Statistical Computing and Graphics

Analysis and Display of Hierarchical Life-Time Data

Galit SHMUELI and Ayala COHEN

We discuss the analysis and display of data from multistage systems or processes where subjects/items go through a sequence of stages in a certain order. The purpose of collecting such data (in the form of *duration times* between the various stages) may be to study the pattern of moving between the stages, to compare the progress of subgroups, or to assess the relation between the different duration times. The methods are based on graphical displays of the duration times between the stages. The graphical methods: survival-curves, boxplots, follow-up (FU) plots, and multidimensional scaling (MDS) are adopted to handle the specific features of such data. We illustrate the different methods on the duration times between promotions in an academic ranking system.

KEY WORDS: Duration times; Graphical methods; Hierarchical-successive data.

1. INTRODUCTION

In many fields data are collected on duration times between successive stages, in order to learn about the pattern of moving between stages. Clear presentations of such data are an important preliminary step in revealing the main features of the data.

We consider the adjustment of several methods for the presentation of duration-time data. In particular, we focus on data that consist of three main characteristics.

1. *Hierarchical structure*: reaching a certain stage requires going through all lower stages.

2. *Sequential nature*: moving from stage to stage is done in a certain order and it is not possible to skip any stage.

3. *(Possibly) right censoring*: since at the time of recording not all subjects reach the final stage, they will have a right-censored promotion profile. (Left censoring is also possible, but is not considered in this article. In our example there are no left censored data.)

Such data appear in many areas. In the organizational context we may consider the duration times between successive promotions of individuals (such as between ranks in the army, ranks in the academic world, and so on). An organization might be interested in the typical duration-time pattern, or in the relation between the duration times in lower and higher ranks. Other interests might focus on the comparison of the promotion rate of subgroups (for example males versus females). In medicine, many multistage illnesses proceed from stage to stage in a certain order. By recording the periods between stages of some illness for different patients, we may learn more about the development of that illness. Furthermore, the analysis of such data may assist in assessing the effect of drugs that are aimed at lengthening the periods between the illness stages. The duration times for patients who have received the drug could then be compared with those of a control group (who has not received it).

In many biological phenomena the periods between developmental stages of organisms are of interest, as a whole or for specific stages. In some cases, such data could be extremely important. For example, a physician may decide to intervene in a pregnancy according to the time elapsed until a particular fetus developmental stage.

In the presentation of such data, we are interested both in the time spent in each stage (univariate analysis) as well as in the entire "promotion profile," that includes the duration times in all stages until the last achieved (multivariate or longitudinal analysis). We borrow the term "promotion profile" (or simply "profile") from the organizational context to describe all the duration times as a whole.

If we have k stages in the system, there are $(k-1)$ duration times (in stages $1, 2, \ldots, (k-1)$). Since at the end of the study subjects (or generally, items) might be recorded at different stages, we can form $(k-1)$ data subsets, each including full and right-censored duration times. In order to display each of these $(k-1)$ censored samples, we use survival curves and a variation of boxplots. Both methods take into account the uncensored as well as the censored information. The advantage in displaying each duration time separately is that we can use existing EDA (exploratory data analysis) methods. The disadvantage is that we lose the longitudinal information contained in the full promotion profiles by dividing them into separate parts.

In order to display the *entire* promotion profiles we use multivariate methods (follow-up plots and multidimensional scaling) by adapting them to the special characteristics of these data. We compare and discuss the advantages and drawbacks of the different methods.

All the previously mentioned methods display important aspects of duration-time data. We focus on four aspects:

- Characterizing the distribution of the time spent in each stage.
- Inferring about existing relations between the times spent in the different stages.
- Comparing "promotion" rates of subgroups.
- Identifying outliers (with extreme promotion profiles).

Galit Shmueli is a Ph.D. Student, and Ayala Cohen is Professor, Faculty of Industrial Engineering and Management, The Technion, Haifa, Israel (Email: ieayala@ie.tehnion.ac.il).

