# Survival Analysis and Data Mining

**Qiyang Chen**
*Montclair State University, USA*

**Alan Oppenheim**
*Montclair State University, USA*

**Dajin Wang**
*Montclair State University, USA*

## INTRODUCTION

Survival analysis (SA) consists of a variety of methods for analyzing the timing of events and/or the times of transition among several states or conditions. The event of interest can happen at most only once to any individual or subject. Alternate terms to identify this process include Failure Analysis (FA), Reliability Analysis (RA), Lifetime Data Analysis (LDA), Time to Event Analysis (TEA), Event History Analysis (EHA), and Time Failure Analysis (TFA), depending on the type of application for which the method is used (Elashoff, 1997). Survival Data Mining (SDM) is a new term that was coined recently (SAS, 2004). There are many models and variations of SA. This article discusses some of the more common methods of SA with real-life applications. The calculations for the various models of SA are very complex. Currently, multiple software packages are available to assist in performing the necessary analyses much more quickly.

## BACKGROUND

The history of SA can be roughly divided into four periods: the Grauntian, Mantelian, Coxian, and Aalenian paradigms (Harrington, 2003). The first paradigm dates back to the 17th century with Graunt's pioneering work (Holford, 2002), which attempted to understand the distribution for the length of human life through life tables. During World War II, early life tables' analysis led to reliability studies of equipment and weapons and was called TFA.

The *Kaplan-Meier method,* a main contribution during the second paradigm, is perhaps the most popular means of SA. In 1958, a paper by Kaplan and Meier in the *Journal of the American Statistical Association* "brought the analysis of right-censored data to the attention of mathematical statisticians" (Oakes, 2000, p. 282). The Kaplan-Meier *product limit method* is a tool used in SA to plot survival data for a given sample of a survival study. Hypothesis testing continued on these missing data problems until about 1972. Following the introduction by Cox of the *proportional hazards model,* the focus of attention shifted to examine the impact of survival variables (covariates) on the probability of survival through the period of third paradigm. This survival probability is known within the field as the *hazard function.*

The fourth and last period is the Aalenian paradigm, as Statsoft, Inc. (2003) claims. Aalen used a martingale approach (exponential rate for counting processes) and improved the statistical procedures for many problems arising in randomly censored data from biomedical studies in the late 1970s.

## MAIN THRUST

The two biggest pitfalls in SA are (a) the considerable variation in the risk across the time interval, which demonstrates the need for shorter time intervals, and (b) censoring. Censored observations occur when a loss of observation occurs. This most often arises when subjects withdraw or are lost from follow-up before the completion of the study. The effect of censoring often renders a bias within studies based upon incomplete data or partial information on survival or failure times.

There are four basic approaches for the analysis of censored data: complete data analysis, the imputation approach, analysis with dichotomized data, and the likelihood-based approach (Leung, Elashoff, & Afifi, 1997). The most effective approach to censoring problems is to use methods of estimation that adjust for whether an individual observation is censored. These likelihood-based approaches include the *Kaplan-Meier estimator* and the *Cox-regression,* both popular methodologies. The *Kaplan-Meier estimator* allows for the estimation of survival over time, even for populations that include subjects who enter at different times or drop out.

Having discovered the inapplicability of multiple regression techniques due to the distribution (exponential vs. normal) and censoring, Cox assumed "a multiplicative relationship between the underlying hazard function and the log-linear function of the covariates" (Statsoft, Inc., 2003) and arrived at the assumption that the underlying hazard rate (rather than survival time) is a function of the independent variables (covariates) by way of a nonparametric model.

As SA emerged and became refined through the periods, it is evident even from the general overview herein that increasingly more complex mathematical formulas were being applied. This was done in large measure to account for some of the initial flaws in the research population (i.e., censoring), to provide for the comparison of separate treatments, and to take entirely new approaches concerning the perceived distributions of the data. As such, the calculations and data collection for the various models of SA became very complex, requiring the use of equally sophisticated computer programs.

In that vein, software packages capable of performing the necessary analyses have been developed and include but are not limited to SAS/STAT software (compares survival distributions for the event-time variables, fits accelerated failure time models, and performs regression analysis based on the proportional hazards model) (SAS, 2003). Also available is the computer software NCSS 2004 statistical analysis system (2003).

## Multiple Area Applications

The typical objective of SA in demography and medical research centers on clinical trials designed is to evaluate the effectiveness of experimental treatments, to model disease progression in an effort to take preemptive action, and also to estimate disease prevalence within a population. The fields of engineering and biology found applicability of SA later. There is always a need for more data analysis. The information gained from a successful SA can be used to make estimates on treatment effects, employee longevity, or product life. As SA went through more advanced stages of development, business-related fields such as economics and social sciences started to use it. With regard to a business strategy, SA can be used to predict, and thereby improve upon, the life span of manufactured products or customer relations. For example, by identifying the timing of risky behavior patterns (Teredata, 2003) that lead to reduced survival probability in the future (ending the business relationship), a decision can be made to select the appropriate marketing action and its associated cost.

Lo, MacKinlay, and Zhang (2002) of MIT Sloan School of Management developed and estimated an econometric model of limit-order execution times. They estimated versions for time-to-first-fill and time-to-completion for both buy and sell limit orders and incorporated the effects of explanatory variables such as the limit price, limit size, bid/offer spread, and market volatility. Through SA of actual limit-order data, they discovered that execution times are very sensitive to the limit price but are not sensitive to limit size. Hypothetical limit-order executions, constructed either theoretically from first-passage times or empirically from transaction data, are very poor proxies for actual limit-order executions.

Blandón (2001) investigated the timing of foreign direct investment in the banking sector which, among other things, leads to differential benefits for the first entrants in a foreign location and to the problem of reversibility. When uncertainty is considered, the existence of some ownership-location-internalization advantages can make foreign investment less reversible and/or more delayable. Such advantages are examined, and a model of the timing of foreign direct investment specified. The model is then tested for a case using duration analysis.

In many industries, alliances have become the organization model of choice. Having used data from the *Airline Business* annual surveys of airline alliances, Gudmundsson and Rhoades (2001) tested a proposed typology predicting survival and duration in airline alliances. They classified key activities of airline alliances by their level of complexity and resource commitment in order to suggest a series of propositions on alliance stability and duration. The results of their analysis indicate that alliances containing joint purchasing and marketing activities had lower risk of termination than alliances involving equity.

Kimura and Fujii (2003) conducted a Cox-type SA of Japanese corporate firms using census-coverage data. A study of exiting firms confirmed several characteristics of Japanese firms in the 1990s. They found that in order to increase the probability of survival, an efficient concentration on core competencies, but not excessive internalization in the corporate structure and activities, is vital to a company. They also found that via carefully selected channels, a firm's global commitment helps Japanese firms be more competitive and more likely to survive.

SA concepts and calculations were applied by Hough, Garitta, and Sánchez (2004) to consumers' acceptance/rejection data of samples with different levels of sensory defects. The lognormal parametric model was found adequate for most defects and allowed prediction of concentration values corresponding to 10% probability of consumer rejection.