European Journal of Population **13:** 223–241, 1997.

Survival Trees: An Alternative Non-Parametric Multivariate Technique for Life History Analysis

ALESSANDRA DE ROSE and ALESSANDRO PALLARA
ISTAT – Istituto Nazionale di Statistica, Rome, Italy (Address for correspondence: A. DE ROSE, ISTAT – Direzione Centrale Statistiche su Popolazione e Territorio, Via A. Ravà, 150, I-00142 ROMA, ITALY; Tel: ++39/6/549001; Fax: ++39/6/5943257)

Received 15 January 1997; accepted in final form 29 April 1997

De Rose, A. and Pallara, A., 1997, Survival Trees: An Alternative Non-Parametric Multivariate Technique for Life History Analysis.

Abstract. In this paper an extension of tree-structured methodology to cover censored survival analysis is discussed. Tree-based methods (also called recursive partitioning) provide a useful alternative to the classical survival data analysis techniques, such as the semi-parametric model of Cox, whenever the main purpose is defining groups of individuals, either with complete or censored life history, having different survival probability, based on the values of selected covariates. The essential feature of recursive partitioning is the construction of a decision rule in the form of a binary tree. Trees generally require fewer assumptions than classical methods and handle non standard and non linear data structures efficiently. Tree-growing methods make the processes of covariate selection and grouping of categories in event history models explicit. An example concerning the analysis of time to marriage of Italian women is presented.

De Rose, A. et Pallara, A. 1997. Des fonctions de séjour arborescentes: une méthode alternative, non paramétrique et multivariée pour l'analyse des biographies.

Résumé. Cet article propose une extension des méthodes utilisant des structures arborescentes pour réaliser une analyse des durées de séjour qui peuvent être interrompues. Les méthodes basées sur des arborescences fournissent une alternative aux méthodes d'analyses biographiques, telles que le modéle semi-paramétrique de Cox. Cela est possible toutes les fois que l'objectif principal est de définir des groupes d'individus d'après leurs biographies complètes ou interrompues, telles que l'événement a des probabilités différentes de survenir en fonction des variables sélectionnées. L'essentiel de ces méthodes est d'établir une règle de décision sous la forme d'un arbre de données binaires. De tels arbres peuvent être élaborés en faisant moins d'hypothèses que dans les méthodes classiques et permettent de traiter efficacement des données de structure plus complexe et non linéaire. Ces méthodes explicitent clairement le processus de sélection des variables et de regroupement des catégories, dans les modèles biographiques. Elles sont présentées avec un exemple d'application à l'âge au mariage des femmes italiennes.

1. Introduction: the discrimination problem in life history analysis[1]

One of the main goals of empirical applications in demography and social sciences is searching for factors which explain heterogeneity among individuals. Typically, one of the results that a social scientist expects from multivariate statistics technique

is the identification of distinct patterns of behaviour which are uniquely defined by a single or a combination of individual and/or background characteristics. For example, in previous work, the timing and intensity of marriage of Italian women have been showed to be highly differentiated with respect to birth cohort, level of education, professional status, urban/rural residence and geographic division (Blossfeld, De Rose, 1992; Castiglioni, Dalla Zuanna, 1994). As a result of this empirical evidence, a group of women can be considered as composed from a certain number of subsets according to characteristics which explain the heterogeneity of their marital behaviour: we expect that women who have a similar familial, socio-economic and cultural profile will also show a similar attitude toward marriage, and this pattern will be different from the one observed in any other subset of women.

The recent developments in the collection of individual biographies make the task even more ambitious. Hence, there is a great deal of information on life history together with the characteristics of a subject. It is possible to follow the evolution over time of a certain demographic behaviour and relate it to other life experiences as well as to fixed or time-varying covariates.

Modeling different paths of behaviour in relation to time and selected covariates is a very difficult task. A helpful contribution is obtained through the well known techniques of multivariate survival data analysis. The interest centres around a group of individuals for each of whom is defined an event (failure) occurring after a length of time, called the failure time. Usually an important concern is with the distribution of failure times in that group, but, more often, one wishes to compare the failure times in two or more groups to see, for example, whether the waiting time before marriage is systematically longer in one group than in an other. Groups are characterized by values of individual explanatory variables.

In this case, the main result from the widely used model for censored survival data analysis, either fully parametric or semi-parametric, is the estimation of the effect of selected covariates or their interactions on the individual "risk" of marrying. It is then possible to calculate a survival curve for any group of individuals with expected important differences in survival, corresponding to the persistence of the condition of not being married.

Whenever the main purpose of the analysis is defining groups of individuals with distinct patterns of behaviour, either with a complete or censored life history, the non-parametric approach proposed in this paper, based on recursive partitioning, is a useful alternative the classical survival data analysis. It provides a superior means for prognostic classification. The essential feature of the technique for recursive partitioning is the construction of a prediction rule in the form of a binary tree. Although the earlier developments of this approach goes back to the work on the AID (Automatic Interaction Detection) program in the early 1960's at the University of Michigan (Sonquist, Baker and Morgan, 1973) recursive partitioning methods have achieved widespread popularity in many fields of statistical data analysis only during the last decade, following the contribution of Breiman, Friedman, Olshen and Stone on the CART (Classification and Regression Trees)

methodology (Breiman et al., 1984). Tree-based methods, originally introduced as an alternative to parametric methods in discriminant analysis and linear regression, have been extended more recently to censored survival analysis (Gordon and Olshen, 1985; Segal, 1988; Ciampi and Thiffault, 1989; Davis and Anderson, 1989; Leblanc and Crowley, 1992). Trees generally require fewer assumptions than classical methods and handle efficiently non-standard and non-linear data structure, exploiting low local dimensionality of functions. Interactions are readily recognized and no problems arise in dealing with variables of continuous, discrete or mixed type, since no assumptions are imposed on the covariates' probability structure.

The final result is a decision rule which gives information which is easily understood and interpreted regarding the predictive structure of the data. Important inputs are quickly picked out and the number of subsets (groups) created is in some sense a minimum. The tree-shaped diagram is a very powerful way of showing the outcome of the analysis. It can be used to draw meaningful patterns of behaviour throughout the individual life history. For each group it is possible to obtain the survival function, hazard rate and relative risk estimates.