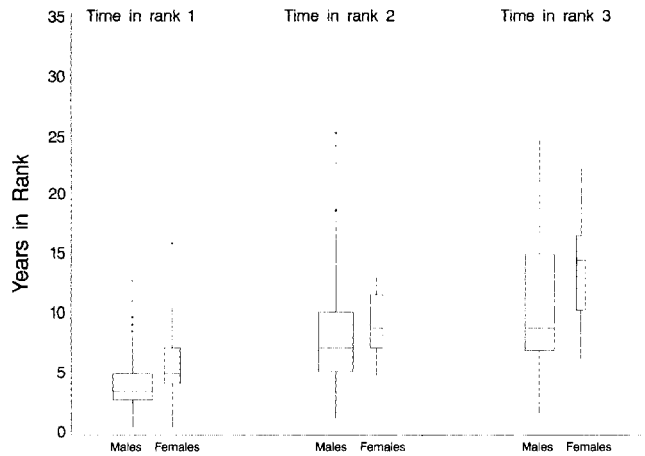

Figure 1. Side-by-Side Boxplots for Males and Females, in Each Rank. The higher the rank the longer the duration time until promotion. Females' durations in all ranks appear longer than males'.

In order to illustrate the use of each of the methods, we apply them to an academic career promotion example. Although our inference relates to a specific example, the purpose of this article is to survey various techniques and their adjustment for handling such data. We follow the same example throughout the article in order to demonstrate a methodology for a real problem.

The article is organized as follows: Section 2 introduces an example of promotion data in a university. Section 3 describes univariate display methods (boxplots and survival curves). Sections 4 and 5 include methods that display the entire promotion profiles (FU-plots and MDS). Section 6 gives concluding remarks.

2. AN EXAMPLE: PROMOTIONS IN A UNIVERSITY

We demonstrate the use of various methods on data from our university. The data consist of our academic staff promotion profiles. The academic ranking system in our university consists of four ranks: lecturer, senior lecturer, associate professor, and full professor. For each staff member the data include the duration-time (in months) in each rank until his/her current rank, and the duration-time in the current rank as well.

This research was initiated by our university management to find out whether any differences exist in the promotion patterns of female versus those of male faculty members. The management objective was to describe the current faculty rank distribution and dwell time in each rank without giving any interpretations or reasons for them. Thus, only the promotion profiles and gender were provided for this statistical study. Further analyses clearly require additional information and are beyond the scope of this article.

3. BOXPLOTS AND SURVIVAL CURVES

Boxplots and survival curves are two effective methods for displaying each duration time separately. By displaying the duration-times between k different stages as $(k-1)$ boxes on one plot, we can compare measures such as location and dispersion. In addition, it is easy to compare boxes that are plotted for subgroups.

In order to create separate datasets from a system with k stages, we divide the data into $(k-1)$ subsets, each containing the duration in a certain stage. In many cases some or all of the subsets will consist of right-censored data, for those who were at that stage at the recording time.

When censoring is present, descriptive methods that are based on the estimated CDF (or equivalently, on the survival function $S = 1 - \text{CDF}$) are preferable to methods that are based on the empirical probability/density function (such as histograms and stem-and-leaf plots). The reason is that CDF-based methods take into account the censored information (for a censored observation we know that it obtains *at least* a certain value). If we refer to the duration times in a certain stage as the survival times in that stage, then the survival function is:

$$S(t) = \Pr(\text{a person stays in the stage longer than } t). \quad (1)$$

where t is some time unit. We can estimate $S(t)$ from a censored sample by either parametric or nonparametric methods. Parametric methods imply fitting some known distribution to the data, and then using diagnostics such as Q-Q plots to assess the fit (e.g., Nelson 1982). If fitting a theoretical distribution is not attempted, nonparametric estimators of the survival function are easy to apply. A well-known nonparametric estimator is the Kaplan–Meier estimator (Kaplan and Meier 1958):

$$\hat{S}(t) = p_1 \times p_2 \times \cdots \times p_t, \quad (2)$$

where p_i denotes the proportion of subjects who survived time-unit i, out of those who have survived time-unit $(i-1)$ (Lee 1992).

