weighting factors is the composite estimator that uses the value $\pi_j$ as specified in (1). This estimator constitutes a reference in our study, and it is called the **theoretical composite** estimator denoted by (*theor*). Nevertheless, this estimator is not feasible in practice because it depends on the weighting factor $\pi_j$, a value that in turn depends on unknown population parameters.

There are several procedures for estimating these population parameters, all of which lead to different small area estimators. In the present study, we investigate the **classic composite** estimator (*class*) and the **alternative composite** estimator (*altern*) which are described below:

### *2.1 Classic composite estimator*

The classic composite estimator assumes that the areas share the same within-area variance and a common estimate for the squared bias. Specifically, we assume components of variance specification $\hat{\theta}_j \sim N(\theta_j, \sigma^2)$ , $j = 1,2,\ldots,J$ with $\theta_j \sim N(\theta_*, b^2)$.

Here we use a weighted mean of the sample variances from each area as an estimate of the baseline data variance. Thus we define the pooled within variance

$$\bar{s}^2 = \frac{\sum_{j=1}^{J} (n_j - 1) s_j^2}{(n - J)} \tag{2}$$

in which $n$ is the size of the entire sample, $n_j$ is the sample size of the small area (in our real population example, the county) and $s_j^2$ is the sample variance of the baseline data of the small area $j$. Under the assumption that $\sigma_j^2 = \sigma^2$for all of $j$, the estimator of $\sigma_j^2$ is $\bar{s}^2/n_j$.

For the squared bias $(\theta_* - \theta_j)^2$, we define the common estimator

$$b^2 = \frac{1}{J} \sum_{j=1}^{J} \left(\hat{\theta}_j - \hat{\theta}_*\right)^2 \tag{3}$$

which is the mean squared difference of the direct and indirect estimators.

We could also have used a weighted mean of the individual biases; however, the properties of each bias estimator are somewhat different. Specifically, in the case in which we preferred to use the weighted mean of the individual biases, $b^2$ would be the estimator of a combination of variances *between* and *within* groups.

Thus, the estimator of $\pi_j$ is:

$$\hat{\pi}_j^c = \frac{\bar{s}^2/n_j}{\bar{s}^2/n_j + b^2}, \tag{4}$$

and the composite estimator obtained through the sample data is

$$\tilde{\theta}_j^c = \hat{\pi}_j^c \hat{\theta}_* + \left(1 - \hat{\pi}_j^c\right) \hat{\theta}_j \tag{5}$$

### 2.2 Alternative composite estimator

Another way of calculating the composite estimator uses direct estimators of each area's variance and bias. In this way the estimator of $\pi_j$ is:

$$\hat{\pi}_j^a = \frac{s_j^2/n_j}{\left(\hat{\theta}_j - \hat{\theta}_*\right)^2} \tag{6}$$

Note that $\left(\hat{\theta}_j - \hat{\theta}_*\right)^2$is biased for $(\theta_j - \theta_*)^2$, but is unbiased for $\sigma_j^2 + b_j^2$, as

$$\begin{aligned} E\left(\hat{\theta}_j - \hat{\theta}_*\right)^2 &= E\left(\hat{\theta}_j - \theta_j + \theta_j - \hat{\theta}_*\right)^2 = \\ &= E\left(\hat{\theta}_j - \theta_j\right)^2 + E\left(\theta_j - \hat{\theta}_*\right)^2 + 2E\left(\hat{\theta}_j - \theta_j\right)\left(\theta_j - \hat{\theta}_*\right) = \\ &= \sigma_j^2 + b_j^2 \end{aligned}$$

The composite estimator obtained through the sample data has the same form as (5), with $\hat{\pi}_j^a$ replacing $\hat{\pi}_j^c$;

$$\tilde{\theta}_j^a = \hat{\pi}_j^a \hat{\theta}_* + \left(1 - \hat{\pi}_j^a\right) \hat{\theta}_j \tag{7}$$

If necessary, the weight $\hat{\pi}_j^a$ is truncated to one.

Thus we consider five estimators: the direct $\hat{\theta}_j$, the indirect $\hat{\theta}_*$ (based on the entire sample), the theoretical composite $\tilde{\theta}_j^t$ based on the optimal weights in expression (1), the classic composite $\tilde{\theta}_j^c$ based on the weights of expression (4) and the alternative composite $\tilde{\theta}_j^a$ based on the weights of expression (6).

## 3 Monte Carlo study in an artificial population

In this section, we investigate the estimators defined in Section 2 using Monte Carlo methods. By generating artificial populations we explore the effect that different population characteristics have on the behaviour of the estimators. Some of the conclusions drawn in this section are validated through a Monte Carlo simulation in the context of a real population.

The artificial population is defined by the following components of variance model

$$x_{ij} = a + z_j + y_{ij} \qquad i = 1,2,\ldots,n_j \quad j = 1,2,\ldots,J$$

where $ij$ denotes individual $i$ in the area $j$, $n_j$ denotes the number of individuals in the sample of area $j$, and $J$ denotes the number of areas considered. In addition, assume that $z_j$ is distributed with a mean 0 and variance $b^2$, independent of $y_{ij}$, which has mean 0and variance $\sigma(x)_j^2$. The specific values of the model parameters and the characteristics of the design used in the study are $b^2 = 1$, $a = 10$ and $J$= 8, with identical sample sizes, $n_j = n^*$. The common sample size $n^*$ varies between 5 and 45. We consider two settings for $\sigma(x)_j^2$: i) common for all the areas, $\sigma(x)_j^2$= 30; and ii) values specific for each area; $\sigma(x)_j^2$ varies from 30 to 240.

We studied the effect of a change in the value of the within-area variance (for $b^2 = 1$). The variation of $\sigma(x)_j^2$ influences the value of the intra-class correlation *cci*, which varies between 0.05 and 0.30. The variation in the distribution, of $z_j$ (variation among groups) and $y_{ij}$ (variation within the area), is also considered. At the same time, we investigated different types of distribution within and between areas, as well as the total number of areas in the population.

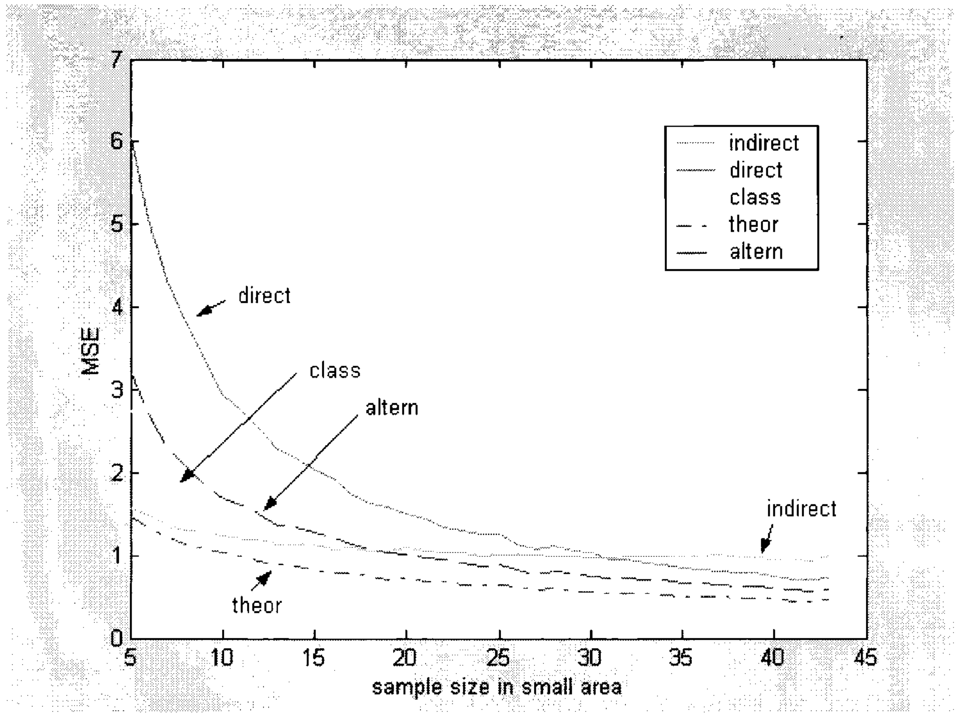

*Figure 1: MSE of the estimators as a function of the sample size n*: identical within-area variances.*

In summary, we considered the following factors:

*i)* Sample size $n^*$(small, medium or large)
*ii)* Coefficient of intra-class correlation
*iii)* Homogeneity of the variances within areas
*iv)* Distribution of the within and between variation

The number of Monte Carlo replications for each combination of factors considered is 6000. In all the cases, we estimated the mean parameter for a specific area. We assess the MSE when estimating this area parameter.

The results of the Monte Carlo study are discussed next. Figures 1-4 show the variation of the MSE of the different estimators analysed (*theor*, *class*, *altern*, *direct* and *indirect*) when we change either the sample size $n^*$ (Figures 1 to 3) or the magnitude of the intra-class correlation coefficient *cci* (Figure 4).

Figure 1 shows the results in the case of normality with identical within-area variances, when we vary the common sample size $n^*$. The population values used are $b^2 = 1$ and $\sigma(x)_j^2 = 30$, common to all the areas. From Figure 1 we conclude that the MSE is minimal for the *theor* estimator, maximal in the case of the *direct* estimator (except for large sample sizes, in which the MSE of the *direct* estimator can be less than that of the indirect estimator), and that the combined classic and alternative estimators have almost identical MSE, with an intermediate value between the *theor* and direct estimators. MSE are in a wide range for small sample sizes, but it is largely bridged as the sample size grows. The indirect estimator has a MSE greater than that of any of the other estimators for large sample size, but for small sample sizes it behaves similarly to the *theor* estimator. Nevertheless, the wide range of sample sizes for which the *indirect* estimator is better than the *direct* estimator should be noted.

Now consider the case in which there is heterogeneity in the within-area variance $\sigma(x)_j^2$. The results are presented in Figure 2. Here $\sigma(x)_j^2$ varies between 30 and 240 in the eight areas considered. We can consider the two extreme cases in which the area examined has variance 30 and 240, respectively.

This figure shows how the classic estimator improves considerably its performance with respect to the other alternative feasible estimators. Note that the MSE of *class* is, for all sample size, very close to the MSE of *theor*. For the smallest sample sizes, the *class* even improves slightly the performance of the *theor* estimator[1]. It is remarkable that the *altern,* which is mean to account for variation of within-area variance and square bias, performs slightly worst than *class* for all sample sizes considered

The results in Figure 2 correspond to the case in which the area examined has the smallest variances, 30, compared to the largest, 240. In Figure 3, we show the results for the complementary case when the variance of the area examined is 240.

[1] This is due to the fact that for small sample sizes and this non-identical within-area case, the variance of the indirect estimator is not negligible. The denominator of *class* of expression (4) has a positive bias that partially accounts for such variance.