The paper is organised as follows. In the next section (2) the fundamentals of tree methodology are outlined; in Section 3 an application of the technique to real data from a survey on the progression to marriage among adult women in Italy is illustrated; in the final section (4) some comments are presented on the main advantages and problems related to tree-structured methodology for censored survival analysis.

2. Fundamentals of tree methodology

Consider a general random pair ($\mathbf{X}$, Y), having joint distribution F($\mathbf{x}$, y), where $\mathbf{X}$ is a set of *features* or predictor variables and Y denotes the '*outcome*' variable. The predictor variables can be both numerical (or ordinal, i.e. they take discrete or continuous values from an ordered set) and categorical (values not having any natural ordering). At this level of generality, Y can be deemed as the object of prediction, considered random. One usually wishes to obtain an estimate of a theoretical quantity which is a characteristic of the distribution of Y, depending on the particular setting. Hence, in the classical regression setting, Y denotes an (uncensored) response and the parameter of interest is the conditional expectation $h(\mathbf{x}) = E(Y|\mathbf{X} = \mathbf{x})$ of Y for any given realization of $\mathbf{X}$, i.e. the regression function. In the survival setting, the outcome of primary interest is duration of survival, time to failure or some other censored outcome and the parameter of interest is usually the hazard function of the survival distribution. The usual models for survival analysis assume that the characteristics of a subject (covariates) may influence the underlying hazard function that determines the survival distribution. The objective of a survival tree is to distinguish among classes of individuals in terms of the values taken by their covariates, so that each class is homogeneous in survival

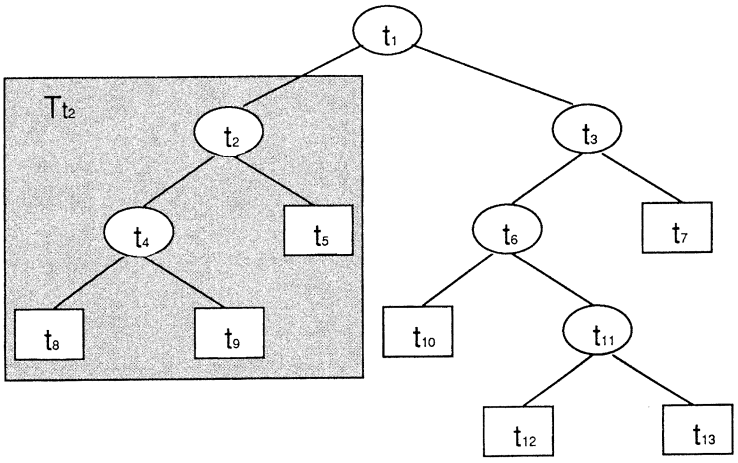

Figure 1. Example of a tree structure.

experience. The following discussion serves to introduce some notation and to specify the problem under consideration.

A (sample) regression tree is obtained by recursively partitioning a data set (training sample) consisting of N samples $(\mathbf{x}_i, y_i)$, i = 1, . . . , N from $(\mathbf{X}, Y)$ into subsets called nodes through a sequence of binary splits, which take the form of linear conditions on the levels of one or more feature variables. The splitting criterion creates subgroups of progressively increasing homogeneity with respect to the response.

In order to introduce the basic notion of a *binary tree*, in Figure 1 an illustrative example of such a structure is presented.

Using tree terminology (Breiman et al., 1984, sec. 10.1) one will distinguish between *non terminal nodes* (those enclosed in a circle – in Figure 1, t_1, i.e. the root node, t_2, t_3, t_4, t_6 and t_{11} . . .), namely nodes which have their *direct* descendant and *terminal nodes* (comprised in a rectangle), i.e. nodes which are no longer partitioned. A *branch* T_t of a tree T consists of the node t and all descendants of t in T. In Figure 1 the branch T_{t2}, e.g., comprises the set of nodes in the shaded area. A *pruning* operation on a tree T consists of declaring an internal (non terminal) node t terminal and deleting all descendants of t. A (pruned) *subtree* T' of T is got from T by one or more pruning operations.

A split s partitions node t into a left node t_L and a right node t_R. The set of candidate splits is obtained from the answer to questions such as "Is $x_m \leq c$?" for c any real number (univariate splits), "Is $\sum_m a_m x_m \leq c$?" (linear combination splits) and "Does $x_m \in B'$?" (if x_m is a categorical variable), where B' ranges over all disjoint subsets of $B = \{b_1, b_2, \ldots, b_L\}$, i.e. the set of different values of the categorical variable.

In practice, since the construction of the classifier is based on a sample, there is a finite number of values, at most N, that each x_m can take; so there are at most N-1 different splits given by $\{\text{Is } x_m \leq c_n?\}$, where the c_n's are the midpoints between consecutive distinct observed values of x_m. For a categorical variable x_m, if x_m takes on L distinct values, then 2^{L-1} splits are defined on the values of the variable.

At each step in tree-growing procedure, the splitting criterion selects among all possible splits the one which optimises some measure of dispersion in the response distribution, resulting in smaller and smaller subsets of data showing progressively increasing homogeneity. A numerical criterion is then needed to evaluate *goodness-of-split*. In what follows the least squares criterion discussed in CART monograph is briefly presented. Extensions proposed to cover situations in which the response variable is subject to censoring will be discussed later.

Let N(t) denote the number of samples in node t. The predicted value of the response variable Y in node t is

$$\overline{y}(t) = \frac{1}{N(t)} \sum_{x_i \in t} y_i$$

i.e. the average value within that node. The sum of squares in node t is

$$SS(t) = \sum_{x_i \in t} [y_i - \overline{y}(t)]^2$$

The least squares criterion for choosing splits is

$$f(s, t) = SS(t) - SS(t_L) - SS(t_R)$$

Given a class S of splits the optimal split s^* is defined by

$$f(s^*, t) = \max_{s \in S} f(S, t)$$

where the maximum is taken over all admissible splits.