When censoring is not too heavy or when most of the censored observations have high values compared to the uncensored data, the 25th, 50th, and 75th percentiles, required for a box-plot can be computed from $\hat{S}(t)$, and an ordinary boxplot can be formed. In other cases, where not

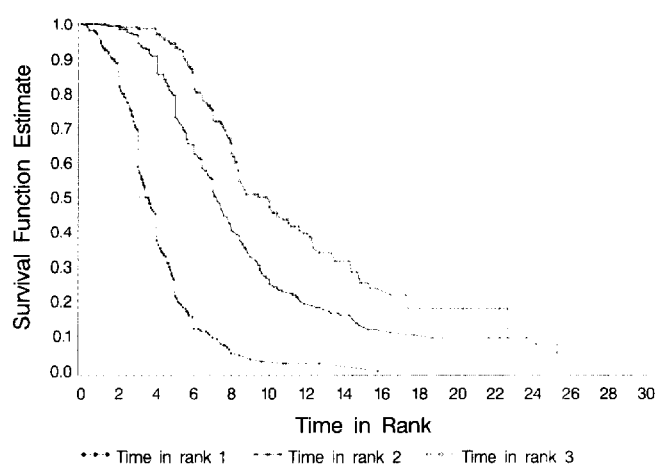

Figure 2. Survival Curves for the Duration Time in Each Rank. The lower the rank, the steeper the curve, meaning a shorter duration time.

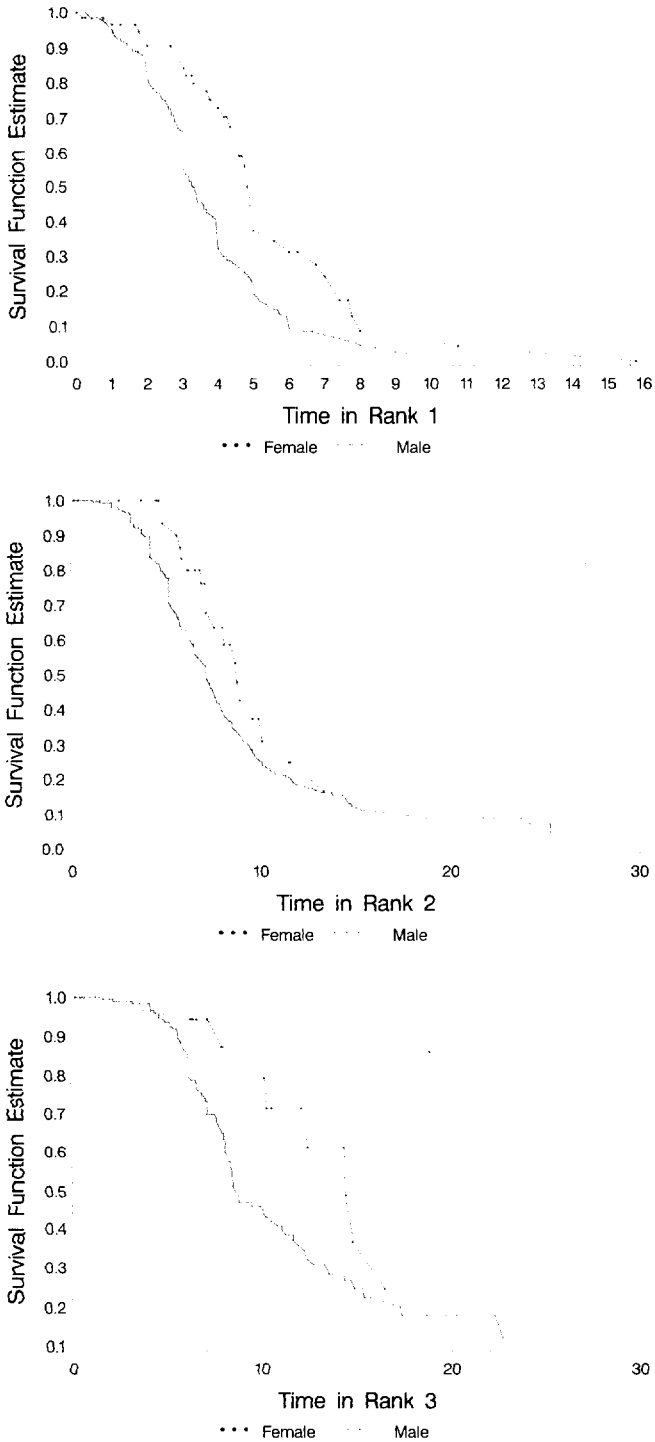

Figure 3. Survival Curves for Males and Females, in Each Rank. Males' curves are steeper than females' curves for each rank, meaning shorter duration-times in all ranks.

