Traffic Distribution Fitting: A Systematic Methodology

RUSSELL THOMPSON AND WILLIAM YOUNG

Traffic distributions are an integral part of many traffic analyses. The correct determination of these distributions is therefore essential. This paper presents a methodology for the determination of the appropriate distribution. The methodology incorporates the data investigation, model selection, parameter estimation, and goodness-of-fit stages in distributions fitting. Many traffic analysts pay scant attention to the first two steps. The methodology is formulated within a microcomputer program, TRANSTAT, that has been developed to determine appropriate traffic distributions. TRANSTAT incorporates many recent developments in exploratory data analyses, expert systems, and outlier analyses. The paper describes the program and illustrates its application to the determination of parking duration curves.

Traffic engineering data analysis often requires the investigation of statistical distributions (*1*). For instance, distributions are needed for traffic simulation, checking of distributional assumptions in statistical analysis, and the determination of outliers (non-typical observations, inconsistent with the other values in the sample). This paper emphasizes a methodological approach for determining appropriate distributions.

The probability distribution fitting process can be thought of as a set of interconnected steps. The steps in the process are:

- Data collection,
- Data investigation,
- Model selection,
- Parameter estimation,
- Goodness-of-fit determination, and
- Application.

It should be noted that curve fitting is an iterative process. It is a search of possible alternatives for an appropriate explanation of the data. The backward link between goodness-of-fit testing and data investigation should continue until a "representative" model has been found. The data investigation and model selection stages of this process are often given scant attention, with analysts moving straight from data collection to parameter estimation. This paper emphasizes the need to consider these steps fully.

The overlooking of the data investigation and model selection stage often results from the time needed to prepare graphs and plots of the data. The formalization of the methodology, therefore, owes much to recent developments in microcomputers and computer graphics. These developments have created the rapid interactive environment that allows the data investigation and model selection stages to become an important part of the distribution fitting methodology. The discussion of the methodology is therefore couched in terms of a microcomputer package (TRANSTAT). TRANSTAT has been developed to fit common univariate distributions to traffic data. The need for such a package was identified largely from the inadequacy of the existing commercial statistical packages (*2*). Most packages did not include any facilities to fit distributions to data. Some microcomputer statistical packages do allow probability densities to be fitted (*e.g.*, STATGRAPHICS, SPSS/PC+). However, the treatment is usually very limited. TRANSTAT contains unique modules relating to outlier analysis, model guidance and diagnostic plots. Also, critical regions as well as test statistics are presented for all tests. These features render TRANSTAT superior compared with conventional analysis packages. TRANSTAT provides a comprehensive and user-friendly package to aid the investigation and modeling of statistical distributions.

This paper describes TRANSTAT, emphasizing the need to take a good look at the data before attempting model fitting, then illustrates its application to the determination of the distribution of parking duration times.

TRANSTAT OVERVIEW

TRANSTAT is written in the Microsoft's QuickBASIC computer-programming language and is designed to run on IBM PC-XT/AT microcomputers. Data input is via an ASCII file, and individual data values must be separated by at least one space. There is a data limitation of 2000 observations. The file input data facility allows data to be directly interfaced with automatic data collection devices enabling the transfer of data without manual intervention. This can speed up analysis considerably and allow observations to be measured with a high degree of accuracy. TRANSTAT is available from the Department of Civil Engineering, Monash University, Australia, for $100 (Aust.).

A conscious effort has been made to make TRANSTAT as efficient as possible. Calculation speed has been given high priority. For instance, when sorting data, the Quicksort routine has been used (*3*). This is accepted as being the fastest general sorting procedure available for computers.

Color graphics have been used extensively throughout TRANSTAT to enhance the modeling process. The package allows either the Enhanced Graphics Adapter (EGA) or the standard Color Graphics Adapter (CGA) to be used. EGA

Monash University, Clayton, Victoria, Australia 3168.

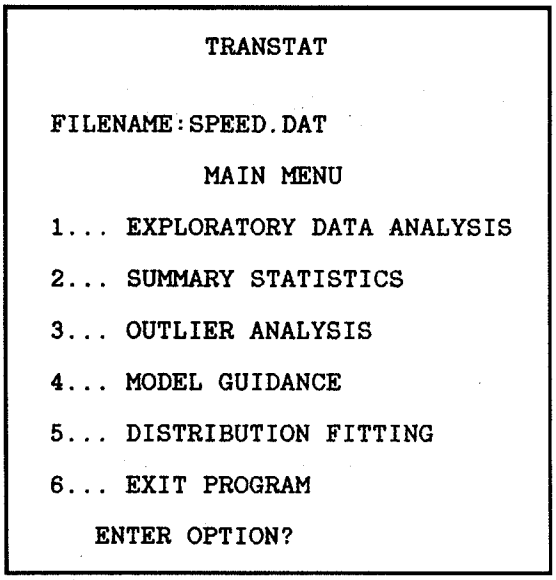


FIGURE 1 TRANSTAT's main menu.

graphics offer a higher resolution output with many more colors.