Thus a regression tree partitions the sample space into M-dimensional rectangles using hyperplanes of the form $x_m = c_n$, where M is the number of predictor variables.

The sequence of splits generates a set of admissible trees having an increasing number of terminal nodes, $\{T_1, T_2, \ldots, T_N\}$ say, in which the subscript denotes the number of terminal nodes in each tree. As the number of terminal nodes in a tree increases, both the number of observations and the sample variance of the response variable are decreasing or at least not increasing, until the extreme case of a very large tree T_N with one observation in each terminal node and zero variance. However, usually a 'stop splitting' rule, involving a minimum size of a node >1, will arrest the process of splitting.

Indeed, a tree with too large a number of terminal nodes may result in a sample-dependent prediction rule. Therefore, some criterion is needed to determine the 'right-sized' tree. The procedure proposed in CART is called *minimal*

cost-complexity pruning. The basic idea of minimal cost-complexity pruning is to penalize trees for having large numbers of terminal nodes. The penalty per terminal node is represented by a real non-negative number α and the error complexity measure of tree T is defined as:

$$R_\alpha(T) = R(T) + \alpha \mid \tilde{\tilde{T}} \mid, \tag{1}$$

where R(T) is an estimate (resubstitution estimate) of the unknown expected mean square prediction error of tree T, $R^*(T)$ say, and $\mid \tilde{\tilde{T}} \mid$ is the number of terminal nodes of T. In the context of least squares regression R(T) is simply the sum over all terminal nodes of the within node sum of squares averaged by N, i.e.

$$R(T) = \frac{1}{N} \sum_{t \in T} \sum_{x_i \in t} [y_i - \bar{y}(t)]^2 \tag{2}$$

Then define $T(\alpha)$ as the smallest subtree of T_{max} – where T_{max} denotes the large tree obtained from the process of splitting downward – such that:

$$R_\alpha[T(\alpha)] = \min_{T \leq Tmax} R_\alpha(T)$$

(the notation $T \leq T_{max}$ indicates that T is a pruned subtree of T_{max}). The cost-complexity measure introduces a trade-off between the size of the tree and how well it fits the data. With $\alpha = 0$, T_{max} is the minimiser of $R_0(T)$. As the penalty per terminal node increases, the number of terminal nodes in the subtree that minimises $R_\alpha(T)$ decreases, finally reducing to the root node. Breiman et al. (1984) show that for every value of α there exists a unique smallest minimising subtree $T(\alpha)$ and that the process of pruning upward yields a nested sequence of trees with a decreasing number of terminal nodes corresponding to an increasing sequence of values of the penalty per terminal node. The optimal tree in the sequence of minimal cost-complexity trees is the tree with the lowest estimate of the expected risk $R^*(T)$. Selecting the optimal tree, therefore, requires an accurate estimate of the error sum of squares for each tree. Using resubstitution estimate is likely to give an overoptimistic estimate of $R^*(T)$, because the same data are used both for tree-growing and for evaluating pruning process. More accurate estimates can be obtained through a genuine sample of observations, assumed as being generated from the same joint distribution as the training sample but not included in tree-growing process (test sample), or using some resampling method such as cross-validation (Breiman et al., 1984).

The above discussion illustrates the essential aspects of regression tree methodology. For a thorough understanding of recursive partitioning, the definitive reference of Breiman et al. (1984) on the CART method is recommended. For a review of several aspects of CART, including a discussion of some developments of tree methodology, see Pallara (1992).

Several extensions of tree-based methods in the survival setting have been proposed. They represent an interesting alternative to traditional methods for the analysis of censored survival data, such as the linear proportional hazards model of Cox (1972). Methods for tree construction in the survival data setting differ (among each other and with respect to existing techniques for uncensored responses) in the way they pursue splitting and for the criterion for selecting tree size.

Gordon and Olshen (1985) first attempted to adapt the CART algorithm to the censored data problem. At any given node, the proposed splitting method uses distance measures between estimated Kaplan-Meier survival functions in candidate left and right daughter nodes. As for the optimal tree selection they use an analogous of least squares criterion (2) to introduce a within-node estimate of error resulting from the minimisation of the survival function with a finite mass point, therefore allowing for an extension to censored setting of the cost-complexity measure defined by (1). Segal (1988) adapted the CART tree-growing strategy in the survival data setting, by replacing goodness-of-split criterion with measures of node separation based on rank statistics. This involves substituting minimal cost-complexity pruning with a new pruning schema, whereas selection of the final tree is carried out by means of an *ad hoc* procedure, since cross-validation is not applicable to the sequence of subtrees obtained through the revised pruning algorithm. Davis and Anderson (1989) describe a CART-like algorithm for survival analysis which uses a partitioning criterion based on exponential log-likelihood loss. With this splitting criterion the value assigned to a node is the estimate of the hazard within the node. CART tree pruning and final selection strategy is incorporated in the proposed method. The modifications required include the increase of the minimum terminal node size of the maximal tree, an alternative estimate of the hazard in cross-validation to prevent zero estimates in some node and a procedure for final tree selection based on a pseudo-chi-square test statistic. A simulation experiment shows good results in terms of the proposed performance measures. Leblanc and Crowley (1992) propose a tree growing and pruning method for censored survival data which uses the first step of a full-likelihood estimation of the proportional hazards model. Final tree selection is obtained by cross-validation, choosing the tree which minimises, among all pruned subtrees, expected one-step deviance between the log-likelihood for the saturated model (a model that allows one parameter for each observation) and the maximised log-likelihood when the baseline cumulative hazard function is known. In a simulation experiment, this method shows similar performances as the method of Davis and Anderson (1989), also in the case of exponential survival times, when the latter is expected to perform the best.

A major development in tree-structured survival analysis has been represented by RECPAM (RECursive Partition and AMalgamation) approach (Ciampi et al., 1988; Ciampi and Thiffault, 1989; Ciampi, 1991). Tree growing in RECPAM is seen as a general strategy to predict a parameter (called *criterion*) of a statistical model on the basis of a set of predictor variables. The criterion may be the hazard function

of a survival distribution but it may also be the probability of class membership, as it is in a discriminant analysis problem or a conditional expectation, as in ordinary regression. The distribution may not be completely specified by the criterion, i.e. allowance is made for the presence of nuisance parameters in the specification of the distribution.