all quartiles are observed, we plot the observed ones and extend the sides of the box to the largest failure time (Gentleman and Crowly 1991).

We illustrate the use of boxplots on our example and display the duration times in each of the first three consecutive academic ranks (based on the Kaplan–Meier estimator). Figure 1 displays separate boxplots for female and male faculty members for each rank. The duration times are the number of years spent in that rank. The widths of the boxes are proportional to the square root of the number of subjects in that rank (McGill, Tukey, and Larsen 1978). Comparing the duration times in the three ranks we see that both location and dispersion increase with the rank. Comparing males and females, we see that in each rank females tend to spend longer time until promotion, as compared to males in the same rank.

A fuller description of the duration time distribution is obtained by plotting an estimated survival curve, which is $\hat{S}(t)$ versus t (where t in some time-unit). The steepness of the curve corresponds to the rate of "surviving" in that rank before being promoted. The steeper the curve, the probability of being promoted grows faster with time. We compare the rate of promotion of subgroups by plotting their survival curves on the same plot. Figure 2 presents the three survival curves, for the duration times in each of the first three ranks. It shows that the lower the rank, the steeper the curve. The medians of duration times in each rank can be estimated by drawing vertical lines for each curve emanating from $\hat{S}(t) = .5$. The estimated medians are 3.5, 7.5, 10 years for lecturer, senior lecturer, and associate professor, respectively. This indicates that the promotion rate is slower in the higher ranks.

Comparing subgroups is also possible by plotting their different survival curves on one plot. For example, by comparing the curves of females and males in each rank (Figure 3), we see a consistent higher promotion rate for males (= steeper curves).

Both boxplots and survival curves describe the distribution of the duration time in each rank. They are easy to apply and to interpret, but they lose the longitudinal information contained in the promotion profiles. Boxplots are a compact display of each duration time distribution, while the survival curves are more detailed.

We now refer to methods that consider the *entire* promotion profiles (longitudinal view).

4. A VARIATION ON FU PLOTS

Follow-up (FU) plots were originally presented for following up the frequency and patterns of longitudinal data (Lesser, Kohn, Napolitano, and Pahwa 1995). We demonstrate a variation of the FU plot. These modified plots describe the "promotions" of each subject during the study period, and at the same time they display the entire sample promotion pattern. In our version of the FU plot, the vertical axis consists of the time units of the study (such as years since entering the career or developing an illness). Each subject (or item) is represented by a column, which is a vertical line proportional in length to the subject's total time in the system. Dots in each column denote the respective uncensored promotion times of the subject. In fact, these dots divide the line into subintervals with lengths corresponding to the duration times in the different stages. Ordering the subjects from lowest to highest rank/stage on the horizontal axis, reveals various features of the promotion pattern.

In order to illustrate this, we present a FU-plot variation for our academic staff example (Figure 4). The subjects are

sorted by their rank at the end of the study, from first rank (on the left) to the last rank (on the right). Within each rank the subjects are sorted according to the time spent in the university until promotion to *associate professor* (or censoring, for lecturers and senior lecturers). Originally we used different colors for the promotions to each rank, to make the plot clearer and more appealing; here we use a gray scale instead, with three shades: light gray, dark gray, and black.

