Analyzing Shelter Stays and Shelter Stay Patterns Using Administrative Data: An Overview of Practical Methods

Stephen Metraux, MA

and

Dennis Culhane, PhD

Center for Mental Health Policy and Services Research
University of Pennsylvania

Paper presented at the Homeless Services Data Users Meeting
September 13-14, 1999
Washington DC

Draft Paper: Comments Invited

Address Correspondence to:
Steve Metraux
Center for Mental Health Policy and Services Research
3600 Market St. #730
Philadelphia PA 19104-2648
215-349-8487; metraux@cmhpsr.upenn.edu

Abstract

Information on stays and days spent in shelters is among the most basic of data collected through administrative records of homeless facilities. This paper goes over methods by which to analyze this data and which yield results that contribute to understanding the dynamics of how people use shelters. First, the paper shows ways to convert shelter stays, however they may be measured, into standardized units called episodes. The second section covers measures of central tendency and frequency distributions, methods that are likely to be readily familiar to most persons. The third section explains two basic methods of survival analysis, survival curves and hazard curves, while the fourth section describes regression procedures, which are more complex measures of survival analysis. Finally, the fifth section covers two methods for segmenting the shelter population by use pattern: the basic heavy user analysis and the more complex cluster analysis. In surveying all of these methods, frequent examples are provided to illustrate the concepts that are discussed and to facilitate understanding of how the specific method of analysis might be used to provide practical information that would be of use in operating and improving shelter services.

Introduction

Information on stays and days spent in shelters is among the most basic of data collected through administrative records of homeless facilities. This paper goes over methods by which to analyze this data and which yield results that contribute to understanding the dynamics of how people use shelters. A better understanding of shelter utilization patterns can make a shelter or shelter system more responsive to the needs of its clientele as well as to the need to provide more efficient, cost effective services. A range of methods is described here. We have attempted to explain them in a manner that keeps the paper accessible to readers with a minimal background in statistics. However, while some of these methods can be implemented with a minimal background in statistics, others will require that the analyst have a more extensive knowledge of statistical concepts and procedures. Even in these cases, however, we seek to show the general reader, by use of examples from our research, what these methods measure, how they might be applied, and to what purposes the results may be used.

I. Defining a Stay or Episode

The beginning of any analysis of shelter stays or episodes involves defining what is meant by a stay and an episode. There are different ways to define these two terms, and it is imperative that an analysis clearly and unambiguously give operational definitions of the terms. One's use of "stay" or "episode" does not have to be uniform with uses in other studies, but it does have to be clear and consistent in the way one uses it.[1] We make no exceptions to this here, and thus will start by defining our use of these terms.

Stay

[1] For a discussion of the importance of definitional issues as they pertain to homelessness, see Cordray and Pion 1991

Table 1 – Sample Stay Data

Family ID	Adults	Children	Shelter	Start Date	End Date
104	2	3	WC	7/19/96	8/12/96
104	2	5	WC	8/12/96	9/12/96
104	2	5	XA	9/12/96	11/1/97
104	1	5	XA	11/1/97	8/31/98
111	1	1	WA	7/27/96	8/1/96
111	1	1	XD	8/1/96	8/21/96
111	1	1	XD	8/21/96	10/30/96
111	1	1	WA	4/16/97	4/18/97
111	1	1	XD	4/18/97	11/8/97
111	1	1	XD	11/8/97	11/26/97
121	1	2	YD	8/25/92	9/3/92
121	1	2	YA	9/3/92	9/8/92
121	1	2	YA	9/8/92	9/9/92
121	1	2	YB	9/9/92	9/14/92
121	1	2	YB	9/16/92	7/4/93
121	1	2	YB	7/4/93	7/6/93
121	1	2	YA	7/13/93	7/14/93
121	1	2	YB	7/14/93	12/6/93
121	1	4	WC	7/10/96	8/3/96
121	1	4	XC	8/3/96	7/26/97

For the purposes of this paper, we define a shelter "stay" to mean a single unit or record of shelter stay. As such a shelter stay could be one event that is part of a larger series. A shelter stay will be a discrete record that has a beginning date and an end date as it is documented on an observation in a database. What a stay consists of differs by shelter provider; one shelter may, for example, consider a shelter stay to consist of a spell of total uninterrupted time spent in a shelter, another shelter may record stays nightly. In such cases, if a family stays in a shelter from January 1 to January 7, the former shelter will record this as one shelter stay, and the latter shelter will record that as seven shelter stays.[2]

Table 1 provides an example from the stay level database used by New York City's Department of Homeless Services (DHS) to keep records of shelter utilization in its family shelter system. This database logs the times spent by every family in the family shelters that DHS either operates, oversees, or funds. The columns included in Table 1 include an identification number (not found in the DHS database) that corresponds to a specific household; the number of adults and children in the shelter during the stay; a code for the shelter facility in which the stay occurred, and the dates each stay commenced and ended.[3]

[2] This is also potentially ambiguous, as January 1 to January 7 can mean the family entered the shelter on the day of the first and left on the day of seventh, in which case six shelter nights were spent at the shelter, or they may have spent each night from the first and including the seventh in the shelter, in which case the family would have spent seven nights in the shelter. Neither way is incorrect in recording time spent in shelter, but it is necessary to clearly indicate the method by which time in a shelter episode is recorded.

[3] Not all variables listed in this DHS database are included on Table 1. Numbers are used to delineate different families, and details in the records were changed to maintain confidentiality.

While the stays in Table 1 often span multiple days (and in some cases months), there are also instances where an uninterrupted period of time spent in a shelter is parsed into two or more stays. This is often done to note a change in the status during the time a family stays in a shelter. For example, family #104 has four consecutive shelter stays. The first stay ends, and the second stays starts, when two children join the household. The second stay ends, and the third stay starts, when the family gets transferred to another shelter. Finally, the third stay ends and the fourth stay starts when one of the adults in the household leaves the shelter. From Table 1 it is apparent that, in the DHS dataset, extended periods of time which the family spends in the shelter system are often parsed due to administrative reasons rather than due to any breaks in which the family leaves the shelter system.

Episodes

A shelter "episode" represents one or more shelter stays that can be grouped together into one discrete time period of shelter use. Stays are grouped into episodes mainly for two reasons. First, stays that are consecutive or that occur with short time intervals between them can be combined into one episode. Second, grouping stays into episodes can be done to standardize the different ways by which different shelters may record shelter stays.

As was just mentioned, many of the stays on Table 1 are consecutive. Looking at family #104, they have a record of four shelter stays, yet they cannot be considered, on the basis of this data, to be repeat shelter users. Thus one useful function of converting stays into episodes is to collapse consecutive stays into one episode. Such a conversion of the stays in Table 1 leads to the episodes in Table 2.