RECPAM tree construction is made up of three separate steps. The first two closely parallel the CART tree-growing and pruning process. The splitting rule is based on the construction of the Likelihood Ratio Statistic (LRS) of the hypothesis that the two subpopulations differ in terms of the estimated criterion versus the hypothesis that they do not. The statistical models allowed include the exponential, the multinomial and the Cox model. A distinctive feature of RECPAM methodology is the *amalgamation* step. The amalgamation algorithm moves from the set of terminal nodes of the most honest tree obtained by pruning, recombining nodes issuing from different parents in such a way that the resulting clusters are not only homogeneous but also distinct as regards the prediction of the criterion. Selection of a tree from the pruned and the amalgamated sequence is made by means of either a simultaneous testing procedure based on a conservative significance level associated to each partition of the sequence, or by choosing the partition with minimum Akaike Information Criterion (AIC) (Ciampi and Thiffault, 1989). The choice of AIC as a selection criterion represents a computational short-cut with respect to cross-validation, justified by asymptotic equivalence of minimum AIC and *leave-one-out* cross-validation.

Recently, some proposals have emerged (Loh, 1991; Ahn and Loh, 1994; Ahn, 1996), aiming to use a number of features of recursive partitioning methods for fitting traditional parametric and semi-parametric models to survival data. The proposed technique estimates at each node a regression model with some distribution for survival time and uses a regression tree based on the residuals of this regression for checking goodness-of-fit of the regression model. The method proceeds to build a kind of local regressions, much in the same way as in Ciampi (1991), fitting models to homogeneous subgroups of population, each characterized by a significant effect for selected covariates.

3. An empirical application: Progression to marriage

In this section the main features and practical use of the recursive partitioning methods with life history data are illustrated with an example of a two-state survival analysis of data on time to marriage among Italian women.

The sample of women has been obtained from the 1988 Multipurpose Survey on Households and Families, carried out at Italian National Statistical Institute (ISTAT). In what follows a subset of the original data set has been examined, including somewhat less than 1500 observations, with a percentage of censoring equal to 10%. The duration variable is the number of years between the 16th birthday and marriage. The outcome variables are the survival time and the censoring

indicator; the criterion is the hazards function. The statistical model applied is the proportional hazards model proposed by Cox (1972). The following individual characteristics are included as covariates:

BIRTH COHORT	8 five-year classes starting from 1924
WORKING STATUS	1 working at time of interview, 2 housewife, 3 retired
OCCUPATIONAL POSITION (current or past)	0 never worked, 1 manager, 2 white collar, 3 skilled worker, 4 unskilled worker, 5 entrepreneur, 6 self employed
EDUCATION	1 no qualification, 2 primary school, 3 middle school, 4 high school, 5 university degree
AREA OF RESIDENCE	1 North-West, 2 North-East, 3 Centre, 4 South, 5 Islands

It is important to note that the variables selected are measured at the time of interview. Indeed, the survey is designed to be cross-sectional, giving retrospective information only on demographic history of women aged 15 through 64, but no time-dependent data on such aspects as education and professional career. This could lead to incorrect interpretations of the results, unless one attributes to the above covariates the role of *proxies* for attitude and cultural inclination toward family and marriage. If, e.g., a woman was working at time of the interview, this is taken as an indication of a preference for a professional career, which is assumed constant throughout her entire life history. The effect of this stated preference on the propensity to marry will also be assumed to be unchanging.

Figure 2 shows the *honest tree* resulting from the application of RECPAM procedure. Hexagons are the results of the amalgamation step. The numbers inside each node refer to the number of cases (marriage observed) and the total size of the node. Beneath each non-terminal node the splitting statement is reported. Beneath each hexagon the results from Cox hazards regression are presented: the relative risks – with the lowest survivorship group as the reference set – and the median age at marriage.

The interpretation of the results can be obtained directly by observing the hexagons at the very bottom of the tree. A hexagon identifies a collection of disjoint subsets of observations (terminal nodes). Each terminal node is defined by a set of characteristics which are associated with a peculiar pattern of marriage but having a survival experience not sufficiently distinct from the other subsets belonging to the same cluster.

Before commenting upon the results, it could be useful to illustrate the main steps of the tree-building process. The first split of the sample is made on the basis of the variable WORKING STATUS. Among all the possible partitions of women according to the values of the covariates considered, the distinction between

not working and working (now or in the past) gives the best separation of the observations so to maximise the difference in timing of marriage. The descendant left branch of the tree includes women with higher propensity and lower age at marriage, while the right branch includes those who tend to delay their marriage.

Going down the left branch, housewives show a different behaviour according to their BIRTH COHORT: the splitting rule separates out women born after 1938 (left) from those born before (right). A comparable cohort effect appears also in the right branch of the tree (working women) even though at a later stage. For these women, in fact, distinct patterns of timing to marriage are first generated by their EDUCATION, the high school or university degree involving, as expected, a longer duration.

The subsequent split, whichever the branch, is determined by BIRTH COHORT. Even though this covariate contributes significantly to the tree-growing process, at this step only three terminal nodes are generated. Indeed, marriage behaviour of women seems to be quite stable throughout cohorts. Even if the period indicators of nuptiality in Italy are showing a strong pattern of change – i.e. the gross nuptiality rate is decreasing and the mean age at marriage is increasing – if one computes the same figures by cohorts the same trend is not observed. Therefore, differences among women are mostly not due to a cohort effect.

The fourth (and fifth) level of the tree identifies fourteen nodes which are not split further (this is, indeed, the result of the pruning process). At this step the geographical AREA OF RESIDENCE determines many splits, though the results are not so obvious as one could expect: for example, women with the highest propensity to marry (the leftmost terminal node) do not live in the South of Italy; women living on the Islands (Sardinia and Sicily) usually show same behaviour as those in the North or the Centre, while women in North-East have a pattern of marital behaviour which is closer to that of women living in the South rather than in the other regions of Northern Italy, as might have been expected. This aggregation would probably have failed to emerge using a different method of analysis. The popular hazards regression model, for example (see Table I, later in this section), would estimate the effect of each area on the "risk" of marriage, assuming that this risk varies among pre-defined domains and the real overlapping of regions will not be captured.