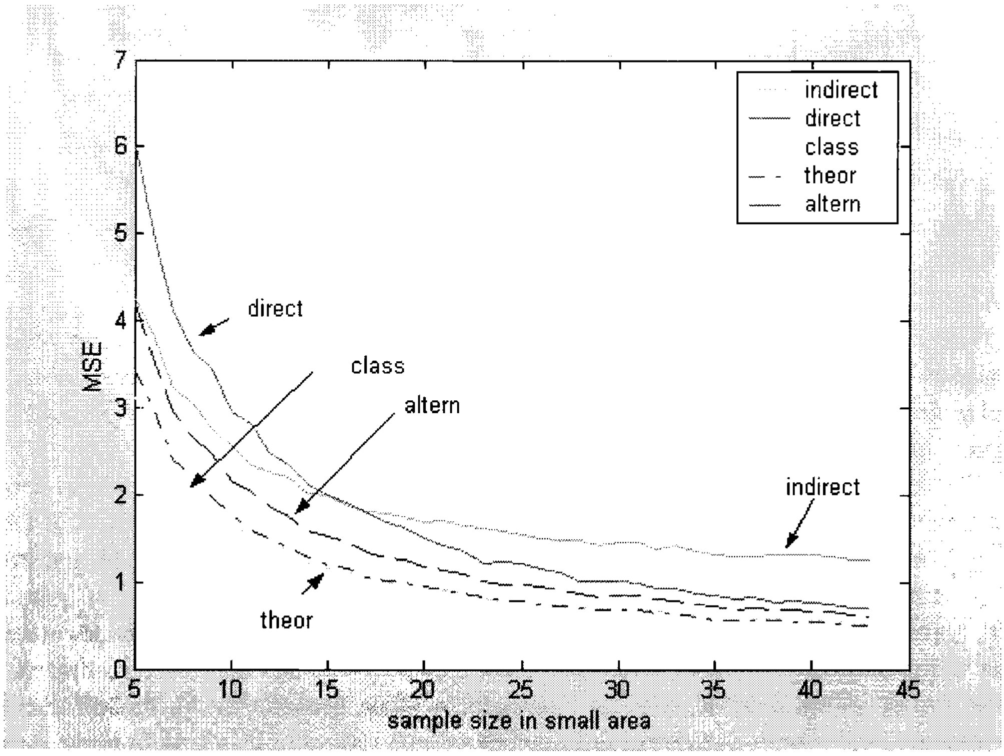

***Figure 2**: MSE of the estimators as a function of the sample size n*: non-identical within-area variances (area examined, the one with the smallest variance).*

In contrast to Figure 2, the indirect estimator exhibits behaviour quite similar to that of the optimal composite estimator (*theor*), while the two feasible composite estimators, *class* and *altern*, are close to each other.

One aspect of the population that could affect the behaviour of the different estimators is the ratio *r* of variance within the areas with respect to the total variation. The intra-class correlation coefficient *cci*, is equal to 1-*r*. Note that *cci* = 0 indicates that the entire variation comes from among the areas, while *cci*=1 indicates that the entire variation is within the areas. Figure 4 shows the variation in the MSE when we vary *cci* while maintaining the sample size constant. The population parameters used in the simulation are $b^2 = 1$ and$\sigma(x)_j^2$, constant for all the areas, with values that fluctuate between 2.5 (*cci* close to zero) and 50 (*cci* near 0.3). The sample size is $n^* = 10$ and the total number of areas is *J*= 20.

The MSEs of the different estimators now converge as *cci* (the "area effect") increases. The theoretical combined estimator (*theor*) outperforms the other estimators, even though its advantage over the direct estimator decreases toward zero as the *cci* increases. The MSE of the indirect estimator remains constant despite variation of the *cci*, and the classic composite estimator (*class*) outperforms the direct estimator and

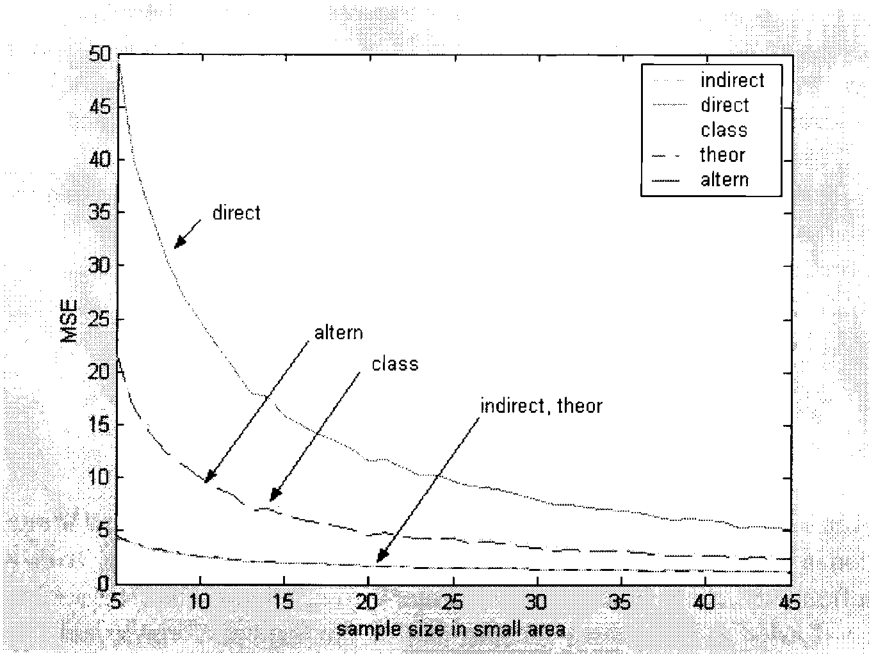

***Figure 3**: Variation of the MSE of the estimators as a function of the sample size n*, with heterogeneity in variances (area examined, the one with the greatest variance).*

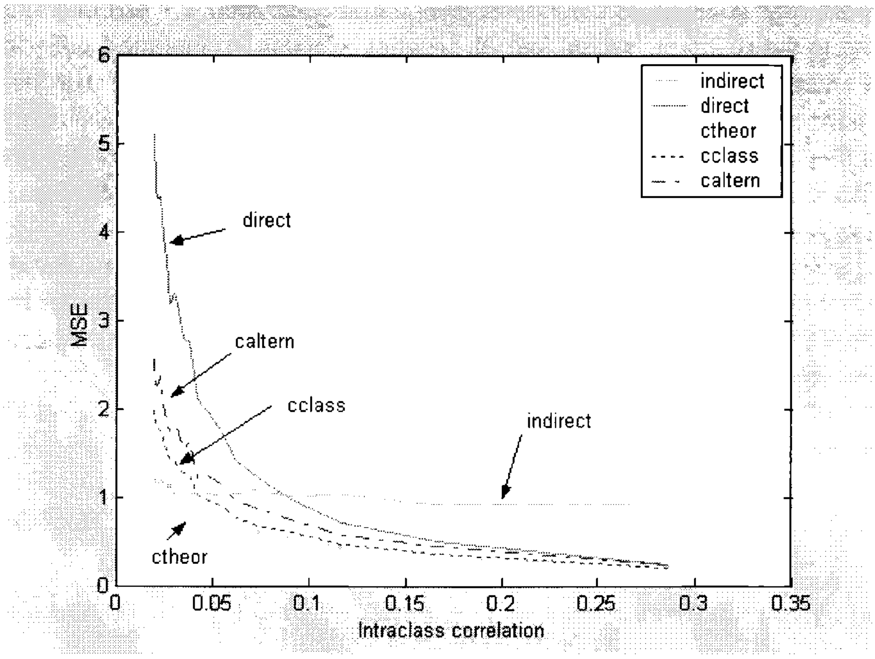

***Figure 4**: Variation in the MSE of the different estimators with regard to the intra-class correlation coefficient cci.*

converges with the theoretical composite (*theor*) as the intra-class correlation increases. The alternative and classic composite estimators exhibit similar behaviour, with a slight advantage for the classic. There is a substantial difference among the MSEs of the different estimators, with the indirect clearly outperforming the direct and the classic composite estimators when *cci* is small.

Without documenting further simulations in the same detail, we note that the same conclusions were obtained when: *a*) we violate the normality assumption of the within and between area variation (letting this distribution be a chi-squared distribution with 1 degree of freedom, i.e. a highly right skewed distribution); and, *c*) when we increase the value of $J$ of the number of areas of the population.

## 4 Simulation study on a real population

In this section we study the behaviour of the composite estimator through a Monte Carlo simulation in which we extract multiple samples from a known population. To do this, we use data from the Labour Force Census of Enterprises affiliated with the Social Security system in Catalonia. This census contains data on the number of employees from each enterprise surveyed who are registered with Social Security. The census was carried out in each of the four quarters between the years 1992 and 2000 (inclusive). We limit the analysis to one year, 2000.

This database contains 243,184 observations from year 2000, divided into 12 groups according to the economic sector, and 41 counties (Catalan "comarques"), the location of a few enterprises was not clarified, they have been excluded from this analysis.

We have eliminated the sector-based classification and have focused solely on the division by counties. Table 1 shows the number of enterprises per county and the mean and variance of individual affiliates per enterprise. The distribution of enterprises is quite uneven, as it is mainly concentrated in densely populated areas.

Next we present the results of the simulation for four sampling designs that differ in size, and compare the behaviour of the five estimators. The sizes of the samples are 10%, 5%, 1.68 % (this is the size used by Idescat in various surveys) and 1% of the population.

### 4.1 Design of the Monte Carlo simulation

Let $x_{j_k}$ be the number of salaried workers in county $j$ and enterprise $k$ . This is referred to as the *baseline data*. The total number of counties in Catalonia is $J$.

The parameters of interest are $\theta_j = \left(\sum_{k=1}^{N_j} x_{jk}\right) \Big/ N_j$, $j = 1,2,\dots J$, the mean number of salaried workers per enterprise in each county, $N_j$ are the numbers of surveyed enterprises in county $j$. With any sample we have a direct estimator $\hat{\theta}_j \sim N\left(\theta_j, var(\hat{\theta}_j)\right)$