Table 2 – Episodes Derived from Collapsing Consecutive Stays in Table 1

Family ID	Maximum Adults In HH	Maximum Children in HH	Last Shelter Before Exit	Start Date	End Date
104	2	5	XA	7/19/96	8/31/98
111	1	1	XD	7/27/96	10/30/96
111	1	1	XD	4/16/97	11/26/97
121	1	2	YB	8/25/92	9/14/92
121	1	2	YB	9/16/92	7/6/93
121	1	2	YB	7/13/93	12/6/93
121	1	4	XC	7/10/96	7/26/97

The two main differences between Table 1 and Table 2 are that, first, 30 observations for three families get reduced to seven observations, and, second, that the headings for the second through fourth columns are relabeled. Collapsing the consecutive stays from Table 1 into the episodes on Table 2 represents a much more intuitive way to view a time period spent in a shelter. One can, for example, now readily say that, on Table 2, two out of the three families experienced repeat shelter episodes. However, one disadvantage to collapsing stays in this fashion is that information is lost in this process. In Table 2, one cannot know all the shelters that a family used during the course of their episode, nor can one know that family #104 had two changes in household composition during their six-week long shelter episode. In converting stays to episodes, one needs to weigh the advantages to the disadvantages involved in the

conversion, and one needs to mention this to the reader along with explaining the criteria used for deriving episodes.

One can also collapse episodes even further by combining episodes where the time interval between them is less than a specified duration. In Table 2, for example, family #121 leaves a shelter on September 14 only to return to the same shelter two days later, on September 16. The family does this again on July 6 of the subsequent year, returning to the same shelter one week later. Finally, the family leaves on December 6, and does not return again for 3½ years, and this time to a different shelter. This latter gap between shelter episodes is almost sure to be qualitatively different from the first two gaps, which are better characterized as respites (perhaps spent in a motel or with extended family) than as extended exits from shelter. Thus to say that Family #121 experienced four separate shelter episodes may be literally true, but may be misleading as the first three episodes can also be considered as one episode.

Table 3 shows the stays in Table 1 when they are collapsed into episodes using a "30-day Exit" criterion. This means that any two stays that occur within 30-days of each other are collapsed into one episode. Doing this presupposes that any gap between stays of less than 30-days does not constitute an extended shelter exit, and so should not mark the beginning of a new shelter episode. While this helps in differentiating between extended shelter exits and what might be considered as respites from shelter, determining the proper gap to use as an exit criterion will always be, to an extent, arbitrary. Currently a 30-day exit criterion is the most widely used in the homeless research literature.[4] One can, however, change the length of this gap to suit one's particular situation, keeping in mind that, regardless of what that criterion is, one needs to inform the reader of the criterion selected and the rationale for selecting it.

Table 3 – Stays Derived from Collapsing Episodes in Table 1 Using a "30-Day Exit" Criterion

Family ID	Maximum Adults In HH	Maximum Children In HH	Last Shelter Before Exit	Start Date	End Date
104	2	5	XA	7/19/96	8/31/98
111	1	1	XD	7/27/96	10/30/96
111	1	1	XD	4/16/97	11/26/97
121	1	2	YB	8/25/92	12/6/93
121	1	4	XC	7/10/96	7/26/97

Selecting an exit criterion of anything greater than one day leads to an added loss of information with regards to knowing how many nights a household actually spent in a shelter. In Tables 1 and 2, one can calculate that family #121 spent 459 days in shelters during its first eight stays in Table 1 (or 1st three episodes in Table 2), but one could not derive this from the episode data on Table 3. This is disadvantageous when it is necessary to keep a precise count of shelter nights consumed.

[4] See, for example, Metraux & Culhane 1999; Culhane & Kuhn 1998; Piliavin, Wright, Mare, and Westervelt 1996; Koegel & Burnam 1994.

Differences in exit length criteria can produce different results. Compared to the 30-day exit criterion used for Table 3, a more stringent exit criterion (such as the 1-day criterion used in Table 2) will lead to an increased number of episodes. Conversely, a more extended exit criterion, such as 6 months, would reduce family #111's two shelter episodes (Table 3) to one episode. Different exit criteria can also produce differing results in more complex data analyses, as is shown in Culhane and Kuhn (1998). In this study, Culhane and Kuhn look at what factors are associated with exiting a shelter episode using two different logistic regression models, one model using a 1-day and the other using a 30-day exit criterion. In assessing the likelihood of exit among female substance abusers, they find that:

> [U]nder a 1-day gap, women with substance abuse problems are 16.1 percent *more* [emphasis original] likely to exit on a given day. This effect, however, becomes insignificant when exits are defined by 30 days, suggesting that, while women with substance abuse problems do leave shelter more quickly than others, they are returning to the shelter more quickly as well. (34)

Episodes using 1-day exit criterion can also be compared to episodes using longer exit criteria. As was previously mentioned, using the 1-day exit criterion can provide a total number of shelter days used, such as the 459 days for family #121's first eight shelter stays (see Table 1). Using a 30-day criteria (see Table 3), these stays collapse into one episode spanning 497 days. Not all the days in this latter episode were spent in shelters. The degree to which a household spent time in shelters during their episode can be derived through dividing the 1-day exit total of 459 days by the 30-day exit total of 497 days. This yields a proportion of .924, meaning that family #121 spent 92% of their first shelter episode, under the 30 day exit criterion, actually staying in a shelter.

This brings up an important distinction of which to be aware: the difference between a shelter episode and a homeless episode. Usually data, such as what is presented here, will only be able to measure shelter episodes. One of the rationales for using extended exit criteria (exit criteria that are greater than 1-day) is that, when a household leaves a shelter for a short period of time, even though they left the shelter, they may still remain homeless. Thus episodes with extended exit criteria can be better said to measure a "homeless episode," rather than shelter episodes. The "shelter-to-episode duration ratio" then becomes a measure of how concurrent shelter and homeless episodes relate to each other.

Different ways to group stays into episodes are discussed in this section; the way which is most preferable to structure episodes will depend on the circumstances and reporting needs specific to a situation. In our experience, a one-day exit criterion is often more useful from a system management or administrative perspective, where measuring and managing daily bed supply and costs is of primary concern. Alternatively, a 30 day exit criterion is often more useful from a research perspective, where understanding the dynamics of homelessness, and not discrete periods of shelter use, is of primary concern. Converting stay into episodes, regardless of the specific criteria used to do so, provides a standardized measure of length of time spent in a shelter, and thus facilitates comparisons of shelter use across different shelters. Different ways of structuring episodes often, however, lead to different results when reporting episode-related statistics, results that can be differentiated into shelter and homeless episodes. A shelter episode is always a homeless episode, but this is not necessarily so *vice-versa*. Finally, as has been stressed throughout this section, notwithstanding the range of options that are available to group

stays into episodes, it is always necessary to be explicit in explaining how one groups stays into episodes.

II. Frequency Distributions and Measures of Central Tendency

Having shown ways to convert sets of stays into a uniform set of episodes, the next step is to go over ways of presenting basic measures of central tendency and frequency distributions that can be used to describe and compare aggregate groups of stays. Included here are such basic statistical measures as mean and median that are familiar to even those persons who are generally averse to statistics, as well as frequency distributions that provide material for easily understandable figures and graphs.