As to the results for OCCUPATIONAL POSITION, the relevance of this covariate also for housewives has to be noted. Among the women who were not working at the time of the interview, many had worked at some time in their life. Even though they have been sent left by the first split they end up in the rightmost of the terminal nodes in this part of the tree, eventually grouping, at the amalgamation step, in a cluster of women working at the time of the interview (right branch of the tree). It would be useful to have some retrospective information on the real time-of-the-event value of the covariates. The assumption which is made herein is that these values are the same as those observed at the time of the interview, which is not, of course, entirely satisfactory.

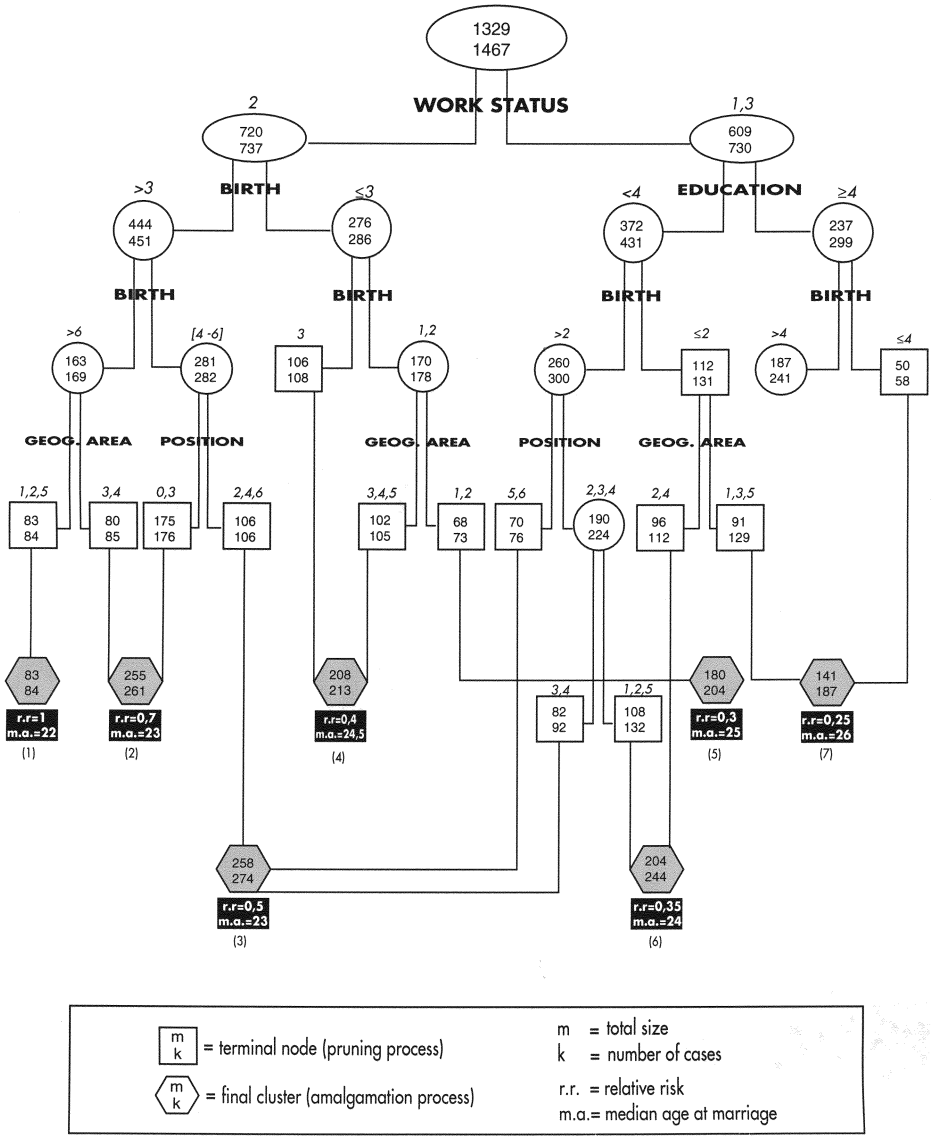

Figure 2. Tree structure for progression to marriage.

The amalgamation step concludes the tree building process, joining those groups which do not show very different survivorship to marriage, even if descending from distinct branches. Namely, different subsets of individual characteristics, which identify different profiles, can lead to homogeneous patterns of marriage behaviour. Among the seven clusters created, the pattern of survivorship is the most heterogeneous. For these groups the relative risk and the median age at marriage are calculated, using as reference group the terminal node at the very left

(housewives, born after 1953, living in the North or in the Islands). Moving from left to right we observe female profiles associated with increasing survivorship to marriage and higher median age. Women working at the time of the interview or in the past, with higher education (high school or university), either born before 1944 (no further specifications) or, if born after 1944, live in the North-West, the Centre or the Islands are less prone to an early marriage. It has to be noted that out of the fourteen terminal nodes which result from the pruning process there are only three that during the amalgamation step are clustered with terminal nodes descending from a different second level parent, namely that cross from one side to the other of the tree structure as defined from the split of the root node. This shift is in some sense conditional upon the effect of birth cohort. Indeed, the common feature of the women grouped in the third 'risk' group in Figure 2 is a birth date after 1933, while those gathered in the cluster next to the lowest risk group were all born before 1933. This outcome is in accordance with previous evidence (Pinnelli and De Rose, 1995) that women born after the first years of the 30s changed their nuptial behaviour with respect to earlier generations, being more likely to marry and at a lower age. The usefulness of the tree structure is that it shows a kind of local effect of the covariate, conditional to a given path along the tree.

An idea of the characteristics of the other amalgamated groups can be easily obtained from the observation of the tree in Figure 2.