*Table 1*: *Population values of area means, bias and variances.*

| County | Population size | $\theta_j$ | $(\theta_j - \theta_*)^2$ | $\sigma(x)_j^2$ |
|---|---|---|---|---|
| **Alt Camp** | 1282 | $8.73^a$ | 0.09 | 3250.37 |
| **Alt Empordà** | 4712 | 5.28 | 14.11 | 294.27 |
| **Alt Penedès** | 3052 | 8.91 | 0.02 | 1686.24 |
| **Alt Urgell** | 745 | 4.71 | 18.70 | 158.25 |
| **Alta Ribagorça** | 140 | 4.59 | 19.73 | 205.38 |
| **Anoia** | 3264 | 7.86 | 1.37 | 801.64 |
| **Bages** | 5698 | 8.24 | 0.63 | 1356.90 |
| **Baix Camp** | 5530 | 6.47 | 6.59 | 479.54 |
| **Baix Ebre** | 2237 | 6.31 | 7.41 | 534.40 |
| **Baix Empordà** | 4634 | 5.44 | 12.92 | 425.17 |
| **Baix Llobregat** | 20541 | 9.73 | 0.48 | 1642.46 |
| **Baix Penedès** | 2197 | 5.26 | 14.23 | 171.82 |
| **Barcelonès** | 88331 | 10.63 | 2.55 | 10314.88 |
| **Berguedà** | 1397 | 5.44 | 12.90 | 196.15 |
| **Cerdanya** | 788 | 3.71 | 28.34 | 71.93 |
| **Conca de Barberà** | 611 | 8.29 | 0.56 | 1388.95 |
| **Garraf** | 3466 | 6.28 | 7.62 | 685.91 |
| **Garrigues** | 516 | 5.24 | 14.42 | 96.89 |
| **Garrotxa** | 1909 | 7.51 | 2.33 | 419.72 |
| **Gironès** | 6369 | 9.82 | 0.62 | 2037.47 |
| **Maresme** | 11718 | 6.46 | 6.64 | 605.07 |
| **Montsià** | 1918 | 5.61 | 11.73 | 246.00 |
| **Noguera** | 1128 | 5.12 | 15.30 | 93.29 |
| **Osona** | 5494 | 7.09 | 3.77 | 774.65 |
| **Pallars Jussà** | 410 | 4.37 | 21.76 | 130.37 |
| **Pallars Sobirà** | 272 | 4.06 | 24.76 | 55.46 |
| **Pla d'Urgell** | 1106 | 6.59 | 5.95 | 271.85 |
| **Pla de l'Estany** | 1160 | 6.07 | 8.79 | 143.37 |
| **Priorat** | 254 | 4.11 | 24.26 | 180.17 |
| **Ribera d'Ebre** | 620 | 5.71 | 11.07 | 418.72 |
| **Ripollès** | 959 | 7.87 | 1.35 | 875.92 |
| **Segarra** | 594 | 10.87 | 3.35 | 8171.41 |
| **Segrià** | 7096 | 7.74 | 1.69 | 714.23 |
| **Selva** | 4586 | 7.11 | 3.70 | 610.20 |
| **Solsonès** | 508 | 5.58 | 11.93 | 157.58 |
| **Tarragonès** | 7440 | 9.42 | 0.15 | 1675.66 |
| **Terra Alta** | 297 | 4.25 | 22.87 | 40.28 |
| **Urgell** | 1178 | 6.28 | 7.59 | 312.25 |
| **Val d'Aran** | 503 | 5.28 | 14.08 | 270.11 |
| **Vallès Occidental** | 26683 | 10.34 | 1.71 | 3026.89 |
| **Vallès Oriental** | 11795 | 8.45 | 0.34 | 832.68 |

The mean number of affiliates in the whole of Catalonia, $\theta_*$, is 9.04.
The parameter *cci* is 0.0008.

*Table 2*: *Results of the simulation. Medium sample size (N = 24,295).*

| | | Sample distribution means | | | | Weights | | | Root mean square deviation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Sample size** | **Direct** | **Theoretical composite** | **Classic composite** | **Alternative composite** | **Theoretical** | **Classic estimate** | **Alternative estimate** | **Direct** | **Indirect** | **Theoretical composite** | **Classic composite** | **Alternative composite** |
| **Alt Camp** | 128 | 8.50 | 9.03 | 8.93 | 7.74 | 1.00 | 0.66 | 0.58 | 21.46 | 0.26 | 0.26 | 3.55 | 3.39 |
| **Alt Empordà** | 471 | 5.28 | 5.44 | 6.70 | 5.51 | 0.04 | 0.37 | 0.09 | 0.55 | 14.25 | 0.53 | 2.61 | 1.19 |
| **Alt Penedès** | 305 | 8.89 | 9.03 | 8.99 | 8.61 | 1.00 | 0.47 | 0.77 | 5.09 | 0.19 | 0.19 | 1.65 | 1.19 |
| **Alt Urgell** | 74 | 4.78 | 5.21 | 8.04 | 5.14 | 0.10 | 0.77 | 0.20 | 2.09 | 18.83 | 1.93 | 11.68 | 3.70 |
| **Alta Ribagorça** | 14 | 4.66 | 6.53 | 8.78 | 4.27 | 0.43 | 0.94 | 0.22 | 13.76 | 19.86 | 8.28 | 17.83 | 6.42 |
| **Anoia** | 326 | 7.76 | 8.58 | 8.37 | 8.08 | 0.64 | 0.45 | 0.63 | 2.13 | 1.53 | 0.85 | 1.03 | 1.52 |
| **Bages** | 569 | 8.27 | 8.87 | 8.55 | 8.33 | 0.79 | 0.33 | 0.66 | 2.30 | 0.79 | 0.62 | 1.24 | 0.94 |
| **Baix Camp** | 553 | 6.44 | 6.74 | 7.34 | 6.83 | 0.12 | 0.34 | 0.23 | 0.72 | 6.74 | 0.65 | 1.30 | 1.35 |
| **Baix Ebre** | 223 | 6.32 | 6.98 | 7.81 | 6.77 | 0.24 | 0.54 | 0.36 | 2.01 | 7.56 | 1.63 | 3.03 | 2.84 |
| **Baix Empordà** | 463 | 5.42 | 5.66 | 6.80 | 5.73 | 0.07 | 0.38 | 0.14 | 0.81 | 13.06 | 0.76 | 2.55 | 1.75 |
| **Baix Llobregat** | 2054 | 9.76 | 9.30 | 9.68 | 9.32 | 0.63 | 0.13 | 0.78 | 0.71 | 0.65 | 0.37 | 0.57 | 0.53 |
| **Baix Penedès** | 219 | 5.23 | 5.43 | 7.33 | 5.50 | 0.05 | 0.54 | 0.09 | 0.70 | 14.36 | 0.66 | 4.89 | 1.21 |
| **Barcelonès** | 8833 | 10.63 | 10.13 | 10.56 | 9.99 | 0.31 | 0.03 | 0.45 | 1.07 | 2.73 | 0.94 | 0.97 | 1.21 |
| **Berguedà** | 139 | 5.41 | 5.76 | 7.78 | 5.87 | 0.10 | 0.65 | 0.20 | 1.23 | 13.04 | 1.11 | 6.03 | 2.37 |
| **Cerdanya** | 78 | 3.73 | 3.90 | 7.76 | 3.95 | 0.03 | 0.76 | 0.06 | 0.88 | 28.46 | 0.86 | 17.04 | 1.59 |
| **Conca de Barberà** | 61 | 8.33 | 9.01 | 8.96 | 7.46 | 0.98 | 0.80 | 0.56 | 21.65 | 0.72 | 0.71 | 2.10 | 5.28 |
| **Garraf** | 346 | 6.25 | 6.82 | 7.50 | 6.84 | 0.21 | 0.44 | 0.38 | 1.66 | 7.76 | 1.36 | 2.34 | 2.98 |
| **Garrigues** | 51 | 5.22 | 5.66 | 8.36 | 5.78 | 0.11 | 0.82 | 0.24 | 1.70 | 14.55 | 1.51 | 10.16 | 3.21 |
| **Garrotxa** | 190 | 7.57 | 8.28 | 8.44 | 7.84 | 0.49 | 0.58 | 0.56 | 2.24 | 2.49 | 1.24 | 1.48 | 1.68 |
| **Gironès** | 636 | 9.85 | 9.16 | 9.62 | 9.22 | 0.84 | 0.31 | 0.84 | 2.87 | 0.80 | 0.65 | 1.51 | 1.18 |
| **Maresme** | 1171 | 6.47 | 6.65 | 7.00 | 6.70 | 0.07 | 0.20 | 0.15 | 0.47 | 6.79 | 0.45 | 0.71 | 0.84 |
| **Montsià** | 191 | 5.59 | 5.93 | 7.60 | 5.96 | 0.10 | 0.58 | 0.19 | 1.12 | 11.87 | 1.01 | 4.56 | 2.07 |
| **Noguera** | 112 | 5.09 | 5.30 | 7.83 | 5.40 | 0.05 | 0.69 | 0.09 | 0.70 | 15.44 | 0.66 | 7.86 | 1.11 |
| **Osona** | 549 | 7.09 | 7.62 | 7.77 | 7.47 | 0.27 | 0.34 | 0.42 | 1.18 | 3.93 | 0.93 | 1.16 | 1.71 |
| **Pallars Jussà** | 41 | 4.36 | 4.96 | 8.35 | 4.90 | 0.13 | 0.85 | 0.22 | 2.77 | 21.88 | 2.45 | 16.29 | 5.13 |
| **Pallars Sobirà** | 27 | 4.00 | 4.38 | 8.52 | 4.57 | 0.08 | 0.90 | 0.17 | 1.86 | 24.89 | 1.69 | 20.22 | 3.82 |
| **Pla d'Urgell** | 110 | 6.60 | 7.31 | 8.31 | 7.08 | 0.29 | 0.69 | 0.44 | 2.10 | 6.10 | 1.58 | 3.42 | 2.85 |
| **Pla de l'Estany** | 116 | 6.09 | 6.45 | 8.13 | 6.53 | 0.12 | 0.68 | 0.26 | 1.01 | 8.94 | 0.93 | 4.67 | 1.99 |
| **Priorat** | 25 | 4.16 | 5.27 | 8.57 | 4.00 | 0.23 | 0.90 | 0.12 | 6.92 | 24.38 | 5.48 | 20.25 | 3.48 |
| **Ribera d'Ebre** | 62 | 5.65 | 6.93 | 8.36 | 5.93 | 0.38 | 0.80 | 0.32 | 5.45 | 11.21 | 3.64 | 7.63 | 4.42 |
| **Ripollès** | 95 | 7.79 | 8.87 | 8.70 | 7.71 | 0.87 | 0.72 | 0.64 | 8.41 | 1.51 | 1.28 | 1.63 | 3.53 |
| **Segarra** | 59 | 10.88 | 9.07 | 10.04 | 6.89 | 0.98 | 0.80 | 0.43 | 129.05 | 3.54 | 3.46 | 17.53 | 23.75 |
| **Segrià** | 709 | 7.69 | 8.19 | 8.11 | 8.08 | 0.37 | 0.29 | 0.55 | 0.91 | 1.85 | 0.61 | 0.69 | 1.04 |
| **Selva** | 458 | 7.14 | 7.64 | 7.88 | 7.53 | 0.26 | 0.38 | 0.42 | 1.20 | 3.85 | 0.94 | 1.22 | 1.59 |
| **Solsonès** | 50 | 5.64 | 6.34 | 8.44 | 6.11 | 0.21 | 0.82 | 0.31 | 2.92 | 12.07 | 2.43 | 8.61 | 3.64 |
| **Tarragonès** | 744 | 9.46 | 9.06 | 9.38 | 9.00 | 0.94 | 0.28 | 0.83 | 2.04 | 0.32 | 0.30 | 1.14 | 0.93 |
| **Terra Alta** | 29 | 4.26 | 4.52 | 8.49 | 4.60 | 0.06 | 0.89 | 0.11 | 1.24 | 23.00 | 1.18 | 18.35 | 2.25 |
| **Urgell** | 117 | 6.28 | 6.99 | 8.17 | 6.64 | 0.26 | 0.68 | 0.33 | 2.35 | 7.74 | 1.82 | 4.18 | 2.62 |
| **Val d'Aran** | 50 | 5.30 | 6.34 | 8.41 | 5.51 | 0.28 | 0.83 | 0.27 | 4.96 | 14.22 | 3.71 | 10.25 | 4.53 |
| **Vallès Occidental** | 2668 | 10.31 | 9.80 | 10.20 | 9.76 | 0.40 | 0.10 | 0.63 | 0.97 | 1.90 | 0.72 | 0.82 | 1.09 |
| **Vallès Oriental** | 1179 | 8.43 | 8.83 | 57 | 8.62 | 0.67 | 0.2 | 0.71 | 0.60 | 0.51 | 0.29 | 0.43 | 0.43 |