Statistics are ways of transcending individual observations and to obtain an understanding of a more general phenomenon such as shelter stays. The primary strength of basic statistics such as mean and median is the simplicity with which they describe such aggregates. One number can, in this case, describe a whole set of shelter stays. It can be assumed that lay readers of reports concerning shelter utilization are familiar with such statistics and what they purport to measure, and their general accessibility also makes them more amenable for dissemination through the mainstream media.

Measures of Central Tendency

The measures that will be covered here are the mean, the median, quartiles, and the standard deviation. Basic overviews of these statistics will be provided here; should the reader wish further information he or she can consult any one of a number of introductory statistics texts.

The *mean*, or arithmetic average, is computed by adding together all the values of x (in this case number of days stayed) in a group of observations and then dividing by n, the number of observations. The formula is

$$\text{Mean} = (1/n)(x_1 + x_2 + x_3 + \ldots + x_n).$$

The main weakness of the mean is that it is sensitive to outlying observations. This means that a few extreme measures (such as shelter stays exceeding 2 years in length) can have a disproportionate effect on the mean, skewing it in the direction of a higher value. Thus while the mean is a valid measure of central tendency, outlying observations may affect the mean more than they might affect other measures of central tendency.

The *median* is the midpoint, or the middle value of a group of observations when the observations are in sequential order for the variable that is being measured. The observation that serves as the median value is found by counting $(n + 1)/2$ observations[5] up from the bottom when there is an odd number of observations (the 50^{th} percentile value), or taking the mean of the values for the two centered observations when there is an even number of observations. This measure is less sensitive to outlying values than the mean.

[5] "*n*" again is the total number of observations in the set.

Quartiles reflect the values that separate a set into four groups with equal numbers of observations. In a sequentially ordered dataset, the first quartile would represent the value for the observation that separates the first 25% of observations from the rest of the set; the second quartile would be equivalent to that between the 25the percentile and the median; and the third quartile would be the value of the observation that separates the last 25% of the observations from the rest of the set. In another way of looking at this, the median (or second quartile) separates a set into two groups of observations, and the first and third quartiles are the medians of the two resulting groups. Quartiles give some indication of how spread out the measures of a variable are, and whether or not the values are evenly distributed.

Standard Deviation, the last measure of central tendency considered here, is another measure of spread, this time measuring the degree by which a group of measurements is spread around the mean. The formula for this is more unwieldy than that of the previous measures discussed. For the variable x measured in n observations, the standard deviation, or s, is the square root of the variance, or s^2, which is derived:

$$s^2 = [1/(n\text{-}1)] * [(x_1\text{-}X)^2 + (x_2\text{-}X)^2 + \ldots + (x_n\text{-}X)^2]$$

where X equals the mean of x.

Table 4 illustrates the use of such statistics, again using data from the DHS family shelter stay database. In this table, the episodes in the DHS database have been divided into three sets, by exit dates in 1994, 1995, and 1996. For each of these three sets, statistics are reported using both 1-day and 30-day exit criteria, as outlined in the previous section.[6] Thus Table 4 enables comparisons both across years and between two different criteria for determining exits. On the table, "Total Episodes" represents all episodes that ended in the given year and "Total Days" is the sum of days covered in these episodes.[7] The other statistics are derived as described in this section.

Briefly interpreting the results on Table 4 shows a large degree of fluctuation in most of the statistics across years. Looking at the 1-Day Criterion results, total episodes declined in 1995 and increased in 1996, while total days increased in 1995 but decreased in 1996. These two trends are borne out in the mean and the median episode lengths, which were much higher in 1995 than in either 1994 or 1996. Consistent with a higher median, 1995 also has higher quartile levels, again suggesting that while the number of episodes decreased in this year, the length of stays for the 1995 episodes increased across the board, among both the proportionately shorter and longer staying clients. As a proportion of the mean, the standard deviation for 1995 is lower than the other two years, suggesting that proportionately there was less spread among the episode lengths in that year. Finally, the difference between mean and median episode lengths suggests that the longer stays skew the mean substantially in all of the years except 1995.

---

[6] I will use SAS version 6.12 for the data analysis, although the statistics and frequencies in this section can be computed using other statistical software programs or spreadsheet programs.

[7] Days covered is computed (Start Date – End Date) + 1.

Table 4 – Measures of Central Tendency for Episodes Ending in 1994-1996 Using 1-Day And 30-Day Exit Criteria

	1-Day Exit Criterion	30-Day Exit Criterion
Stays Ending In: S		

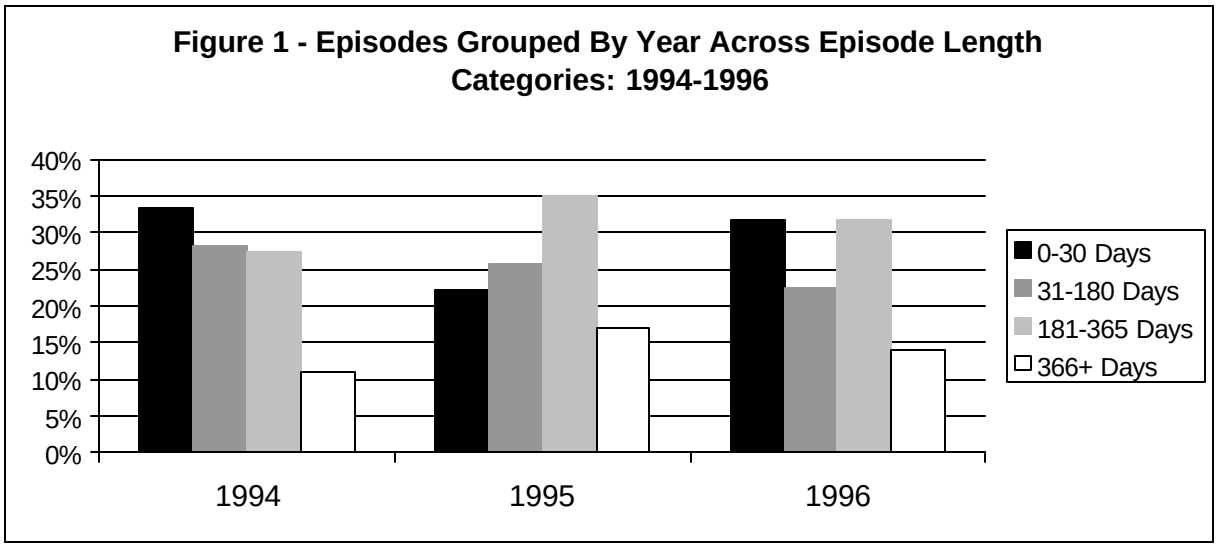


Table 5 reports frequency distributions for the sets of stays used in Table 4 (using 1-Day exit criterion). The groups are set at 1-30 days, 31-180 days, 181-365 days, and more than 365 days. The results show that in 1994, the proportions of shorter episodes fell while the