Figure 3 displays the Kaplan-Meier estimates of the survival function (Kalbfleisch and Prentice 1980) for the seven subgroups finally identified by recursive partitioning and amalgamation. It is interesting to observe that a distinct pattern of survival between groups belonging to the two sides of the tree (clusters 1–4 versus clusters 5–7) can be recognized throughout the age span, while the survival curves are the most distant in the nearby of the age interval most exposed to the 'risk' of marriage, namely for ages between 20 through 28.

In order to provide a basis for comparison the classical Cox regression hazards model was also fitted to the data. The relative risks in the final model are reported in Table I.

The results of Cox regression suggest that women born after 1933 have an increasing risk of marriage compared to women in earlier birth cohorts, while an apparent trend turnover is observed only for the youngest generation (1959–63), but note that at the time of the interview the upper limit of age in this cohort was only 29. Working condition also had a significant impact on the propensity to marry, such as obtained using recursive partitioning, so that women who were working at time of interview were less likely to marry and married later than housewives. Unlike the tree regression results, education does not display any significant effect in Cox regression: indeed, looking at Figure 4 it can be easily deemed that this covariate is likely to violate the proportionality assumption (the survival curves overlap each other and cross in many points of time), so that it fails to show any significant effect under the proportional hazards model.

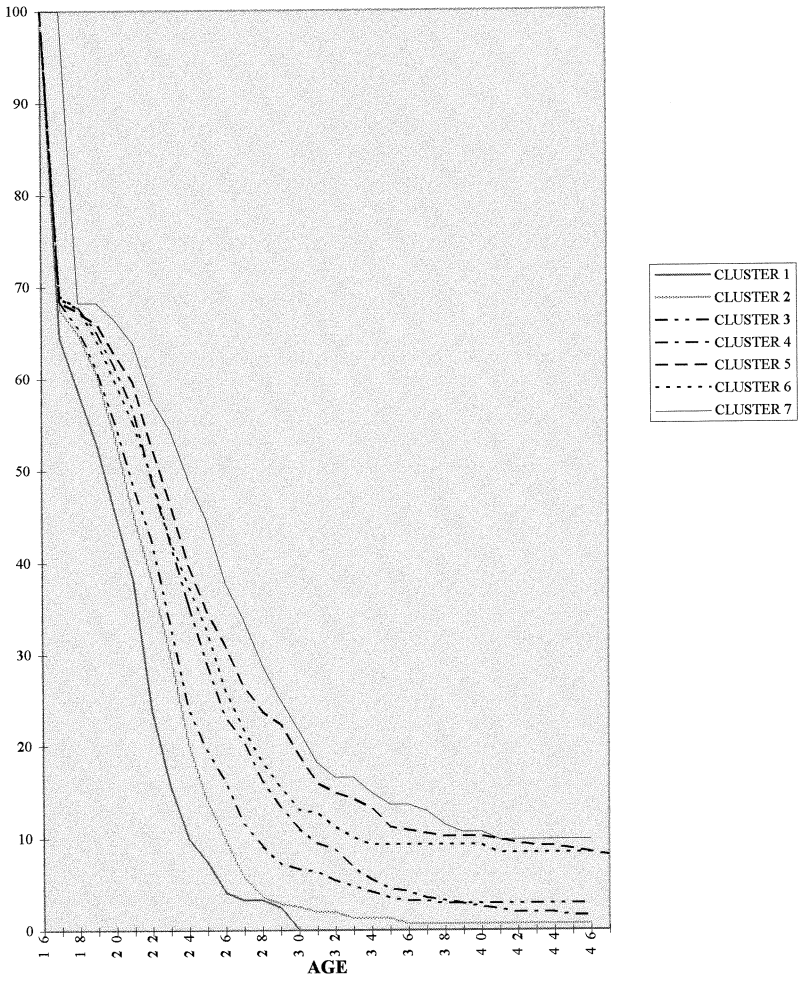

Figure 3. Survival function by clusters. (Kaplan-Meier – %)

The estimated coefficients for the geographical area of residence suggest merely that women living in the North have a distinct pattern of behaviour compared with women living in any other region of Italy. The Cox model fails to show the kind of "local" overlapping among geographical areas which has been recognized during tree-growing process. Finally, as for occupation, this covariate fails to show any significant effect, while it plays some role in the tree-growing process, although the interpretation of its effect is not completely clear.

On the basis of the significant main effects of the Cox model, one would conclude that women in work, living in the North, have the lowest risk of marrying. This risk increases inversely with birth cohort, with earlier generations showing higher survivorship. These conclusions are not completely consistent with the findings

Table I. Results of Cox regression

COVARIATE	RELATIVE RISKS (* significant at p-value = 0.95)
BIRTH COHORT	
1924–1928	1
1929–1934	1.09
1934–1938	1.42*
1939–1943	1.77*
1944–1948	1.96*
1949–1953	2.04*
1954–1958	2.47*
1959–1963	2.10*
WORKING STATUS	
Not working	1
Working	0.63*
Retired	0.91
EDUCATION	
University degree	0.93
High school diploma	0.93
Middle school diploma	0.95
Primary school	1.03
No qualification	1
AREA OF RESIDENCE	
North-West	0.68*
North-East	0.76*
Center	1
South	0.98
Islands	1.18
OCCUPATIONAL POSITION	
Never worked	1
Manager	0.93
White-collar	0.88
Skilled-worker	1.03
Unskilled-worker	0.93
Enterpreneur	0.84
Self employed	1.17

of the regression tree procedure. In the tree structure the survival prospect of the women gathered in one cluster is not due merely to the direct effect of the individual characteristics, but to some local interactions, conditional to a given path along the tree. Interaction terms can be included in the proportional hazards model in order to estimate the joint effect of covariates on the survival time, but practical