***Table 3**: Results of the simulation. Small sample size (N = 12,059).*

| | | Sample distribution means | | | | Weights | | | Root mean square deviation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Sample size** | **Direct** | **Theoretical composite** | **Classic composite** | **Alternative composite** | **Theoretical** | **Classic estimate** | **Alternative estimate** | **Direct** | **Indirect** | **Theoretical composite** | **Classic composite** | **Alternative composite** |
| **Alt Camp** | 64 | 8.32 | 9.04 | 8.99 | 7.44 | 1.00 | 0.73 | 0.56 | 39.34 | 0.45 | 0.45 | 6.22 | 5.24 |
| **Alt Empordà** | 236 | 5.27 | 5.58 | 7.04 | 5.55 | 0.08 | 0.45 | 0.14 | 1.14 | 14.48 | 1.06 | 4.21 | 1.93 |
| **Alt Penedès** | 153 | 8.88 | 9.04 | 9.02 | 8.30 | 1.00 | 0.55 | 0.69 | 10.42 | 0.37 | 0.37 | 2.81 | 2.51 |
| **Alt Urgell** | 37 | 4.77 | 5.56 | 8.26 | 5.04 | 0.19 | 0.81 | 0.23 | 4.51 | 19.06 | 3.72 | 13.53 | 4.60 |
| **Alta Ribagorça** | 7 | 4.73 | 7.31 | 8.86 | 3.92 | 0.60 | 0.95 | 0.20 | 30.84 | 20.10 | 12.38 | 18.77 | 7.15 |
| **Anoia** | 163 | 7.67 | 8.74 | 8.45 | 7.76 | 0.78 | 0.54 | 0.58 | 4.30 | 1.73 | 1.19 | 1.67 | 2.22 |
| **Bages** | 285 | 8.29 | 8.95 | 8.66 | 8.17 | 0.88 | 0.41 | 0.64 | 5.19 | 0.98 | 0.86 | 2.26 | 1.84 |
| **Baix Camp** | 177 | 6.39 | 6.94 | 7.54 | 6.84 | 0.21 | 0.42 | 0.33 | 1.46 | 6.95 | 1.16 | 2.14 | 2.18 |
| **Baix Ebre** | 112 | 6.40 | 7.43 | 8.06 | 6.74 | 0.39 | 0.62 | 0.42 | 4.52 | 7.78 | 3.02 | 4.44 | 3.90 |
| **Baix Empordà** | 232 | 5.41 | 5.86 | 7.12 | 5.66 | 0.12 | 0.46 | 0.17 | 1.65 | 13.29 | 1.46 | 4.07 | 2.28 |
| **Baix Llobregat** | 1027 | 9.74 | 9.20 | 9.66 | 9.27 | 0.77 | 0.18 | 0.80 | 1.47 | 0.83 | 0.60 | 1.04 | 0.81 |
| **Baix Penedès** | 110 | 5.26 | 5.63 | 7.64 | 5.64 | 0.10 | 0.62 | 0.18 | 1.49 | 14.59 | 1.36 | 6.75 | 2.47 |
| **Barcelonès** | 4417 | 10.67 | 9.89 | 10.56 | 9.66 | 0.48 | 0.05 | 0.69 | 2.25 | 2.89 | 1.65 | 1.88 | 1.98 |
| **Berguedà** | 70 | 5.42 | 6.06 | 8.03 | 5.97 | 0.18 | 0.71 | 0.30 | 2.47 | 13.26 | 2.07 | 7.65 | 4.04 |
| **Cerdanya** | 39 | 3.72 | 4.05 | 8.04 | 4.06 | 0.06 | 0.81 | 0.12 | 1.76 | 28.71 | 1.67 | 19.70 | 3.43 |
| **Conca de Barberà** | 31 | 8.24 | 9.03 | 9.04 | 7.06 | 0.99 | 0.84 | 0.53 | 41.31 | 0.92 | 0.90 | 3.56 | 7.64 |
| **Garraf** | 173 | 6.30 | 7.23 | 7.79 | 6.67 | 0.34 | 0.52 | 0.40 | 3.55 | 7.98 | 2.50 | 3.74 | 3.70 |
| **Garrigues** | 26 | 5.25 | 6.03 | 8.52 | 5.88 | 0.21 | 0.86 | 0.35 | 3.57 | 14.78 | 2.88 | 11.42 | 5.25 |
| **Garrotxa** | 95 | 7.50 | 8.50 | 8.53 | 7.70 | 0.65 | 0.65 | 0.58 | 4.13 | 2.69 | 1.66 | 2.02 | 2.30 |
| **Gironès** | 318 | 9.86 | 9.11 | 9.59 | 9.05 | 0.91 | 0.39 | 0.80 | 6.60 | 0.97 | 0.88 | 2.85 | 2.61 |
| **Maresme** | 586 | 6.48 | 6.83 | 7.21 | 6.85 | 0.13 | 0.27 | 0.27 | 1.03 | 7.00 | 0.93 | 1.46 | 1.77 |
| **Montsià** | 96 | 5.59 | 6.21 | 7.87 | 6.01 | 0.18 | 0.65 | 0.27 | 2.43 | 12.10 | 2.02 | 6.20 | 3.31 |
| **Noguera** | 56 | 5.07 | 5.46 | 8.07 | 5.52 | 0.10 | 0.75 | 0.20 | 1.59 | 15.67 | 1.41 | 9.59 | 2.72 |
| **Osona** | 275 | 7.09 | 7.92 | 7.96 | 7.42 | 0.43 | 0.42 | 0.47 | 2.62 | 4.13 | 1.64 | 2.06 | 2.23 |
| **Pallars Jussà** | 21 | 4.39 | 5.42 | 8.49 | 4.57 | 0.22 | 0.88 | 0.20 | 6.02 | 22.13 | 4.77 | 17.76 | 4.78 |
| **Pallars Sobirà** | 14 | 3.97 | 4.67 | 8.61 | 4.58 | 0.14 | 0.91 | 0.25 | 3.73 | 25.14 | 3.15 | 21.34 | 6.46 |
| **Pla d'Urgell** | 55 | 6.64 | 7.73 | 8.48 | 6.99 | 0.45 | 0.75 | 0.49 | 4.93 | 6.32 | 2.83 | 4.40 | 3.84 |
| **Pla de l'Estany** | 58 | 6.05 | 6.71 | 8.31 | 6.67 | 0.22 | 0.74 | 0.39 | 2.20 | 9.16 | 1.75 | 5.76 | 3.42 |
| **Priorat** | 13 | 4.22 | 5.97 | 8.67 | 3.92 | 0.36 | 0.92 | 0.13 | 14.44 | 24.63 | 9.36 | 21.52 | 3.61 |
| **Ribera d'Ebre** | 31 | 5.64 | 7.51 | 8.49 | 5.73 | 0.55 | 0.84 | 0.34 | 11.62 | 11.44 | 5.77 | 8.86 | 5.75 |
| **Ripollès** | 48 | 7.83 | 8.95 | 8.84 | 7.27 | 0.93 | 0.77 | 0.57 | 17.61 | 1.71 | 1.57 | 2.67 | 6.43 |
| **Segarra** | 30 | 10.65 | 9.06 | 10.52 | 6.59 | 0.99 | 0.84 | 0.43 | 248.74 | 3.70 | 3.67 | 50.61 | 31.13 |
| **Segrià** | 355 | 7.66 | 8.41 | 8.22 | 8.00 | 0.54 | 0.37 | 0.60 | 1.85 | 2.05 | 0.98 | 1.27 | 1.58 |
| **Selva** | 229 | 7.15 | 7.94 | 8.08 | 7.53 | 0.42 | 0.46 | 0.50 | 2.61 | 4.06 | 1.61 | 2.10 | 2.25 |
| **Solsonès** | 25 | 5.62 | 6.80 | 8.57 | 5.93 | 0.35 | 0.86 | 0.34 | 6.60 | 12.30 | 4.39 | 9.70 | 5.02 |
| **Tarragonès** | 372 | 9.40 | 9.05 | 9.31 | 8.90 | 0.97 | 0.36 | 0.81 | 4.04 | 0.50 | 0.48 | 1.93 | 1.64 |
| **Terra Alta** | 15 | 4.30 | 4.79 | 8.61 | 4.92 | 0.11 | 0.91 | 0.23 | 2.65 | 23.24 | 2.41 | 19.64 | 5.08 |
| **Urgell** | 59 | 6.40 | 7.48 | 8.39 | 6.64 | 0.41 | 0.74 | 0.40 | 5.36 | 7.95 | 3.37 | 5.46 | 3.63 |
| **Val d'Aran** | 25 | 5.35 | 6.95 | 8.56 | 5.43 | 0.43 | 0.86 | 0.31 | 10.37 | 14.45 | 6.17 | 11.58 | 5.50 |
| **Vallès Occidental** | 1334 | 10.30 | 9.58 | 10.14 | 9.53 | 0.57 | 0.15 | 0.78 | 1.89 | 2.06 | 1.15 | 1.50 | 1.63 |
| **Vallès Oriental** | 590 | 8.42 | 8.92 | 8.64 | 8.52 | 0.81 | 0.27 | 0.70 | 1.31 | 0.70 | 0.50 | 0.85 | 0.73 |