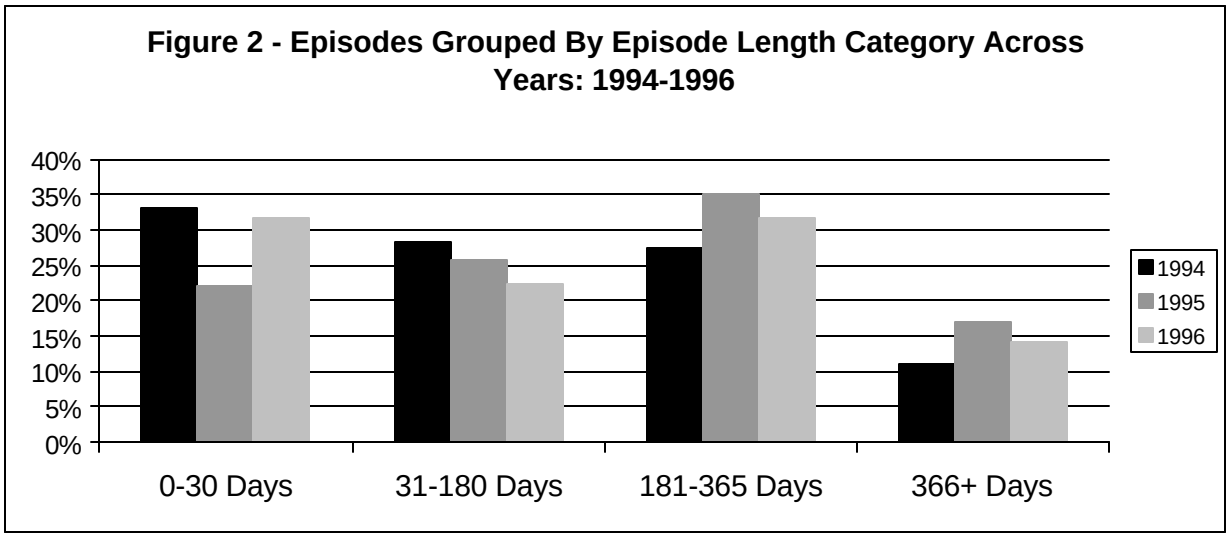
proportions of longer episodes increased. The sharp fall in the "0-30 Days" category helps explain why the mean and median episode lengths, shown on Table 4, are more similar in 1995 than in the other two years.

These results can be better displayed graphically, and two ways in which these results can be graphed are shown in Figures 1 and 2. For Figure 1, the episode length categories are grouped together for each year to reveal substantially different distributions of episodes among these categories within each of the three years. Figure 2 changes the display around for these results, as the years are grouped together by episode length categories. This facilitates a comparison of category size across years. There are clearly differences over these three years, although it is hard to identify any possible trends.

Measures of central tendency and frequency distributions provide straightforward, easy to understand ways to communicate basic trends to persons without resorting to specialized statistical knowledge or complex computational methods. In this section, examples have been given of how the most commonly used measures might be used for describing aggregated groups of shelter episodes, and changes in these groups of episodes over time. Different needs, different questions, and different perspectives will surely lead to different adaptations of these examples, and to different ways for displaying results using this class of statistics.

III. Survival Curves and Hazard Curves

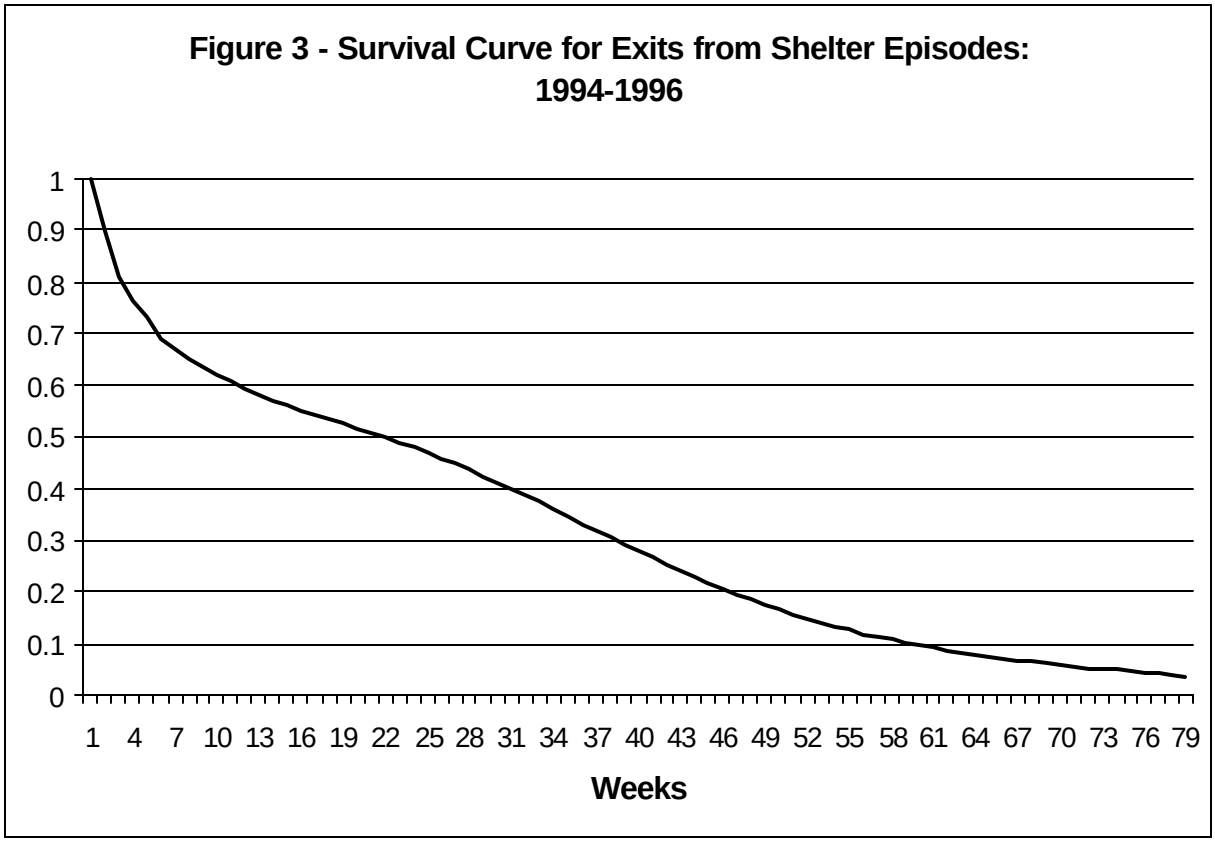
Survival curves and hazard curves are two additional ways to show characteristics and trends in a group of shelter episodes. Both of these curves are somewhat intuitive and understandable to a layperson, yet they are not as well known as the statistics discussed in section II, and thus will



require a degree of explanation when presenting them. Hazard and survival curves are relatively simple to compute, but require statistical software, such as SAS or SPSS to do so. This chapter will describe survival curves and hazard curves, and will offer examples for applying these methods to episode data using the same datasets used in section II.