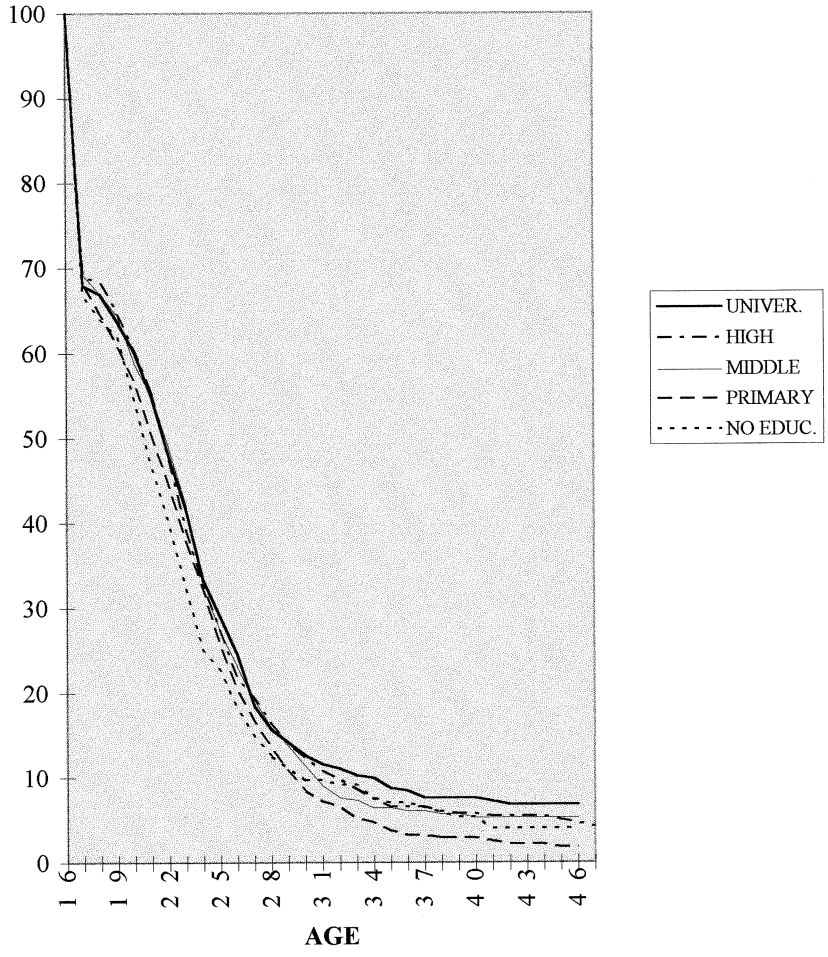

Figure 4. Survival function by level of education. (Kaplan-Meier – %)

experience suggests that in this case the statistical modeling becomes soon very time consuming and the interpretation of the results quite complicated, frequently for a beneficial effect which consists of a low improvement of the fitting.

One likely conclusion of the outcomes of the application is that the two statistical techniques can be deemed to be somewhat complementary for survival data analysis. The Cox proportional hazards model, while useful for evaluating covariate effects, is not adept at identifying homogeneous subgroups of individuals. Recursive partitioning is ideally suited for detection of conditional information, thus helping to define group of individuals with different survival probability.

4. Discussion and concluding remarks

This paper has illustrated tree-structured methodology, a non-parametric method of analysis of multivariate data, which is becoming popular in many fields of applied statistics research (for a list of applications see Breiman et al., 1984; Pallara, 1992).

In particular, the survival trees technique has already proved to be a useful tool for censored data analysis, especially when the primary interest is in defining several groups of individuals with important differences in survival. Some applications have appeared using this approach in medical settings (see, among others, Ciampi et al., 1989; Piette et al., 1992; Bacchetti and Segal, 1995), but, as far as we know, it has never been used for survival data analysis in demography and social sciences.

The advantages afforded by tree procedure are manifold. They require a minimum of prior beliefs to be imposed on the estimation and yet they yield predictions which are seldom less accurate than those obtained using standard statistical methods. The technique is particularly effective if the structural relationships have jump discontinuities or are otherwise non-linear. Moreover, it is extremely robust with respect to outliers, particularly among predictor variables. The tree-growing method makes the processes of covariate selection and grouping of categories in event history models explicit. Lastly, a binary tree is a powerful graphical tool which is easy to interpret and to use.

The purpose of the simplified example proposed in the paper has been to show some of these advantages. The outcome of the procedure allows immediate identification of the "most significant" groups in terms of survival to a certain event as well as singling out non-trivial classifications of individuals. With respect to the classical proportional hazards regression model, the results are rather different and the comparison of the outcomes of the two methods emphasizes some of the above mentioned advantages of tree methodology. The Cox model seems to be better suited for settings, such as in medical sciences, where the scope of the analysis is to evaluate the direct effect of selected covariates on the survivorship prospects of individuals. In socio-demographic studies, where it is of primary interest to represent the complexity of human behaviour and the over-simplified assumptions of standard linear models are rarely satisfied, the notions of local interaction and conditional effect seem to be the most appropriate; in such setting tree-structured survival analysis can prove to be very useful.

On the other hand, there are some limitations of tree-structured methodology. The method yields an approximation which has discontinuities (*piecewise-constant*) at the boundaries of the sub-regions in which the measurement space is subdivided by the splitting process. Moreover, some problems have been observed with the stability of the tree-structure obtained. This problem typically originates from the presence of associations between covariates. A consequence of this in the tree-growing process is that at any given node, there may be a number of splits on different variables having almost the same accuracy. Therefore, choosing between competing splits is somewhat random, because of a noisy component