*Table 4*: *Results of the simulation. Very small sample size (N = 4,100).*

| | | Sample distribution means | | | | Weights | | | Root mean square deviation | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | **Sample size** | **Direct** | **Theoretical composite** | **Classic composite** | **Alternative composite** | **Theoretical** | **Classic estimate** | **Alternative estimate** | **Direct** | **Indirect** | **Theoretical composite** | **Classic composite** | **Alternative composite** |
| **Alt Camp** | 22 | 8.49 | 9.05 | 9.40 | 6.84 | 1.00 | 0.78 | 0.53 | 132.58 | 1.20 | 1.20 | 37.69 | 9.07 |
| **Alt Empordà** | 79 | 5.39 | 6.16 | 7.51 | 5.60 | 0.21 | 0.54 | 0.24 | 3.85 | 15.35 | 3.23 | 7.64 | 3.66 |
| **Alt Penedès** | 51 | 8.98 | 9.05 | 9.19 | 7.92 | 1.00 | 0.63 | 0.66 | 34.77 | 1.11 | 1.11 | 8.35 | 6.22 |
| **Alt Urgell** | 13 | 4.68 | 6.40 | 8.39 | 4.84 | 0.39 | 0.84 | 0.27 | 13.25 | 19.95 | 7.97 | 15.72 | 6.17 |
| **Alta Ribagorça** | 2 | 5.03 | 8.41 | 9.08 | 3.35 | 0.84 | 0.97 | 0.19 | 134.57 | 21.00 | 18.81 | 23.45 | 9.89 |
| **Anoia** | 55 | 7.62 | 8.93 | 8.65 | 7.34 | 0.91 | 0.62 | 0.56 | 14.45 | 2.51 | 2.14 | 4.24 | 4.31 |
| **Bages** | 96 | 8.44 | 9.03 | 8.89 | 7.82 | 0.96 | 0.50 | 0.64 | 16.30 | 1.75 | 1.67 | 6.29 | 3.28 |
| **Baix Camp** | 93 | 6.52 | 7.63 | 7.92 | 6.70 | 0.44 | 0.51 | 0.42 | 4.67 | 7.78 | 3.09 | 4.53 | 3.49 |
| **Baix Ebre** | 38 | 6.33 | 8.11 | 8.29 | 6.36 | 0.65 | 0.69 | 0.44 | 14.19 | 8.61 | 5.54 | 7.22 | 5.29 |
| **Baix Empordà** | 78 | 5.51 | 6.56 | 7.55 | 5.64 | 0.30 | 0.54 | 0.24 | 5.68 | 14.15 | 4.22 | 7.66 | 3.14 |
| **Baix Llobregat** | 346 | 9.74 | 9.12 | 9.65 | 8.97 | 0.91 | 0.25 | 0.81 | 4.67 | 1.54 | 1.38 | 2.89 | 1.98 |
| **Baix Penedès** | 37 | 5.27 | 6.20 | 7.97 | 5.71 | 0.25 | 0.69 | 0.32 | 4.14 | 15.46 | 3.30 | 9.46 | 4.73 |
| **Barcelonès** | 1490 | 10.69 | 9.49 | 10.46 | 9.22 | 0.73 | 0.08 | 0.90 | 6.77 | 3.58 | 3.34 | 4.84 | 3.59 |
| **Berguedà** | 24 | 5.31 | 6.76 | 8.21 | 5.77 | 0.39 | 0.76 | 0.37 | 7.24 | 14.13 | 4.62 | 9.97 | 5.83 |
| **Cerdanya** | 13 | 3.70 | 4.57 | 8.25 | 3.87 | 0.16 | 0.84 | 0.14 | 5.08 | 29.63 | 4.36 | 22.75 | 4.20 |
| **Conca de Barberà** | 10 | 7.75 | 9.05 | 9.16 | 6.17 | 1.00 | 0.87 | 0.47 | 105.46 | 1.68 | 1.67 | 11.31 | 14.50 |
| **Garraf** | 58 | 6.25 | 7.96 | 8.08 | 5.99 | 0.61 | 0.61 | 0.33 | 10.89 | 8.82 | 4.92 | 6.72 | 4.34 |
| **Garrigues** | 9 | 5.23 | 6.87 | 8.64 | 5.47 | 0.43 | 0.88 | 0.38 | 9.86 | 15.65 | 6.03 | 13.24 | 7.51 |
| **Garrotxa** | 32 | 7.44 | 8.81 | 8.68 | 7.24 | 0.85 | 0.72 | 0.58 | 11.73 | 3.48 | 2.74 | 3.74 | 4.84 |
| **Gironès** | 107 | 9.85 | 9.08 | 9.63 | 8.63 | 0.97 | 0.48 | 0.75 | 21.02 | 1.68 | 1.62 | 7.42 | 5.34 |
| **Maresme** | 198 | 6.56 | 7.35 | 7.52 | 6.72 | 0.32 | 0.35 | 0.35 | 3.39 | 7.83 | 2.59 | 3.64 | 2.56 |
| **Montsià** | 32 | 5.66 | 7.00 | 8.15 | 5.86 | 0.40 | 0.72 | 0.35 | 7.31 | 12.96 | 4.84 | 8.83 | 4.95 |
| **Noguera** | 19 | 5.09 | 6.05 | 8.29 | 5.82 | 0.24 | 0.80 | 0.39 | 4.97 | 16.55 | 3.78 | 11.99 | 6.13 |
| **Osona** | 93 | 7.16 | 8.46 | 8.20 | 7.06 | 0.69 | 0.51 | 0.49 | 7.86 | 4.94 | 3.25 | 4.52 | 3.45 |
| **Pallars Jussà** | 7 | 4.20 | 6.44 | 8.60 | 4.52 | 0.46 | 0.90 | 0.28 | 15.43 | 23.03 | 9.00 | 19.66 | 7.57 |
| **Pallars Sobirà** | 5 | 4.06 | 5.61 | 8.70 | 4.20 | 0.31 | 0.92 | 0.26 | 11.54 | 26.05 | 8.02 | 23.04 | 8.03 |
| **Pla d'Urgell** | 19 | 6.65 | 8.35 | 8.63 | 6.34 | 0.71 | 0.80 | 0.46 | 13.77 | 7.14 | 4.84 | 6.27 | 6.31 |
| **Pla de l'Estany** | 20 | 6.11 | 7.43 | 8.51 | 6.34 | 0.45 | 0.79 | 0.45 | 7.11 | 10.00 | 4.24 | 7.85 | 5.50 |
| **Priorat** | 4 | 4.08 | 7.31 | 8.77 | 3.92 | 0.65 | 0.94 | 0.20 | 41.22 | 25.54 | 15.93 | 23.59 | 6.64 |
| **Ribera d'Ebre** | 10 | 5.54 | 8.32 | 8.67 | 5.13 | 0.79 | 0.87 | 0.35 | 36.97 | 12.29 | 9.32 | 11.60 | 8.24 |
| **Ripollès** | 16 | 7.66 | 9.02 | 8.94 | 6.28 | 0.98 | 0.82 | 0.47 | 54.19 | 2.49 | 2.40 | 6.01 | 11.37 |
| **Segarra** | 10 | 11.06 | 9.06 | 11.92 | 5.99 | 1.00 | 0.87 | 0.45 | 825.17 | 4.37 | 4.37 | 345.24 | 33.31 |
| **Segrià** | 120 | 7.72 | 8.76 | 8.43 | 7.72 | 0.78 | 0.45 | 0.61 | 5.97 | 2.83 | 2.06 | 3.31 | 3.08 |
| **Selva** | 77 | 7.18 | 8.46 | 8.32 | 7.20 | 0.68 | 0.55 | 0.52 | 7.93 | 4.87 | 3.13 | 4.34 | 3.83 |
| **Solsonès** | 9 | 5.57 | 7.64 | 8.67 | 5.56 | 0.59 | 0.88 | 0.38 | 18.29 | 13.16 | 7.68 | 11.34 | 7.02 |
| **Tarragonès** | 125 | 9.33 | 9.06 | 9.33 | 8.29 | 0.99 | 0.44 | 0.71 | 12.31 | 1.23 | 1.21 | 5.01 | 4.79 |
| **Terra Alta** | 5 | 4.35 | 5.57 | 8.71 | 4.73 | 0.26 | 0.92 | 0.32 | 8.88 | 24.15 | 6.69 | 21.40 | 8.06 |
| **Urgell** | 20 | 6.37 | 8.18 | 8.58 | 6.43 | 0.67 | 0.79 | 0.47 | 15.96 | 8.79 | 5.87 | 7.82 | 5.99 |
| **Val d'Aran** | 8 | 5.28 | 7.94 | 8.71 | 4.85 | 0.71 | 0.89 | 0.29 | 29.89 | 15.32 | 10.16 | 14.02 | 7.18 |
| **Vallès Occidental** | 450 | 10.27 | 9.30 | 10.09 | 9.26 | 0.80 | 0.21 | 0.80 | 6.13 | 2.76 | 2.25 | 4.20 | 3.26 |
| **Vallès Oriental** | 200 | 8.37 | 9.00 | 8.73 | 8.32 | 0.92 | 0.35 | 0.70 | 4.06 | 1.46 | 1.29 | 2.45 | 1.79 |