Both survival curves and hazard curves are ways of expressing the relationships of observations in a dataset to a specific type of event. An event here is defined as an occurrence that can be measured in time. One example of an event would be an episode exit, with time measured as the number of days from episode entry until an exit occurs. Time until an event, as measured here, would be the same as length of episode stay, which was looked at in section 2. Another example of an event would be shelter reentry, with the event here being a repeat shelter episode and the time measured being the length from an episode exit to subsequent shelter reentry. This second example is different from the first example insofar as, in the first example, all the episodes with exits in 1994 through 1996 have, by definition, an event (i.e., a shelter exit) whereas all these episodes are not necessarily followed by a subsequent shelter episode (i.e., many households don't return for another shelter stay). In statistical parlance, such episodes that do not have events by a certain time period are referred to as "censored" observations.

Survival curves measure the percentage of observations that "survive," or that have yet to experience an event, at a given point in time. This will be further explained with an example. Figure 3 is a survival curve where exit from shelter (using 1-day exit criterion) constitutes the event and the dataset is the combined sets of episodes for 1994 through 1996. The x-axis signifies episode length in weeks and the y-axis shows the percentage of episodes that have not ended by the end of that time period. The survival curve starts with 100% of the episodes in the shelter (as all episodes, by definition, have "survived" the first week), and the percentage still in the shelter declines with each passing week as more episodes end, until week 79 (1½ years from the episode start), where only 4% of the episodes have yet not ended.



From this figure it is possible to get a better idea of the distribution of the duration of episodes. At first the survival curve descends somewhat sharply, showing roughly 30% of the episodes to be shorter than six weeks. On the other end of the curve, which is considerably flatter, 15% of the episodes last longer than a year. Most of the episodes are over by the 22-week mark, and those that are left continue to leave, but at a slower rate and to where there remains a small core group of long-term episodes. Segments of this curve, for example the 30% of the episodes which make up the shortest episodes, or the 15% of the episodes which are the longest episodes, can serve as the basis for forming subgroups which may merit further study or specialized intervention.

Variations of this figure can break down the episodes used here into different subgroups. Examples might include three curves on one graph signifying the three different years in which the episode exits occurred, or two curves which compare episode exits by gender. This can be used to determine different shelter use patterns among different subgroups.

Hazard curves, somewhat less intuitive than survival curves, offer another way to look at event occurrences. Hazard curves measure the conditional hazard rate, or the expected number of events per one-unit interval of time for all observations that have not yet experienced an event. It is defined in terms of the probability that a specific event will happen at a specific time interval. Hazard rates change over time, and a higher hazard rate represents a greater risk that an event will occur to an individual observation during the specified time interval (Allison 1995).

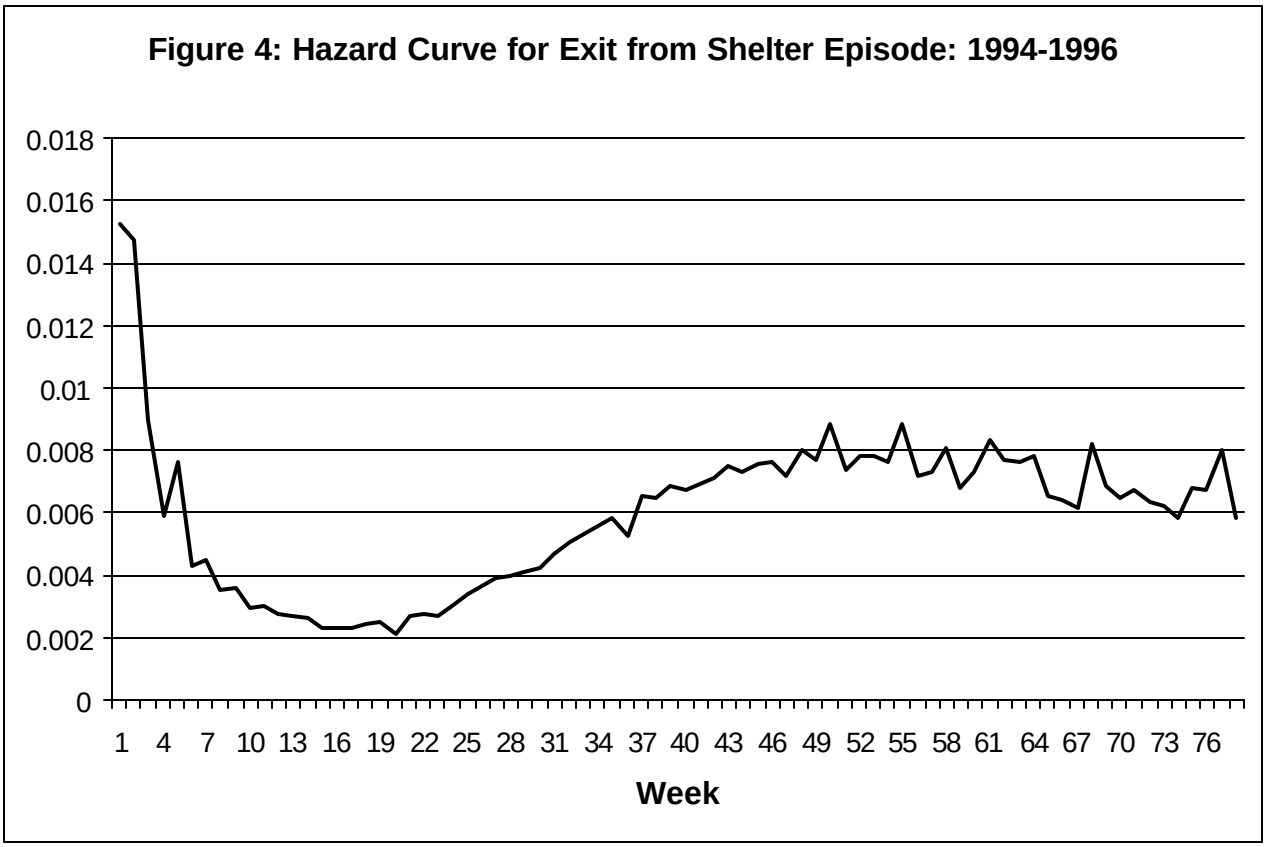


Figure 4 shows the hazard curve for exiting shelters (using 1-day exit criterion) over the same time period as the survival curve in Figure 3. It shows that the greatest hazards for shelter exit occur right after the episode begins, with the hazard falling sharply and then bottoming out at around the twentieth week, whereafter the hazard of exiting starts to increase again. One way to interpret this pattern is that persons are at high risk to exit early from a shelter episode as they are more likely then to make other arrangements, if available, to avoid a prolonged stay. In the absence of making such arrangements, households often stay in the shelter and leave upon receiving a subsidized housing placement, which can, in New York City, occur at any point after 90 days (around 13 weeks) but typically occurs about 9 months to a year into the episode (Metraux 1999). Looking at figure 4, the hazard for exit starts a prolonged increase, after its initial decline, at around the six-month (24 weeks) point, suggesting that housing placements start becoming available a little before the nine month mark.