always existing in sampling data. However, selecting an alternative split which is almost as good will lead to a different evolution of the tree from that node downward, eventually ending up in different tree topology with an almost equivalent amount of information. Illustrative examples of this problem with simulated data from the "waveform recognition" application (well-known in the literature on tree methodology) can be found in Breiman et al. (1984: 156–159) and Pallara (1992: 275–277). Breiman et al. (1984: 159) comment on the outcomes of their experiment that "in practice, tree instability is not nearly as pronounced as in these simulated examples". The algorithms for recursive partitioning comprise auxiliary information on the tree-growing process to examine competing splits at each node and explore alternative tree-structures. As for the consistency of the results of the application proposed in this paper, a few points can be highlighted. (*i*) The split of the root node on the WORKING STATUS is by far the best split. (*ii*) Examining the two nodes at the second level of Fig. 2, the split on BIRTH COHORT (left) is significantly better than the second best, while the split on the right according to EDUCATION has a competing split on OCCUPATIONAL POSITION which has an almost equivalent information content. (*iii*) Out of the four nodes at the third level of the tree structure, the split on the right of each of the two sides of the tree appears to be better defined (all of the splits are on BIRTH COHORT), namely the split on the left of each side has a competing split (based on the AREA OF RESIDENCE) which does almost as well. (*iv*) There is only one split at the fourth level of the tree which has a competing split with an equivalent information content. A smaller experiment, with about 40% of the data included in this application, obtains the same splits at the first and second level of the tree, with two competing splits at second level on the right of the tree, on EDUCATION and OCCUPATIONAL POSITION, having almost equivalent information content. The splitting and pruning process in the case of this reduced size experiment end up in a final pruned tree with eight terminal nodes. Altogether, the tree structure examined in this application appears fairly stable, at least for nodes near the top of tree. This is due, to some extent, also to the circumstance that the number of predictors used in this application is not very large. In applications with a larger number of predictors, the associations between the measured variables worsen the problem of tree instability.

Another shortcoming in the tree growing process is that in applications including both continuous and categorical variables among predictors, splits on continuous variables tend to predominate. For a discussion on these and some other critical issues regarding tree structured methodology see Breiman et al. (1984), Pallara (1992).

Specifically, concerning the use of tree-methodology to analyse censored survival data, it seems useful to go on extending the application of the technique to more complex settings, e.g. including in the analysis time-dependent covariates (Bacchetti and Segal, 1995; Huang et al., 1995). Overall, it is believed that tree-

structured methods represent a powerful addition to the tools for solving data analysis problems in censored survival applications.

Acknowledgements

A preliminary version of the paper was presented at the European Population Conference, in the session "Event History Analysis in Demography: Methodology", Milan 4–8 September 1995. Helpful comments from Nico Keilman, Ian Diamond and one referee are gratefully acknowledged. This research was partly supported by a grant from the *Ministero dell'Università e Ricerca Scientifica o Tecnologica* (research project "Indagine su controllo e aspettative della fecondità in Italia", MURST 40%).

Note

[1] The paper is the result of a joint research of A. De Rose and A. Pallara. The two authors share the responsibility for Sections 1 and 4; Section 2 is attributable to A. Pallara and Section 3 to A. De Rose.

References

Ahn, H., 1996. 'Log-normal regression modeling through recursive partitioning', *Computational Statistics & Data Analysis* 21: 381–398.

Ahn, H. and Loh, W.Y., 1994. 'Tree-structured proportional hazards regression modeling', *Biometrics* 50: 471–485.

Bacchetti, P. and Segal, M.R., 1995. 'Survival trees with time-dependent covariates: Application to estimating changes in the incubation period of AIDS', *Lifetime Data Analysis* 1: 35–47.

Blossfeld, H.P. and De Rose, A., 1992. 'Educational expansion and changes in entry into marriage and motherhood: the experience of Italian women', *Genus* 3-4: 73–89.

Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., 1984. *Classification and Regression Trees*, Wadsworth International, Belmont, CA.

Castiglioni, M. and Dalla Zuanna, G., 1994. 'Innovation and tradition: reproductive and marital behaviour in Italy in the 1970s and 1980s', *European Journal of Population* 10(2): 107–141.

Ciampi, A., Hogg, S.A., McKinney, S. and Thiffault, J., 1988. 'RECPAM: a computer program for recursive partitioning and amalgamation for censored survival data and other situations frequently occurring in biostatistics. I. Methods and program features', *Computer Methods and Programs in Biomedicine* 26: 239–256.

Ciampi, A., Thiffault, J. and Sagman, U., 1989. 'RECPAM: a computer program for recursive partitioning and amalgamation for censored survival data and other situations frequently occurring in biostatistics. II. Applications to data on small cell carcinoma of the lung (SCCL)', *Computer Methods and Programs in Biomedicine* 30: 283–296.

Ciampi, A. and Thiffault, J., 1989. 'Pruning trees for censored survival data: The RECPAM approach', *Communications in Statistics – Theory and Methods* 18(9): 3373–3388.

Ciampi, A., 1991. 'Generalized regression trees', *Computational Statistics and Data Analysis* 12: 57–78.

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

Davis, R.B. and Anderson, J.R., 1989. 'Exponential survival trees', *Statistics in Medicine* 8: 947–961.

Gordon, L. and Olshen, R.A., 1985. 'Tree structured survival analysis', *Cancer Treatment Reports* 69(10): 1065–1069.

Huang, X., Chen, S. and Soong, S-J., 1995. 'Piecewise proportional hazards survival trees with time-dependent covariates, computationally intensive statistical methods: Proceedings of 26th symposium on the interface', Research Triangle Park, North Carolina, 242–246.

Kalbfleisch, J.D. and Prentice, R.L., 1980. *The Statistical Analysis of Failure Time Data*. Wiley.

Leblanc, M. and Crowley, J., 1992. 'Relative risk trees for censored survival data', *Biometrics* 48: 411–425.

Loh, W.Y., 1991. 'Survival modeling through recursive stratification', *Computational Statistics & Data Analysis* 12: 295–313.

Pallara, A., 1992. 'Binary decision trees approach to classification: A review of CART and other methods with some applications to real data', *Statistica Applicata* 4(3): 255–285.

Piette, J.D., Intrator, O., Zierler, S., Mor, V. and Stein, M.D., 1992. 'An exploratory analysis of survival with AIDS using a non-parametric tree-structured approach', *Epidemiology* 3: 310–318.

Pinnelli, A. and De Rose, A., 1995. 'Recent changes in the process of family formation in Italy', in H.P. Blossfeld (ed), *The Family Formation in Modern Societies and the New Role of Women*. Westview Press, 174–190.

Segal, M.R., 1988. 'Regression trees for censored data', *Biometrics* 44: 35–47.

Sonquist, J.A., Baker, E.L. and Morgan, J.N., 1973. *Searching for Structure* (rev. ed.). Ann Arbor: Institute for Social Research, University of Michigan.