*Table 5*: *Results of the simulation. Very small sample size (N = 2,431).*

|  |  | Sample distribution means |  |  |  | Weights |  |  | Root mean square deviation |  |  |  |  |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
|  | **Sample size** | **Direct** | **Theoretical composite** | **Classic composite** | **Alternative composite** | **Theoretical** | **Classic estimate** | **Alternative estimate** | **Direct** | **Indirect** | **Theoretical composite** | **Classic composite** | **Alternative composite** |
| **Alt Camp** | 13 | 8.22 | 9.09 | 9.47 | 6.59 | 1.00 | 0.79 | 0.51 | 184.43 | 2.16 | 2.16 | 65.13 | 13.40 |
| **Alt Empordà** | 47 | 5.33 | 6.49 | 7.64 | 5.54 | 0.31 | 0.57 | 0.27 | 6.80 | 16.59 | 4.98 | 9.83 | 4.27 |
| **Alt Penedès** | 31 | 8.98 | 9.09 | 9.33 | 7.50 | 1.00 | 0.65 | 0.60 | 59.28 | 2.06 | 2.06 | 15.00 | 8.95 |
| **Alt Urgell** | 7 | 4.63 | 7.07 | 8.55 | 4.37 | 0.55 | 0.87 | 0.24 | 27.31 | 21.24 | 11.92 | 18.19 | 7.16 |
| **Alta Ribagorça** | 1 | 4.75 | 8.71 | 9.31 | 9.09 | 0.91 | 0.97 | 1.00 | 235.46 | 22.29 | 20.38 | 34.97 | 22.29 |
| **Anoia** | 33 | 7.63 | 9.02 | 8.74 | 7.00 | 0.95 | 0.64 | 0.51 | 23.69 | 3.55 | 3.23 | 6.39 | 6.53 |
| **Bages** | 57 | 8.43 | 9.08 | 8.97 | 7.57 | 0.97 | 0.53 | 0.61 | 24.59 | 2.75 | 2.66 | 8.35 | 4.94 |
| **Baix Camp** | 55 | 6.54 | 8.00 | 8.09 | 6.67 | 0.57 | 0.54 | 0.44 | 8.08 | 8.93 | 4.60 | 6.77 | 4.32 |
| **Baix Ebre** | 22 | 6.52 | 8.49 | 8.51 | 6.08 | 0.77 | 0.71 | 0.41 | 28.98 | 9.77 | 7.68 | 10.93 | 7.91 |
| **Baix Empordà** | 46 | 5.58 | 7.05 | 7.76 | 5.72 | 0.42 | 0.57 | 0.31 | 9.77 | 15.38 | 6.37 | 10.23 | 4.48 |
| **Baix Llobregat** | 205 | 9.76 | 9.13 | 9.70 | 8.93 | 0.94 | 0.28 | 0.77 | 8.80 | 2.43 | 2.27 | 4.95 | 3.12 |
| **Baix Penedès** | 22 | 5.24 | 6.60 | 8.10 | 5.64 | 0.35 | 0.71 | 0.34 | 6.24 | 16.71 | 4.62 | 11.36 | 5.37 |
| **Barcelonès** | 883 | 10.78 | 9.40 | 10.44 | 9.22 | 0.82 | 0.10 | 0.92 | 12.40 | 4.39 | 4.64 | 7.77 | 4.44 |
| **Berguedà** | 14 | 5.42 | 7.34 | 8.37 | 5.46 | 0.52 | 0.78 | 0.36 | 13.25 | 15.36 | 7.17 | 12.15 | 6.90 |
| **Cerdanya** | 8 | 3.75 | 5.03 | 8.35 | 3.87 | 0.24 | 0.85 | 0.16 | 8.90 | 31.00 | 7.04 | 24.76 | 4.93 |
| **Conca de Barberà** | 6 | 7.68 | 9.09 | 9.40 | 5.61 | 1.00 | 0.88 | 0.41 | 182.38 | 2.68 | 2.67 | 28.73 | 20.25 |
| **Garraf** | 35 | 6.15 | 8.27 | 8.18 | 5.83 | 0.72 | 0.63 | 0.33 | 18.47 | 9.98 | 6.54 | 9.83 | 5.87 |
| **Garrigues** | 5 | 5.31 | 7.48 | 8.74 | 5.09 | 0.57 | 0.89 | 0.34 | 19.47 | 16.90 | 9.02 | 15.11 | 8.51 |
| **Garrotxa** | 19 | 7.50 | 8.94 | 8.81 | 7.07 | 0.90 | 0.74 | 0.57 | 20.45 | 4.54 | 3.91 | 5.91 | 7.06 |
| **Gironès** | 64 | 9.94 | 9.11 | 9.67 | 8.37 | 0.98 | 0.51 | 0.70 | 36.91 | 2.56 | 2.51 | 11.68 | 8.21 |
| **Maresme** | 117 | 6.51 | 7.64 | 7.66 | 6.71 | 0.44 | 0.38 | 0.40 | 5.05 | 8.98 | 3.48 | 5.20 | 3.46 |
| **Montsià** | 19 | 5.66 | 7.46 | 8.30 | 5.81 | 0.52 | 0.74 | 0.38 | 12.11 | 14.17 | 6.72 | 11.10 | 6.30 |
| **Noguera** | 11 | 5.11 | 6.53 | 8.42 | 5.57 | 0.36 | 0.82 | 0.38 | 8.06 | 17.80 | 5.55 | 13.97 | 7.00 |
| **Osona** | 55 | 7.12 | 8.68 | 8.32 | 6.97 | 0.79 | 0.54 | 0.50 | 13.97 | 6.04 | 4.48 | 7.10 | 5.05 |
| **Pallars Jussà** | 4 | 4.24 | 7.15 | 8.72 | 4.04 | 0.60 | 0.91 | 0.25 | 26.85 | 24.34 | 12.63 | 22.06 | 8.56 |
| **Pallars Sobirà** | 3 | 4.17 | 6.28 | 8.79 | 3.95 | 0.43 | 0.93 | 0.24 | 22.26 | 27.39 | 12.54 | 25.07 | 10.03 |
| **Pla d'Urgell** | 11 | 6.60 | 8.61 | 8.77 | 6.25 | 0.81 | 0.82 | 0.47 | 22.90 | 8.28 | 6.30 | 8.58 | 7.49 |
| **Pla de l'Estany** | 12 | 6.17 | 7.85 | 8.62 | 6.26 | 0.58 | 0.80 | 0.46 | 12.61 | 11.18 | 6.16 | 9.67 | 6.87 |
| **Priorat** | 3 | 4.14 | 7.67 | 8.79 | 3.80 | 0.71 | 0.93 | 0.21 | 59.33 | 26.87 | 18.82 | 25.34 | 7.19 |
| **Ribera d'Ebre** | 6 | 5.40 | 8.59 | 8.74 | 4.77 | 0.86 | 0.88 | 0.31 | 56.59 | 13.50 | 11.14 | 14.75 | 9.97 |
| **Ripollès** | 10 | 7.91 | 9.08 | 9.15 | 5.87 | 0.98 | 0.83 | 0.42 | 93.34 | 3.52 | 3.45 | 11.40 | 15.94 |
| **Segarra** | 6 | 11.22 | 9.10 | 12.60 | 5.71 | 1.00 | 0.88 | 0.43 | 1278.82 | 5.16 | 5.15 | 701.02 | 36.85 |
| **Segrià** | 71 | 7.77 | 8.90 | 8.58 | 7.55 | 0.86 | 0.48 | 0.60 | 10.14 | 3.87 | 3.14 | 5.34 | 4.28 |
| **Selva** | 46 | 7.19 | 8.68 | 8.44 | 7.00 | 0.78 | 0.57 | 0.51 | 12.88 | 5.96 | 4.38 | 6.36 | 5.14 |
| **Solsonès** | 5 | 5.36 | 8.07 | 8.76 | 5.19 | 0.73 | 0.89 | 0.35 | 24.51 | 14.38 | 9.10 | 13.53 | 8.64 |
| **Tarragonès** | 74 | 9.37 | 9.10 | 9.39 | 8.21 | 0.99 | 0.48 | 0.71 | 20.14 | 2.13 | 2.12 | 7.74 | 6.64 |
| **Terra Alta** | 3 | 4.25 | 6.04 | 8.79 | 4.54 | 0.37 | 0.93 | 0.32 | 14.08 | 25.47 | 9.03 | 23.09 | 10.89 |
| **Urgell** | 12 | 6.40 | 8.49 | 8.67 | 6.11 | 0.77 | 0.80 | 0.45 | 25.50 | 9.95 | 7.45 | 9.61 | 6.99 |
| **Val d'Aran** | 5 | 5.19 | 8.29 | 8.83 | 4.63 | 0.79 | 0.89 | 0.28 | 49.26 | 16.56 | 12.34 | 17.78 | 8.46 |
| **Vallès Occidental** | 267 | 10.28 | 9.25 | 10.08 | 9.07 | 0.87 | 0.24 | 0.80 | 10.44 | 3.59 | 3.20 | 6.37 | 4.43 |
| **Vallès Oriental** | 118 | 8.40 | 9.06 | 8.84 | 8.15 | 0.95 | 0.38 | 0.67 | 6.91 | 2.44 | 2.26 | 4.23 | 2.62 |

for each county j and an indirect estimator $\hat{\theta}_* \sim N\left(\theta_*, var(\hat{\theta}_*)\right)$, which is common to all the counties.

If the variance of $x_{jk}$ is $\sigma(x)_j^2$, then $var(\hat{\theta}_j) = \sigma(x)_j^2 \big/ n_j$ , where $n_j$ is the number of sample observations in county $j$.

Our simulation exercise allows us to develop an optimal *theoretical composite* estimator, since we can evaluate expression (1).

We also evaluate a *classic composite* estimator and an *alternative composite* estimator as defined in Section 2.

We replicate 1,000 proportional samples from the enterprise census and apply the five estimators. The results are summarised in Tables 2 to 5.

### 4.2 Results of the simulation

Tables 2 through 5 summarise the results of the simulations for four scenarios. These scenarios differ in the sample size. In Table 2, the sample size is large, ***large*** sample size: precisely 24,295 observations in each total sample, which corresponds to 10% of the population. In Table 3, 5% of the population is sampled, resulting in 12,059 sample observations (***medium***-sized). The third sample represents slightly more than 1.68% of the population, yielding an average of 100 county observations (***small*** sized). However, the sample was extracted proportionally and the observations per county are distributed between a minimum of two in the county of *Alta Ribagorça* and a maximum of 1,490 in the county of *el Barcelonès*. The total number of observations from Catalonia is 4,100. Table 5 shows the fourth sample, which represents 1% of the population (***very small*** sized). The total number of observations is only 2,431.

To illustrate the form of the distribution of the MSE across counties, in Figures 5 we show the distributions of the MSEs for the four feasible estimators and two contrasting sample sizes: n =24295, a large sample; and n = 4100, a small sample. Overall we can say that these distributions of MSEs have the following common characteristics:

1. They are asymmetrical
2. They have extreme values (very noticeable in the case of the direct estimator)
3. They reveal a high degree of variation

These characteristics make it difficult to evaluate the different estimators based solely on their mean MSEs, especially given the presence of skew distributions and extreme values. For this reason we have decided to mix three comparison criteria, allowing us to make a more refined evaluation than just comparing simple means. These criteria are:

1. Comparison of the mean MSEs .
2. Comparison of the median of MSEs.
3. Comparison of the percentage of counties with lower MSEs (this criterion will be used for each pair of estimators)

***Table 6**: Statistics on the distribution of the MSEs for each estimator, by sample size.*

| **ESTIMATORS (n = 24,295)** | **direct** | **indirect** | **com teor** | çom clas | **com alt** |
|---|---|---|---|---|---|
| **mean** | 6.44 | 9.14 | 1.48 | 5.98 | 2.89 |
| **variance** | 399.01 | 64.14 | 2.33 | 39.22 | 12.88 |
| **average** | 1.86 | 7.56 | 0.94 | 3.03 | 1.99 |
| **Minimum value** | 0.47 | 0.19 | 0.19 | 0.43 | 0.43 |
| **Maximum value** | 129.05 | 28.46 | 8.28 | 20.25 | 23.75 |

| **ESTIMATORS (n = 12,059)** | **direct** | **indirect** | **com teor** | çom clas | **com alt** |
|---|---|---|---|---|---|
| **mean** | 12.82 | 9.35 | 2.48 | 7.98 | 4.16 |
| **variance** | 1,478.30 | 64.50 | 5.65 | 83.73 | 21.12 |
| **average** | 3.73 | 7.78 | 1.65 | 4.40 | 3.42 |
| **Minimum value** | 1.03 | 0.37 | 0.37 | 0.85 | 0.73 |
| **Maximum value** | 248.74 | 28.71 | 12.38 | 50.61 | 31.13 |

| **ESTIMATORS (n = 4,100)** | **direct** | **indirect** | **com teor** | çom clas | **com alt** |
|---|---|---|---|---|---|
| **mean** | 41.45 | 10.17 | 4.78 | 18.57 | 6.35 |
| **variance** | 16,316.69 | 65.45 | 13.88 | 2,723.66 | 24.62 |
| **average** | 11.54 | 8.61 | 3.78 | 7.82 | 5.34 |
| **Minimum value** | 3.39 | 1.11 | 1.11 | 2.45 | 1.79 |
| **Maximum value** | 825.17 | 29.63 | 18.81 | 345.24 | 33.31 |

| **ESTIMATORS (n = 2,431)** | **direct** | **indirect** | **com teor** | çom clas | **com alt** |
|---|---|---|---|---|---|
| **mean** | 66.38 | 11.29 | 6.48 | 30.91 | 8.33 |
| **variance** | 39,254.52 | 67.72 | 18.23 | 11,342.14 | 36.74 |
| **average** | 20.14 | 9.77 | 5.15 | 11.10 | 6.99 |
| **Minimum value** | 5.05 | 2.06 | 2.06 | 4.23 | 2.62 |
| **Maximum value** | 1,278.82 | 31.00 | 20.38 | 701.02 | 36.85 |

In Tables 6 and 7 the results of the synthesis can be seen, along with other complementary data, allowing the estimators to be evaluated. Based on the tables, we conclude:

1. For all sample sizes and for any of the three criteria used, the best estimator is the **theoretical composite** estimator. This result is as expected. Although not so important in practice, since this estimator is not accessible in real life applications. It is useful as a benchmark.
2. The best estimator among the four feasible ones is the **alternative composite**. For the four sample sizes and the three evaluation criteria (twelve combinations), the alternative composite estimator is better. The only exception to this is when we have a large sample size and we use the criterion of the counties with lowest MSE. In that case, the direct estimator is better than the alternative composite estimator.