The main advantage to survival and hazard curves is that they show near continuous changes for the event of interest over time, in contrast to the cruder time breakdowns found in the measures of central tendency and the frequency distributions described earlier. As mentioned earlier, the same hazard and survival methods can be applied to other events such as shelter reentry. In other words, these would be the preferred means for examining the distribution in the time periods between repeat shelter episodes or stays. These measures can also be computed and displayed separately by year, to examine trends in the distribution from one year to the next. Similarly, one can compute these measures for different populations of shelter users, i.e. families versus singles, men versus women, substance users versus non-substance users. Again, note that

applying these curves to episodes using a 30-day exit criterion would produce different results. Moreover, it would require making a decision rule as to whether to use the sheltered days as the basis for the distribution, or the entire homeless episode days (sheltered and unsheltered).

IV. Regression Techniques

Hazard curves and survival curves belong in a category of methods, known collectively as survival analysis or event history analysis, that describe an event while taking time into consideration. Included among survival analysis methods are several different regression techniques, which are used to estimate the effect of different factors upon an event over time. Thus, given the data that we have been using, it would be possible, using regression techniques, to see if a factor such as, for example, having a greater number of children in a household would increase the likelihood of that household having a longer shelter stay. Survival analysis regression can also accommodate censoring, where not all observations experience events, such as would occur in estimating the effect of different factors on the likelihood of experiencing a subsequent shelter episode after exiting a shelter episode.

Regression techniques are powerful means by which to examine what factors contribute to a particular outcome. Not only can these techniques be used to estimate the effects of a particular factor (covariate) of interest on an outcome (dependent variable), it can estimate these effects while holding constant, or "controlling for," other covariates that may distort the effects of the relationship between the dependent variable and the covariate of interest.

Regression techniques are, however, considerably more sophisticated than the methods and statistics discussed so far in this paper. Some training is required to interpret the results, often rendering analyses using regression techniques of limited use for some audiences. Many major statistical software packages can perform survival analysis regressions, however a considerable degree of familiarity with these methods is required before one can perform valid analyses, thus limiting the persons who can effectively use regression techniques.[9] The detailed nature of the implementation of these techniques goes beyond the scope of this paper, and instead a limited discussion of two studies is provided, both of which use survival analysis regressions on shelter episode data.

Culhane and Kuhn (1998), as previously mentioned, use shelter episode data to identify the characteristics that predict an exit from shelter episode in the single adult shelter systems in New York City and in Philadelphia. In addition to episode information, they also attach data to the episode that reflect characteristics of the persons who experience the shelter episode: race, sex, age, health and mental health status, and substance use. Using a survival analysis method known as discrete-time logistic regression, they find that, in general, being older, of black race, having a substance abuse or mental health problem, or having a physical disability significantly reduces the likelihood of exiting shelter. The authors show how these results can be used as a basis for targeting specific intervention programs to persons with specific characteristics with the goal of shortening long shelter stays and preventing quick returns to shelter upon exit.

[9] See Allison (1995) for a text on using an array of survival analysis techniques, including regressions.

Metraux and Culhane (1999) examine, using DHS shelter data, family shelter episodes of 8,030 women and their households, and the single adult shelter episodes of 2,444 women, all of whose shelter episodes ended in 1992, to see what factors increased the risk of subsequent repeat shelter episodes by these women. Using a survival analysis method known as Cox regression,[10] particular attention was focused on family characteristics associated with the women during their shelter stays and on stay outcomes – where the women went after they exited their 1992 shelter episodes. Among the findings were that exits to one's own housing, either with government subsidized rents or on the private market, were significantly associated with large decreases in the risk of experiencing subsequent shelter episodes. Certain family characteristics, on the other hand, such as being pregnant or having young children upon the beginning of the shelter episode, being the sole adult in the household, and having a history of domestic violence all increased the likelihood of a subsequent shelter stay. Again there are notable policy implications for shelters here, as providing housing prevents future shelter stays and that family characteristics that are associated with a greater vulnerability for becoming homeless also are associated with repeated shelter episodes and thus a longer homeless "career."

Further reading of these articles would reveal additional details on the methods used, the results, and the policy considerations for shelters borne from these results. It is also instructive, to see how the regression models build on other techniques that we have covered in this paper already. These summaries nonetheless give an idea of the ability of these techniques to integrate many facets of a dataset into one comprehensive analysis. However, while their uses are apparent, such analyses must be used carefully, and efforts must taken to ensure that the results are disseminated clearly and in a way that is understandable when one's audience includes the lay public.

V. Population Segmentation by Stay Pattern

One advantage of the survival and hazard curves described above is that one can compare the curves for various population groups, enabling one, for example, to compare men and women, or households with children and those without children. The resulting distributions are easily displayed and, with some introductory explanation, can be readily interpreted by lay readers or observers. In this way, an analyst can define groups of interest beforehand, or *a priori*, and apply some standard techniques for comparing the distributions of their shelter stays.

Alternatively, an analyst may be more interested in doing some more "exploratory" data analysis, whereby population groups or segments are identified based on the observed distributions in the data, or making *post h*oc comparisons. For example, again using the survival curve in Figure 3, an analyst may observe that there are at least two potentially interesting groups that one may want to learn more about, namely those who exit the shelters quickly (in fewer than 4 weeks), and those who are still there after 70 weeks in the system. Accordingly, an analyst may select the groups whose stays fit the pattern of interest (the short or long stayers), and do a descriptive analysis of their characteristics (age, race, sex, disability status) to better understand who they are, and potential reasons for their pattern of stay. Again, the survival curves are useful for informing such selection or segmentation decisions, and can be quite informative for program and policy analysis.

---

[10] The formal name for this method is Cox proportional hazards model.

In this section, two exploratory data analysis approaches to population segmentation by shelter stay pattern are described. The first, a "heavy user" approach, is the simpler, and is accessible to both relatively novice analysts and to a general audience. The second, cluster analysis, is a more complex multivariate statistical technique, but can permit analysts to obtain a more textured profile of shelter stay patterns.

Heavy User Analysis

The "heavy user" analysis is fairly straightforward. Essentially, it is a cumulative measure of the proportion of shelter days used by a fixed proportion of shelter users, or, conversely, the cumulative proportion of shelter users who consume a fixed proportion of shelter days. In order to do this, first the number of days stayed for each individual shelter user is calculated. Next, the shelter users are grouped by such increments as quintiles, deciles or even in 5% increments of persons, and the proportion of the total user group that this subgroup represents is compared to the proportion of total days stayed this subgroup consumes. This provides a simple means of selecting a percentage of clients (the heaviest users defined, for example, as the top 5% or top 10%), and identifying their proportionate use of system days. Likewise, one could select some proportion of days used (for example, 25% of the days used), and find the corresponding percentile of the population that is using those days. A hypothetical finding using such a method might be that 5% of the heaviest shelter users consume 25% of the total shelter days, or that 50% of the shelter days are consumed by 20% of the sheltered households. Once the criteria for "heavy use" or "heavy users" has been selected, one can conduct studies of the characteristics associated with the heavy users, to try to understand how they differ from other groups of shelter users.