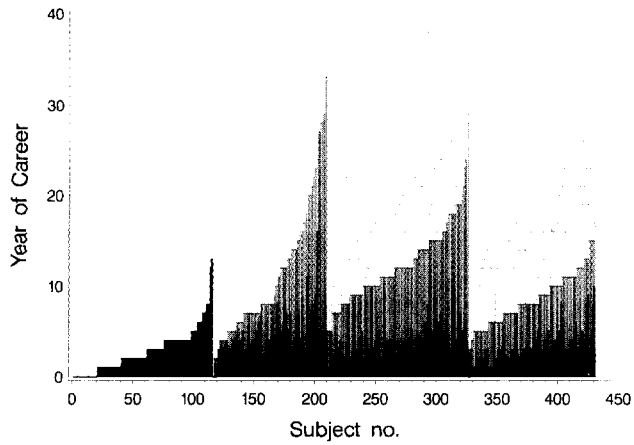

Figure 4. FU Plot for Promotion Data. Subjects in each rank are sorted by their time in the system until promotion to associate professor (or censoring for lower ranks). Time of promotion to senior lecturer is marked by black dots, and to associate professor by light gray dots.

Comparing subjects in ranks senior lecturer (rank 2) and above, we see that the time from entering the system until promotion to associate professor (or until censoring) is longer for lower ranks (denoted by the dark gray area).

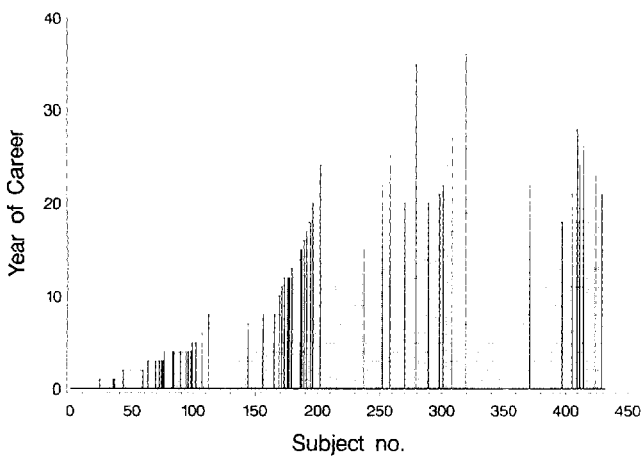

Figure 5. Comparing Males and Females on the FU-Plot. Females (black columns) are located mainly on the right sides in each rank (= longer duration times).

In addition, we infer about the relation between the duration time until promotion to associate professor and to full professor. As can be seen for full professors (rank 4), those who took longer to get promoted to associate professors (denoted by "higher" dark gray areas) tend to have longer columns. This means that they stayed longer as associate professors before being promoted to full professor. The same relation emerges for associate professors (rank 3); namely, subjects with larger dark gray areas tend to have longer columns (if we disregard the very long columns).

Comparing the promotion patterns of subgroups is done by using different colors/shades on the plot. For example, we compare females and males on the above FU plot by denoting columns of females in black and males in gray (Figure 5). Females who are senior lecturers and full professors are generally located on the right-hand side, which corresponds to a slower promotion rate. For associate professors, although the black columns are located more towards the center, many of them have very long columns (which means spending a long time in the university).

Sorting subjects, within each rank, according to some criterion enhances aspects concerning this criterion. For example (Figure 6), sorting the staff members within each rank according to the time spent in the university until promotion to *senior lecturer* (or until censoring, for lecturers), implies that the distribution of the duration time as lecturers is very similar for all four ranks. (This is not surprising since in our university there is an upper time limit allowed for staying in the rank of a lecturer.)

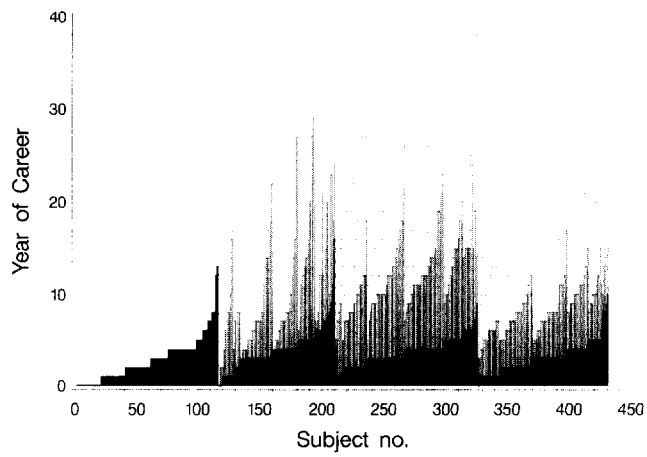

Figure 6. Different Sorting of Subjects on the FU Plot. According to their time in the system until promotion to senior lecturer. The distributions of the duration times as lecturer appear to be the same for subjects in the different ranks.