*Table 7*: *Comparison of the estimators according to the criterion based on the percentage of counties with best MSE*$^2$.

| n=24,295 | direct | indirect | com teor | com clas | com alt |
|---|---|---|---|---|---|
| direct |  | 73.17 | 0.00 | 60.98 | 60.98 |
| indirect | 26.83 |  | 0.00 | 19.51 | 19.51 |
| com teor | 100.00 | 100.00 |  | 100.00 | 95.12 |
| com clas | 39.02 | 80.49 | 0.00 |  | 34.15 |
| com alt | 39.02 | 80.49 | 4.88 | 65.85 |  |

| n=12,059 | direct | indirect | com teor | com clas | com alt |
|---|---|---|---|---|---|
| direct |  | 65.85 | 0.00 | 48.78 | 36.59 |
| indirect | 34.15 |  | 0.00 | 24.39 | 24.39 |
| com teor | 100.00 | 100.00 |  | 100.00 | 90.24 |
| com clas | 51.22 | 75.61 | 0.00 |  | 26.83 |
| com alt | 63.41 | 75.61 | 9.76 | 73.17 |  |

| n=4,100 | direct | indirect | com teor | com clas | com alt |
|---|---|---|---|---|---|
| direct |  | 36.59 | 0.00 | 34.15 | 4.88 |
| indirect | 63.41 |  | 0.00 | 39.02 | 36.59 |
| com teor | 100.00 | 100.00 |  | 100.00 | 70.73 |
| com clas | 65.85 | 60.98 | 0.00 |  | 12.20 |
| com alt | 95.12 | 63.41 | 29.27 | 87.80 |  |

| n=2,431 | direct | indirect | com teor | com clas | com alt |
|---|---|---|---|---|---|
| direct |  | 26.19 | 0.00 | 19.05 | 0.00 |
| indirect | 73.81 |  | 7.14 | 52.38 | 38.10 |
| com teor | 100.00 | 92.86 |  | 97.62 | 54.76 |
| com clas | 80.95 | 47.62 | 2.38 |  | 7.14 |
| com alt | 100.00 | 61.90 | 45.24 | 92.86 |  |

Only if we grant this last criterion as much importance as the other two criteria, or more, can we say that for larger sample sizes the direct estimator is better. This specific advantage in one criterion disappears with the medium sample size (N = 12,059), so that in general the conclusion that the alternative composite estimator is better is warranted.

3. The **direct estimator** exhibits acceptable behaviour for the largest sample size, but its performance declines as sample size is reduced. In effect, for the large sample size (N =24,295), the direct estimator is the best according to the criterion of percentage, the second best according to the criteria of the average, and the third best according to the criteria of the mean MSE (it is surpassed by the two composite estimators, both classic and alternative). Its performance declines considerably in small samples since it is the second best according to the criterion of the average for medium-sized samples (N = 12,059), the third best according to the criteria of percentage and the worst according to the criteria of the mean of the MSEs. For small and very small samples, the direct estimator performs worse than any other estimator for all three criteria.
4. The **classic composite** is the one usually used in small areas. It is an estimator that always performs worse than the alternative composite, but it exhibits certain

interesting results in relation to the other estimators. In brief, for large sample sizes it is better than the indirect, while for small sample sizes it is better than the direct. For medium-sized samples it most likely obtains the best-combined results, since (if we keep aside the alternative estimator) it performs the best in both average and percentage. For small sample sizes it competes with the indirect, since for the small sample the indirect performs better on the average and percentage criteria, but worse on the MSE mean criterion. For the smallest sample size, it is clearly outperformed by the indirect estimator.

5. The last estimator to be examined is the **indirect estimator**, or the synthetic estimator. This "naive" estimator shows its qualities in small samples. Although it performs the worst of all the estimators for samples larger than 10,000, in the sample containing 4,100 it outperforms the direct estimator according to all three criteria used, and in the smallest sample it is the best estimator after the alternative composite.

## 5 Conclusions and research programme

The following general conclusions can be drawn from the Monte Carlo studies on artificial and real populations.

*a)* When the within-area variances are identical, the differences among the MSEs of the estimators examined are great when the (area) sample size is small, and tend to disappear in large sample sizes, although the indirect estimator shows a lesser degree of variation as a result of varying the sample size. Thus, there is a direct relationship between the sample size and convergence of the MSEs of the estimators.

*b)* When the within-area variances are identical, the differences among the MSEs of the direct, theoretical composite, classic composite and alternative composite estimators is large in the case of small intra-class correlation, but it disappears as the intra-class correlation increases. Thus, there is a direct relationship between the intra-class correlation and convergence of the MSEs of the estimators. An increase in the intra-class correlation does not lead to a reduction of the MSE in the indirect estimator.

*c)* As the sample size increases, the behaviour of the MSEs of the indirect estimator reflects a rate much lower than that of the other estimators, both in the improvements in its estimates and in its convergence with the rest. The greatest improvement when faced with increases in sample sizes is that of the direct estimator. The composite estimators have intermediate rates. Thus, each estimator has a different degree of sensitivity to increases in sample sizes.

*d)* There is a sample size below which the indirect estimator (or synthetic estimator), which uses information from all the areas, is the best alternative for

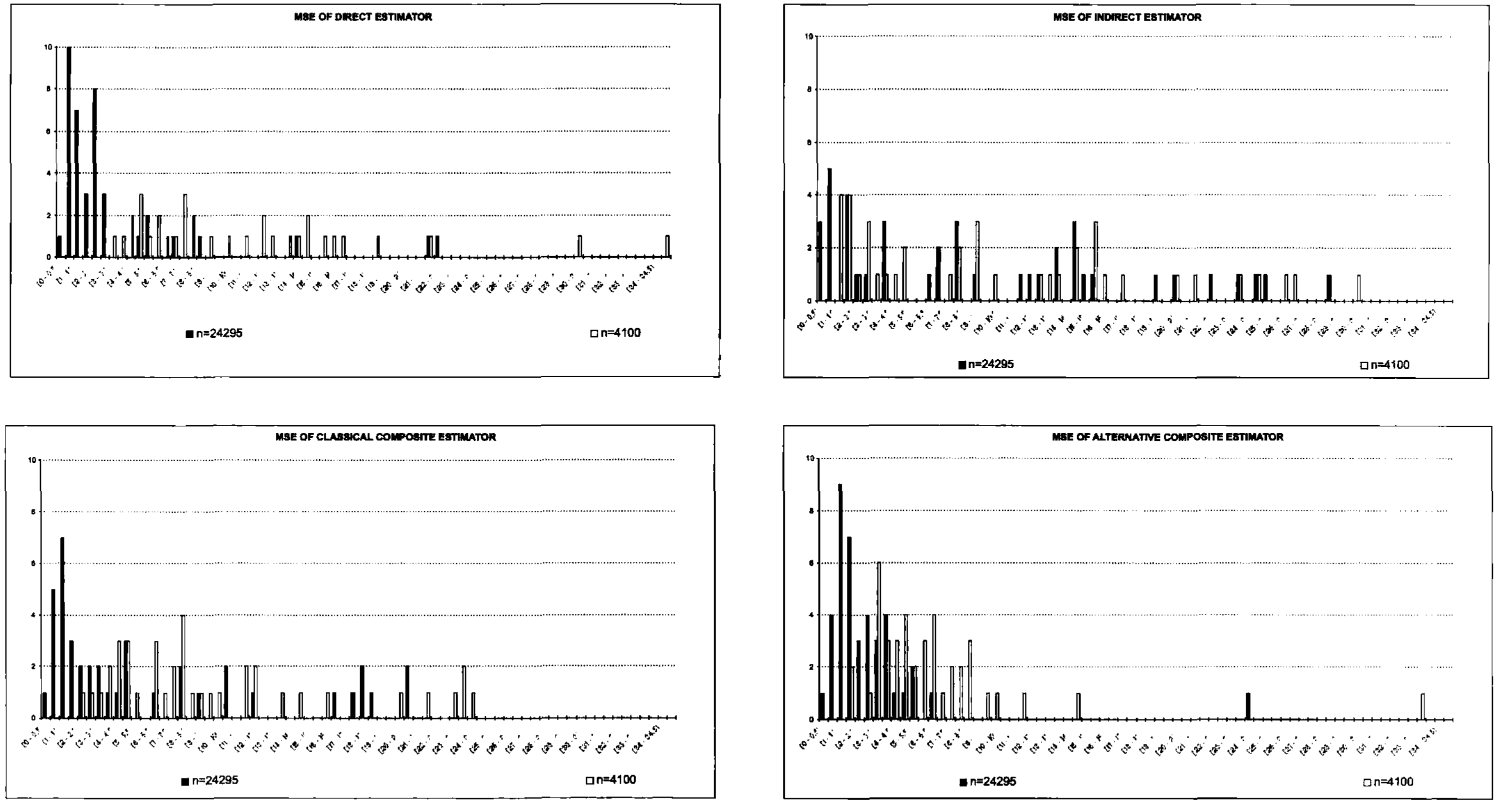

*Figure 5*: *MSE of the feasible estimators.*

estimating a parameter in a small area. In the real population examined in this study, below a certain sample size (specifically, the very small sample size) the best alternative to estimate the mean of a specific area, or the means of all the areas, is the indirect or synthetic estimator.

*e)* In the real population examined, the alternative composite estimator achieves the same degree of precision as a direct estimator with a sample size that is four times larger. In general, this estimator presents the best performance with regard to the MSEs for almost all the sample dimensions examined and for the different criteria applied.

*f)* For small or very small samples, in the empirical population studied, the direct estimator exhibits the worst performance with regard to the MSE. Thus, each of composite estimators considered performs better than the direct estimate.

As extensions to the present study, which constitute a research programme for the immediate future, we shall examine a series of points grouped in two different sections: theoretical developments and simulations, and applications:

### *Theoretical developments and simulations*

1. Estimates of inter-annual variation rates: we wish to replicate the evaluation of the five estimators studied when we examine the most important type of statistics for economic analysis, inter-annual variation rates. This extension could present surprising conclusions given the complexity of the variances of these indicators.
2. Analysis of the estimated weighting factors of the composite estimators: to better understand the comparative performance of the various composite estimator, it would be interesting to analyse in what way the weighting factor estimates are distributed compared to the theoretical weighting factors.
3. Sampling design for small area estimators: how should the sample size $n$ be allocated to the areas when small area estimation is considered? Answering this question through some theoretical development or through simulations is highly relevant to the practical work carried out by statistical organisations that need to provide information both at the area and country-level. In the initial phase it could be enlightening to compare proportional allocation (used here) with fixed and optimal classic allocation (depending on the variances in each stratum).

### *Practical applications*

1. County-level estimates of unemployment rates: in addition to their intrinsic interest for territorial economic analysis, these rates have at least three additional features: we can use sources we have already worked with and with which we are familiar, such as the INE's EPA; this is one of the surveys that has drawn the attention of

recent international literature on small data estimation (Datta *et al.*, 1999), and finally, we will soon have census data on county-wide unemployment when the 2001 census information is disseminated.