This technique has been applied in the area of mental health services research, particularly with the aim of identifying the population segment responsible for using the bulk of state hospital inpatient days. Other public systems, such as jails and general hospitals, similarly conduct such analyses to identify the proportion of people whose demands on the system are greatest. Because "heavy users," by definition, use a highly disproportionate amount of the system days, they are often a compelling target for system managers and policymakers who would like to divert such persons to alternative settings, where costs can be better managed, or the people more appropriately served.

Such would be the case in the area of homeless services. Conducting such an analysis could identify the proportion of persons who, from a shelter manager's point of view, are using the shelter for purposes contrary to the intended "emergency" function of the shelter system. Interventions for this group could then be considered so as to provide more appropriate long-term housing for this group, as well as to reduce overall demand on the shelter system. It is worth noting, however, that "heavy use" is an arbitrary denotation, and that the very nature of a given distribution of shelter stays is likely to show a longer-staying group whose use of system days would be characterized as proportionately "heavy," though their stays may still be considered of an appropriate and expected duration, at least for some segment of the population.

Table 6 - Differentials in Male Shelter Users Population in New York City: 1993-1996

Year	% of Men Who Use 25% of Annual Shelter Resources	% of Men Who Use 50% of Annual Shelter Resources	Median # of Days per Person Used Over the Year	% of Men w/ 180 or More Days in Shelter	% of Men w/ 30 or Less Days in Shelter
1993	6.0%	14.2%	39	17.5%	58.5%
1994	5.7%	13.5%	34	15.9%	60.5%
1995	6.8%	16.0%	54	20.3%	52.7%
1996	6.9%	15.9%	54	21.2%	52.3%

To illustrate different heavy users techniques, we again turn to data from New York City. In table 6, the first two columns show trends in the proportion of the heaviest shelter users among single males who, combined, consume a fixed proportion of shelter resources (i.e., shelter days). The last two columns, in a slightly different measurement, show trends in the proportion of male shelter users who consumed more/less than a fixed number of shelter days in the given year. The center column shows that, over the four year period, changes in the median number of per capita days spent annually in men's shelters fluctuates consistently with the proportion of men who spend over half the year in shelters and inversely with the proportion that spend one month or less of the year in shelters. Looking at the first two columns, as the median stay contracts, so does the proportion of persons who are among the heaviest users. All in all, both measures show that as the median stay increased, so did the proportion of persons designated as heavy users, and when the median stay is shorter, a smaller proportion of the shelter population make up the heavy user group.

This table must be interpreted cautiously, however, as this trend toward increasing proportions of heavy users says nothing about changes in the size of the total shelter population. The proportional increases in long-term stayers noted in 1995 and 1996 could conceivably come from an overall decrease in shelter census, in which case the majority of the population decrease came from the ranks of the short-term stayers. On the other hand, longer shelter stays may just as well accompany an increased number of shelter users, leading to a situation where increased numbers of heavy users is consistent with an overall increase in the shelter population.

More information on heavy and light shelter users is available from Tables 7a and 7b, which again set fixed stay criterion and shows trends, this time both in the number *and* the proportion of persons in the male shelter population who meet each table's criterion. Over 1995 and 1996, both the relative proportion and the overall number of light users in the shelter system decreased and the number and proportion of heavy users increased, both trends in concert an increased rate of median per capita shelter use over these two years (see Table 6). Tables 7a and 7b also illustrate fundamental differences in the "case mix" in the heavy and light users with respect to age and identified mental illness, substance abuse problems, and medical problems.[11]

[11] These three indicators are not absolute measures of mental illness, substance abuse, or physical health among this group of sheltered males, as these indicators are based on interviewer assessment upon intake and are inconsistently applied. Again, caution must be exercised in making conclusions from these examples.

Table 7a - Heavy Male Shelter User Characteristics (stayed 300 days or more in shelters during given year)

Year	Number	% of Total Population	Mental Illness	Substance Abuse	Medical Problems	Median Age
1993	1626	7.4%	9.0%	39.7%	26.0%	41.4
1994	1502	6.7%	9.2%	36.7%	28.2%	43.3
1995	1673	8.5%	7.3%	38.0%	25.3%	43.5
1996	1970	9.3%	6.7%	38.7%	24.9%	43.8

Table 7b- Short Term Male Shelter User Characteristics (stayed 30 days or less in shelter during given year)

Year	Number	% of Total Population	Mental Illness	Substance Abuse	Medical Problems	Median Age
1993	12920	58.5%	5.8%	36.0%	18.4%	35.9
1994	13491	60.5%	5.4%	36.2%	17.9%	36.5
1995	10315	52.7%	6.6%	36.0%	17.6%	36.6
1996	11052	52.3%	6.4%	33.2%	18.2%	36.9

Cluster Analysis

Cluster analysis is a technique that provides the means for parsing of a set of observations into two or more subgroups on the basis of one or more quantitative variables. Where, in heavy user analysis, the researcher creates groups based on one measure, in cluster analysis the observations can be grouped based on multiple variables so that each observation gets placed in the one cluster where the other observations are most similar to it based on the values of the determining variables. The resulting clusters, or subgroups, can then be tested to determine whether or not they are sufficiently different from each other so that the groupings have practical meaning, and the clusters can be labeled on the basis of the similar characteristics.[12] Like survival curves, cluster analysis requires appropriate statistical software and a basic degree of statistical knowledge, and, with some explanation, cluster analysis results are accessible to a broad audience.

In this section we will show how, using cluster analysis, the researcher can utilize the basic stay measures discussed in this paper – days stayed and discrete episodes – to provide a more textured characterization of stays than the other methods we have examined so far. This texture becomes apparent upon revisiting the discussion that accompanies Figure 3, where it was suggested that the survival curve showed two important population groups – short-term stayers and long-term stayers. However the survival curve only measured the rates of exit based on the total number of shelter days per episode; it could not consider that some persons are likely to have experienced multiple, discrete episodes. Thus, these methods did not permit the examination of the degree to which there were long-term stayers whose shelter tenure was drawn out over several episodes. This would represent a second group of long-term stayers, those who enter and exit frequently but who nevertheless accrue many shelter days. The shelter use pattern

[12] In a facetious illustration of this, if one had a dataset in which information on criminal history and physical features were available, one could conceivably use cluster analysis to create three groups that one might label the good, the bad, and the ugly.