5. USING MDS

Multidimensional scaling is a technique to display multivariate data in a lower dimension (usually two dimensions), using as input the distances between each pair of multivariate observations. A simple illustration is the construction of a map (like the map of the USA) from a table of distances between points on the map (such as flying mileages between 100 US cities; Kruskal and Wish 1978).

Using multidimensional scaling (MDS) with promotion-like data means displaying the subjects as dots on a map, in a way that close dots on the map represent subjects with

"close"/similar profiles, and far dots on the map represent subjects with dissimilar promotion profiles. In order to create a distance matrix between subjects (according to their promotion profiles) we must define a distance between two promotion profiles. The choice of a distance function depends on the specific characteristics of the data that we want to emphasize. For example, we could give different weights to duration times in different ranks or we could use different norms (absolute values or squared differences).

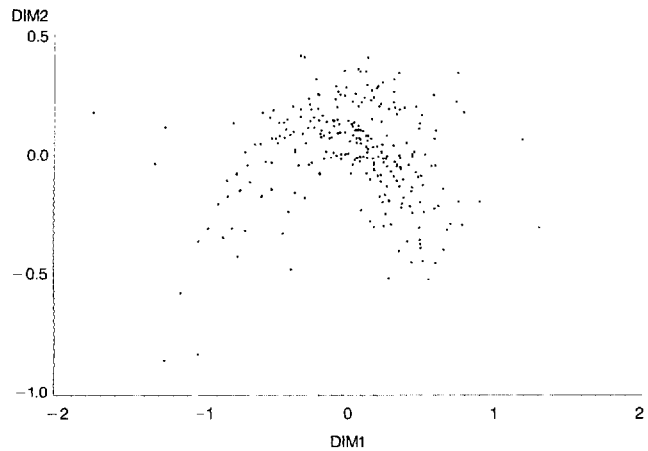

Figure 7. Two-Dimensional Map of Promotion Profiles. Close dots represent subjects with similar promotion profiles.

In order to illustrate the use of MDS for duration times, we present the two-dimensional map created by applying MDS to our promotion example (Figure 7). In this case, we chose an arbitrary distance which is both easy to understand and to compute (see Appendix). The map excludes subjects who were lecturers at the end of the study, since in our example subjects with only one censored duration time add noise rather than valuable information.

The dots form a horseshoe shape, which implies that one dimension could suffice. Other goodness of fit criteria also imply the same. Still, a two-dimensional map is used to give a clearer picture. Outliers appear as dots that are very far from the rest. (Two such outliers are marked on the map, one of which corresponds to a very slow promotion profile and the other to a very fast one.)

We use different shades to find the relation between the location on the map and the duration times in each rank. For example, we use a gray scale on the MDS map to denote the duration time as lecturer (Figure 8). We can now see that subjects are arranged on the horseshoe from the bottom-left, corresponding to short durations, through the top-middle and until the right side corresponding to long durations as lecturers. Similar maps can be formed for the duration times in the other ranks, using a gray scale to mark the duration time.

After finding the meaning of a location on the map, we can compare subgroups by checking their different locations on the map. Figure 9 compares female locations versus the male locations. All females (denoted by black dots) are located on the right half of the horseshoe, corresponding to the slower promotions.