2. County-level estimates of the use of ICT (Information and Communication Technology): Idescat is currently researching this topic through a biannual survey undertaken since 2000, with samples slightly under 4,000 families. Currently, the Secretariat of the Information Society of the Generalitat de Catalunya (the sponsor of these surveys) has requested Idescat to generate a series of county-level estimates; this is therefore a natural point to begin applying small area estimators in official statistics.
3. Estimates of the IPI (Industrial Production Index) for Catalonia and its counties: IPI is a fundamental anchor in short-term economic analyses, and constitutes the first experience Idescat has had with small area estimation (the IPI for Catalonia), using a methodology that was later temporarily adopted by INE for all the autonomous regions within Spain. It is a case in which inter-annual variation rates are applied. In the future we will attempt to apply small area estimators to disaggregate the general IPI index provided by INE as of 2003 for Catalonia in two directions: by industrial sectors and by counties.

## Acknowledgements

The authors are grateful to the following statisticians from Institut d'Estadística de Catalunya: Xavier L6pez, Maribel García, Cristina Rovira and Antoni Contel for their help at several stages of this paper. We are also indebted to Nick T. Longford for detail comments on a previous version of this paper.

## 6 References

Clar, M., Ramos, R. and Suriñach, J. (2000). Avantatges i inconvenients de la metodologia del INE per elaborar indicadors de la producci6 industrial per a les regions espanyoles. *Qüestiió*, 24, 1, 151-186.

Costa, A. and Galter, J. (1994). L'IPPI, un indicador molt valu6s per mesurar l'activitat industrial catalana. *Revista d'Indústria*, 3, Generalitat de Catalunya, 6-15.

Costa, A., Satorra, A. and Ventura, E. (2002). Estimadores compuestos en estadística regional: aplicaci6n para la tasa de variaci6n de la ocupaci6n en la industria. *Qüestiió*, 26, 1-2, 213-243.

Cressie, N. (1995). Bayesian smoothing of rates in small geographic areas. *Journal of Regional Science*, 35 (4), 659-73.

Datta, G. S., *et al.* (1999). Hierarchical Bayes estimation of unemployment rates for the states of the U.S. *Journal of the American Statistical Association*, 94 (448), 1.074-82.

Farrell, P. J., Macgibbon, B. and Tomberlin, T. J. (1997). Empirical Bayes small-area estimation using logistic regression models and summary statistics. *Journal of Business & Economic Statistics*, 15 (1), 101-8.

Ghosh, M. and Rao, J. N. K. (1994). Small area estimation: an appraisal. *Statistical Science*, 9, 1, Statistics Canada, 55-93.

Isaki, C. T. (1990). Small-area estimation of economic statistics. *Journal of Business & Economic Statistics*, 8 (4), 435-41.

Longford, N. T. (2001). Synthetic estimators with moderating influence: the carry-over in cross-over trials revisited, *Statistics in Medicine*, 20, 3.189-3.203.

Pfeffermann, D. and Barnard, C. H. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business & Economic Statistics*, 9 (1), 73-84.

Platek, R., Rao, J. N. K., Särndal, C. E. and Singh, M. P. (eds.) (1987). *Small Area Statistics: an International Symposium*. New York: John Wiley and Sons.

Raghunathan, T. E. (1993). A quasi-empirical Bayes method for small area estimation. *Journal of the American Statistical Association*, 88 (424), 1444-48.

Singh, M. P., Gambino, J. and Mantel, H. J. (1994). Issues and strategies for small area data. *Survey Methodology*, 20, 1, Statistics Canada, 3-22.

Singh, A. C., Mantel, H. J. and Thomas, B. W. (1994). Time series EBLUPs for small areas using survey data. *Survey Methodology*, 20, 1, Statistics Canada, 33-43.

Singh, A. C., Stukel, D. M. and Pfeffermann, D. (1998). Bayesian versus frequentist measures of error in small area estimation. *Journal of the Royal Statistical Society*, b, 60, 377-396.

Thomas, N., Longford, N. T. and Rolph, J. E. (1994). Empirical Bayes methods for estimating hospital-specific mortality rates. *Statistics in Medicine*.

## Resum

Aquest paper compara cinc estimadors d'àrea petita. S'utilitzen mètodes de Monte Carlo primer en un context de població artificial i després en un context de població real. Juntament amb els estimadors directe i indirecte, es considera un estimador compost òptim amb pesos que són funció de valors poblacionals, i dos estimadors compostos amb pesos estimats: un que assumeix homogeneïtat de variàncies dintre àrees i biaix al quadrat, i un altre que considera estimadors específics de variància i biaix. En l'estudi basat en població real, s'observa que entre els estimadors factibles, el millor és aquell que empra estimadors específics de variància i biaix.

*MSC:* 62J07, 62J10, 62H12

*Paraules clau:* Estadística regional, àrea petita, error quadràtic mitjà, estimadors directe, indirecte i compost

# Information for authors and subscriptors

# Information for authors and subscriptors

## Submitting articles to SORT

### Guidelines for submitting articles

SORT accepts for publication only original articles that have not been submitted to any other journal in the areas of Statistics, Operations Research, Official Statistics and Biometrics. Articles may be either theoretical or applied, and may include computational or educational elements. Publication will be exclusively in English, with an abstract in Catalan (abstract translation into Catalan can be done by the journal's staff).

All articles submitted to thematic sections will be forwarded for systematic peer review by independent specialists and/or members of the Editorial Board, except for those articles specifically invited by the journal or reprinted with permission. Reviewer comments will be sent to the primary author if changes are requested in form or content.

When an original article is received, the journal sends the author a dated receipt; this date will appear in the publication as "reception date". The date in which its final version is received will be the "acceptance date".

To submit an article, the author must send it in PDF format to the electronic address of the journal (sort@idescat.es) especifying his e-mail address and/or a phone number. Once the article has been approved, the author must send a final electronic version, following the instructions of the editor responsible for that article. It is recommended that this final version be submitted using a $\LaTeX\,2_\varepsilon$.

In any case, upon request the journal secretary will provide authors with $\LaTeX\,2_\varepsilon$ templates and appropriate references to the MSC2000 classification of the American Mathematical Society. For more information about the criteria for using this classification system, authors may consult directly the Mathematics Subject Classification 2000 (MSC2000) at the website of the Institut d'Estadística de Catalunya (http://www.idescat.es/sort/Normes.stm).

### Publishing rights and authors' opinions

The publishing rights for SORT belong to the Institut d'Estadística de Catalunya since 1992 through express concession by the Technical University of Catalonia. The journal's current legal registry can be found under the reference number DL B-46.085-1977 and its ISSN is 1696-2281. Partial or total reproduction of this publication is expressly forbidden, as is any manual, mechanical, electronic or other type of storage or transmission without prior written authorisation from the Institut d'Estadística de Catalunya.

Authored articles represent the opinions of the authors; the journal does not necessarily agree with the opinions expressed in authored articles.

## Formal guidelines for articles

The first page of the article must contain the following items: title, name of the author(s), professional affiliation and complete address, and an abstract (100-200 words) followed by MSC2000 classification of the American Mathematical Society and the keywords. Before submitting an article, the author(s) would be well advised to ensure that the text uses correct English.

Bibliographic references have to be listed alphabetically at the end of the article, according to the following categories:

**Citations**
Mahalanobis (1936), Rao (1982b)

**Journal articles**
Pfeffermann, D. and Barnard, C. H. (1991). Some new estimators for small-area means with application to the assessment of farmland values. *Journal of Business & Economic Statistics*, 9 (1), 73-84.

**Books**
Platek, R., Rao, J. N. K., Sãrndal, C. E. and Singh, M. P. (eds.) (1987). *Small Area Statistics: an International Symposium*. New York: John Wiley and Sons.

**Parts of books**
Lee, E.W., Wei, L.J. and Amato, D.A. (1992). Cox-type regression analysis for large numbers of small groups of correlated failure time observations. In *Survival Analysis: State of the Art*, J.P. Klein and P.K. Goel (eds.), Kluwer Academic Publishers, 237-247.

**Web files or "pages"**
Nielsen, S. F. (2001). *Proper and improper multiple imputation*
http://www.stat.ku.dk/~feodor/publications/ (10th May 2003).

Multiple publications by a single author are to be listed chronologically. Explanatory notes should be numbered sequentially and placed at the bottom of the corresponding page. Tables and figures should also be numbered sequentially and will be reproduced directly from the submitted originals if it is impossible to include them in the electronic text.

**SORT** ***(Statistics and Operations Research Transactions)***
Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel.: +34-93 412 09 24 or +34-93 412 00 88
Fax: +34-93 412 31 45
E-mail: sort@idescat.es

## How to cite articles published in SORT

Commenges, D. (2003). Likelihood for interval-censored observations from multi-state models. *SORT*, 27 (1), 1-12.

## Subscription form
## SORT *(Statistics and Operations Research Transactions)*

Name ______________________________________________
______________________________________________

Organisation ______________________________________________
______________________________________________

Street Address ______________________________________________
______________________________________________

Zip/Postal code ____________________ City ____________________
State/Country ____________________ Tel. ____________________
Fax ______________ NIF/VAT Registration Number ______________
E-mail ______________________________________________
Date ______________________

Signature

I wish to subscribe to **SORT** ***(Statistics and Operations Research Transactions)*** for the year 2003 (volume 27)

Annual subscription rates:
– Spain: €22 (VAT included)
– Other countries: €25 (VAT included)

Price for individual issues (current and back issues):
– Spain: €9/issue (VAT included)
– Other countries: €11/issue (VAT included)

Method of payment:

□ Bank transfer to account number 2013-0100-53-0200698577

□ Automatic bank withdrawal from the following account number
□□□□ □□□□ □□ □□□□□□□□□□

□ Check made payable to the Institut d'Estadística de Catalunya

Please send this subscription form (or a photocopy) to:

**SORT** ***(Statistics and Operations Research Transactions)***
**Institut d'Estadística de Catalunya (Idescat)**
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

## Bank copy

Authorisation for automatic bank withdrawal in payment for
**SORT** ***(Statistics and Operations Research Transactions)***

The undersigned ______________________________________________

authorises Bank/Financial institution ______________________________

located at (Street Address) ____________________________________

Zip/postal code ______________________ City ______________________

Country _____________________________________________________

to draft the subscription to **SORT** ***(Statistics and Operations Research Transactions)*** from my account

number □□□□ □□□□ □□ □□□□□□□□□□

Date _________________________

Signature

**SORT** ***(Statistics and Operations Research Transactions)***
**Institut d'Estadística de Catalunya (Idescat)**
Via Laietana, 58
08003 Barcelona
SPAIN
Fax: +34-93-412 31 45

## Institut d'Estadística de Catalunya

The Institut d'Estadística de Catalunya (Idescat) is the statistical office of the Government of Catalonia. Its main duty is the management of the Catalan Statistical System by planning, coordinating and standardizing the statistical activity as well as providing statistical technical assistance.

## Secretary and subscriptions to SORT

Institut d'Estadística de Catalunya (Idescat)
Via Laietana, 58
08003 Barcelona (Spain)
Tel. +34 - 93 412 09 24 - 93 412 00 88
Fax. +34 - 93 412 31 45
E-mail: sort@idescat.es

**Publisher:** Institut d'Estadística de Catalunya (Idescat)

ISSN 1696-2281
DL B-46.085-1977
Key title: SORT
Numbering: 1 (december 1977)
http://www.idescat.es/sort

# SORT

Volume 27 (1), January - June 2003
Formerly Qüestiió

## Contents

ISSN: 1696-2281
http://www.idescat.es/sort

9 771696 228009 27002