Table 8 - Cluster Statistics, Demographics, and Treatment Variables in a Model for New York City Single Adult Shelter System Users

	Transitional	Episodic	Chronic	Total
Summary Statistics:				
Number of Clients	59367	6700	7196	73263
Avg. # of Episodes	1.4	4.9	2.3	1.8
Avg. # of Days	57.8	263.8	637.8	133.6
Avg. Days per Episode	42.4	54.4	280.9	75.4
Percent of Client Days Used	35.1%	18.1%	46.9%	100.0%
Percent of Clients	81.0%	9.1%	9.8%	100.0%
Ratio (%Days / %Clients)	43.0%	197.0%	477.0%	100.0%
Demographics:				
Percent White	11.9%	6.1%	9.5%	11.1%
Percent Male	81.5%	81.8%	82.3%	81.6%
Percent Under 30	36.1%	37.7%	23.2%	35.0%
Percent Over 50	8.3%	6.3%	13.9%	8.7%
Treatment Variables:				
Mental Illness	6.5%	11.8%	15.1%	7.8%
Medical	14.2%	19.8%	24.0%	15.7%
Substance Abuse	28.2%	40.0%	37.9%	30.2%
All Three	1.3%	3.0%	3.3%	1.7%

Source: Table from Culhane, Metraux, and Wachter (1999), as adapted from Kuhn and Culhane (1996)

of this group would be different from both the group represented by those with long shelter episodes and the group represented by those with one-time short-term episodes.

Recognizing that this may be an important distinction among the shelter user population, Kuhn and Culhane (1998) applied cluster analysis as an alternative means of specifying stay patterns. In their study of single-adult shelter users in Philadelphia and New York City, they used two shelter utilization measures, number of days and the number of episodes that each individual spent in shelters, to create three clusters. They hypothesized that these three clusters would correspond to the three types of shelter use patterns just described: transitional homelessness, characterized by a single, relatively short-lived period of homelessness; episodic homelessness, whereby persons "drift" in and out of relatively short periods of homelessness, and chronic homelessness, in which persons are homeless for extended, uninterrupted periods of time.[13]

The results of Kuhn and Culhane's cluster procedure did indeed yield three clusters that corresponded to the typology that was described. As shown on Table 8, the three groups not only showed distinct patterns of shelter use, but they also differed in demographic characteristics and incidence of physical and psychological morbidities. While these results were consistent across Philadelphia and New York City, the results will likely vary across localities, and the

[13] Kuhn and Culhane give an extensive overview of research literature that gives theoretical support for typologies of homelessness based on transitional, episodic and chronic patterns. See, among others, Fischer & Breakey (1986); Snow & Anderson (1987); Hopper (1989); and Sosin *et al*. (1990).

distinctiveness of one group from another in terms of stay pattern may be more or less apparent. For example, it may be, for a given locality, that a two-cluster solution is more appropriate, as there is either too little episodic or too little chronic homelessness for this cluster to be sufficiently distinct.

A final advantage of the cluster analysis strategy, similar to that of the "heavy user" approach above, is that, for a given cluster, one can measure the proportion of users and days used by a given cluster. This may be particularly useful for managers who would wish to use such an analysis for reallocating existing resources more efficiently, or for targeting programs intended to reach a particular cluster. For example, the results of the Kuhn and Culhane analysis (Table 8) showed that 10% of the shelter users (the chronic cluster) used 50% of the shelter days, while 80% of the users (the transitional cluster) used about 15% of the days.[14] In terms of policy ramifications, this suggests that the chronic users might be more effectively served by alternative housing programs that get them out of shelters, while the transitional cluster might be the target for prevention programs aimed at keeping them out of shelters altogether. Here cluster analysis provides a means that permits shelter managers to hone specific interventions for specific subgroups of the shelter population, and it offers the promise that interventions with a relatively small group of shelter users has the potential to drastically reduce overall demand for shelter resources.

Conclusion

This paper has focused on methods whereby shelter use data, records of stays and days spent in shelters, can be used to generate statistics that describe different dynamics of shelter utilization. These methods are also designed to be practical in that they provide a means for analysis that contribute to ongoing dialogue on providing more effective shelter services. In outlining these methods, we have attempted to keep our description of the methods general enough so they can be adapted to an array of data collection setups and local circumstances. In addition, hopefully the examples provided, relying primarily on data collected from the New York City shelter system, will facilitate ideas whereby these methods can be successfully adapted elsewhere. This is, of course, far from the final word on this topic, nor does this paper represent a comprehensive set of methods whereby such data can be analyzed. Rather it represents a starting from which the reader can embellish, improvise, and perhaps develop other methods of analysis that supplement those described here.

References

Allison, PA. 1995. *Survival analysis using the SAS® System: A practical guide*. Cary NC: SAS Institute.

Cordray DS & GM Pion. 1991. What's behind the numbers? Definitional issues in counting the homeless. *Housing Policy Debate* 2(3): 587-616.

[14] A study in Columbus, Ohio (Scioto Peninsula Relocation Task Force 1998) found a similar pattern of shelter use among three clusters with similar shelter use patterns.

Culhane, DP and RS Kuhn. 1998. Patterns and determinants of public shelter utilization among homeless adults in New York City and Philadelphia. *Journal of Policy Analysis and Management* 17(1): 23-43.

Culhane, DP, S Metraux and SM Wachter. 1999. Homelessness and public shelter provision in New York City. In MH Schill, ed. *Housing and community development in New York City: Facing the future*. Albany NY: SUNY Press.

Fischer, P and W Breakey (1986). Homelessness and mental health: An overview. *International Journal of Mental Health* 14(4): 6-41.

Hopper, K. 1989. Deviance and dwelling space: Notes on the resettlement of homeless persons with alcohol and drug problems. *Contemporary Drug Problems* 16: 391-414.

Koegel, P and MA Burnam. 1994. The course of homelessness among homeless adults in Los Angeles. Annual meeting of the American Public Health Association, Washington DC, October.

Kuhn, RS and DP Culhane. 1998. Applying cluster analysis to test a typology of homelessness by pattern of shelter utilization: Results from the analysis of administrative data. *American Journal of Community Psychology* 26(2): 207-232.

Metraux, S. 1999. Shelter and housing: Relationships between stay outcomes and individual, family, and shelter stay characteristics among women in New York City family homeless shelters. Presented at the American Sociological Association Annual Meeting. Chicago IL, August.

Metraux, S and DP Culhane. 1999. Family dynamics, housing, and recurrent homelessness among women in New York City homeless shelters. *Journal of Family Issues* 20(3): 371-396.

Piliavin, I, BR Wright, RD Mare and AH Westerfelt. 1996. Exits and returns to homelessness. *Social Service Review* 70(1): 33-57.

Scioto Peninsula Relocation Task Force. 1998. *Rebuilding Lives: A New Strategy to House Homeless Men.* Columbus OH: Community Shelter Board.

Snow D, and L Anderson. 1987. Identity work among the homeless: The verbal construction and avowal of personal identities. *American Journal of Sociology* 18: 129-160.

Sosin, M, I Piliavin and H Westerfelt. 1990. Toward a longitudinal analysis of homelessness. *Journal of Social Issues* 46(4): 157-174.