6. CONCLUDING REMARKS

The different methods presented in this article enhance different aspects of duration-time data, yet all of them show the same main features of the data. These methods reveal the relations between the duration times in the different stages, they are useful for comparing subgroups, and for detecting outliers. In the example presented, all methods showed a positive relation between the promotion rate in

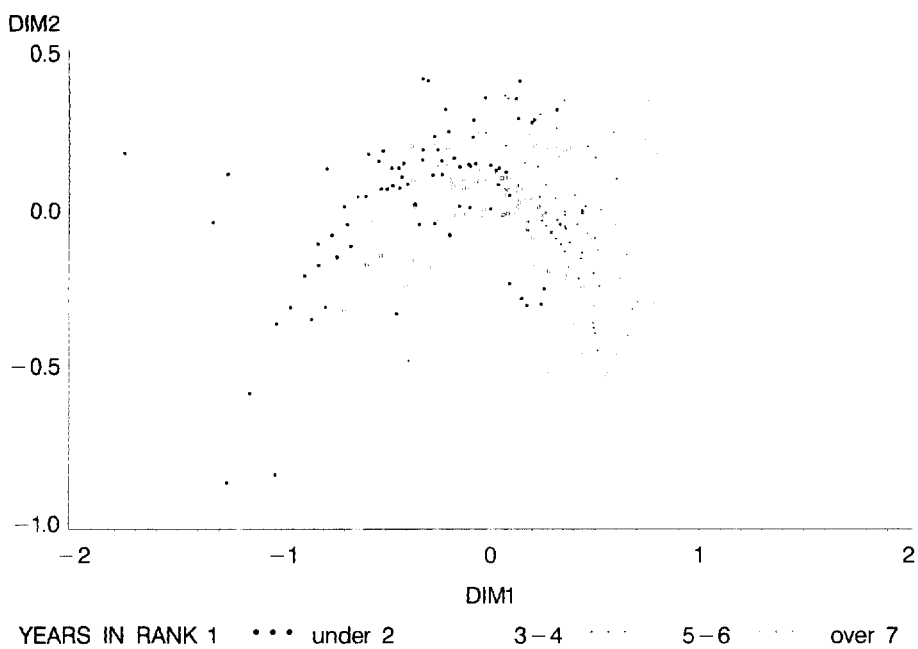

Figure 8. MDS Map According to Time Until First Promotion. Dots towards the left side of the plot represent subjects with quick first promotions.

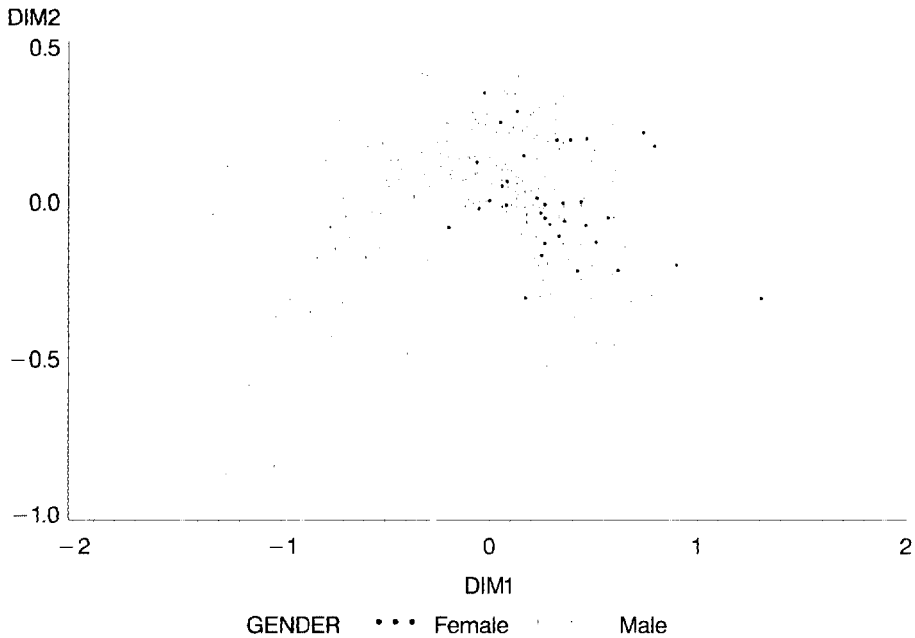

Figure 9. Location of Females Versus Males on the MDS Map. Females are located only on the right side, which corresponds to long durations as lecturers.

higher ranks to that in lower ranks. Comparing male and female faculty members consistently indicated a slower promotion rate of the latter. This finding does not necessarily imply that there is a discrimination, since no consideration was given to covariates. One should recall the example of Freedman, Pisani, and Purves (1980) on the admissions of graduate students at Berkeley in 1973. The seemingly significant difference in the acceptance rates of males versus females which was found there was contradicted when the comparison was done separately for each major. Further investigation of our example can be done provided more data are available. This may involve fitting regression models for each stage separately where the duration time in a certain stage is explained by the duration times in lower stages, and by other covariates (e.g., faculty, number of publications). The regression models should obviously be suitable for censored data (log-linear or Cox proportional hazard models).

Graphical presentations of data serve many purposes. First, the displays in themselves can be very informative and may assist in the understanding of the structure of the data. Second, they are clear and easy to interpret for non-statisticians, and are less bound to be misunderstood than are formal analyses. Third, as in our case, they may imply directions for further investigation of the data.

The methods presented were adjusted to handle the specific characteristics of promotion data; namely, hierarchical sequential data with right censoring. Variations can easily be made to fit different features of promotion data, such as left censoring, time limits on duration times etc. In those cases, though adjustments should be made, our basic deas and methodology still apply.

APPENDIX: A DEFINITION OF DISTANCES BETWEEN PROFILES

In order to apply MDS to our data, we define the distances between each two promotion profiles as illustrated in Figure 10. The plot consists of two promotion profiles: the dashed line represents a subject who was promoted in his/her 2nd, 8th, and 14th year in the system (to senior lecturer, associate professor, and full professor, respectively). The smooth line represents a subject who was promoted in his/her 4th and 10th year in the system (to senior lecturer and associate professor), and at the recording time was still an associate professor (in his 15th year in the university).

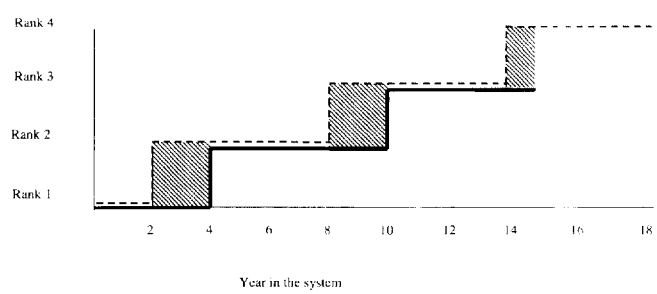

Figure 10. A Distance Between Two Promotion Profiles. The distance between these two profiles is defined by the marked area divided by its common basis.

If we define a "block" as a one-year difference between two consecutive ranks, then the distance between two profiles is defined as the area between their lines (in this case five "blocks") divided by their common basis (15 years).

The distance between the promotion profiles presented in Figure 10 is then $\frac{5}{15} = .333$.

[Received June 1997. Revised February 1998.]

REFERENCES

Freedman, D., Pisani, R., and Purves, R. (1980), *Statistics*, New York: W.W. Norton.

Gentleman, R., and Crowly, J. (1991), "Graphical Methods for Censored Data," *Journal of the American Statistical Association*, 86, 678–683.

Kaplan, E.L., and Meier, P. (1958), "Nonparametric Estimation from Incomplete Observations," *Journal of the American Statistical Association*, 63, 457–481.

Kruskal, J.B., and Wish, M. (1978), *Multidimensional Scaling*, Sage University Paper Series on Quantitative Applications in the Social Sciences, 07-011, Beverly Hills and London.

Lee, E.T. (1992), *Statistical Methods for Survival Data Analysis*, New York: Wiley.

Lesser, M.L., Kohn, N.E., Napolitano, B.A., and Pahwa, S. (1995), "The FU-PLOT: A Graphical Method for Visualizing the Timing of Follow-Up in Longitudinal Studies," *The American Statistician*, 49, 139–144.

McGill, R., Tukey, J.W., and Larsen, W.A. (1978), "Variations of Box Plots," *The American Statistician*, 38, 12–16.

Nelson, W. (1982), *Applied Life Data Analysis*, New York: Wiley.