Interactive menus are used throughout the package in an attempt to make distribution modeling simple. The main menu (Figure 1), allows any one of five modules to be chosen. After selecting an option, another menu or prompt will be presented requesting information. The structure of the main menu is deliberately designed to encourage the distribution-fitting methodology introduced above, and hence good modeling practice, by the ordering of the options.

The following sections of the paper describe the options offered in the main menu. They discuss the application of exploratory data analysis, summary statistics, outlier analysis, model guidance, and distribution fitting (the Chi-squared test and the Kolmogorov-Smirnov test). The first two steps provide a strong indication of the type of distribution. The distribution fitting stage provides statistical measures of the degree of fit between the observed and the expected distributions.

Exploratory Data Analysis

Exploratory Data Analysis (EDA) techniques (*2*) are becoming very popular. These techniques seek to guide the analyst into the appropriate analysis procedure through the visual examination of data. Numerous plots of data can be used to illuminate its structure and subtleties. Tukey (*4*) states that EDA is "numerical detective work" and notes that its strength lies in forcing "us to notice what we never expected to see."

EDA should precede statistical inference and stochastic model-building. Its role is to enhance the analyst's knowledge of the data before any inference is attempted. It is recommended that several plots of the data be produced before any models are fitted. The EDA module appears first on TRANSTAT'S main menu to encourage its early use in the modeling process.

Although there are many standard types of EDA plots, EDA itself is not a set of well defined procedures. It consists of any plot or data analysis technique which attempts to illustrate the structure or subtleties of the data. The plots included in TRANSTAT are commonly used for investigating one dimensional data. The box, quantile, and jitter plots are available in TRANSTAT and provide a valuable tool for model identification. To illustrate the variety of plots the histogram, jitter, and box plots will be discussed.

The frequency histogram (Figure 2) provides an effective method of displaying frequency data. However, since it summarizes the data, it can produce a deceiving picture and care should be taken in its interpretation. This distortion can be caused by the length and number of the class intervals. TRANSTAT offers the user the option of specifying the class size and the start of the first interval to alleviate this problem. It is good practice to vary the class size in order to gain a varied picture of the data's distribution. In practice, about twenty classes over the data's range usually provide a reasonable picture of the distribution.

Chambers et al. (*5*) introduced an alternative to the frequency histogram, a one-dimensional scatter plot. This plot is called a jitter plot. Dots are used to represent observations. The dots are placed in the appropriate position on the major axes. This often results in loss of data due to point overlap. The overlap is reduced by spreading the points randomly on the vertical axis. The latter plot can be used to complement a frequency histogram by providing an undistorted picture of the data's local density. The data's range, density and symmetry can be easily seen from this plot. An example of the combined frequency and jitter plots for different TRANSTAT distributions is presented in Figure 2.

Another useful EDA plot is the box plot. Initially introduced by Tukey (*4*), this plot summarizes the data by plotting the upper and lower quartiles as well as the median. These quartiles are represented by the top and bottom horizontal lines of the rectangle with the median being portrayed by the line within the rectangle. Generally, vertical lines extend from the box to the minimum and maximum values. However, if a very large or small value falls outside the upper quartile plus or minus 1.5 times the inter-quartile range, it is highlighted by being plotted individually. The vertical lines then extend to the next highest or lowest value falling inside this range. This provides a simple method of identifying possible outliers, singling them out for further examination. This plot conveniently summarizes many distributional properties, including symmetry, spread, and range. Figure 3 presents box plots for the distributions available in TRANSTAT. Distribution types can be rejected on the basis of symmetry, range, etc., narrowing the possible model types to be considered.

Box plots can also be used to study numerous "batches" of data. Multiple box plots can also be used to partition sets of data into homogeneous classes. Figure 4 provides an example of the application of this type of plot to the presentation of parking duration data over time. Here, the skewness of the data suggests that the exponential or Erlang may be appropriate.

Summary Statistics

Summary statistics calculated from the sample are important in determining the type of distribution. Most distributions have a constrained range and other properties based on sample statistics.

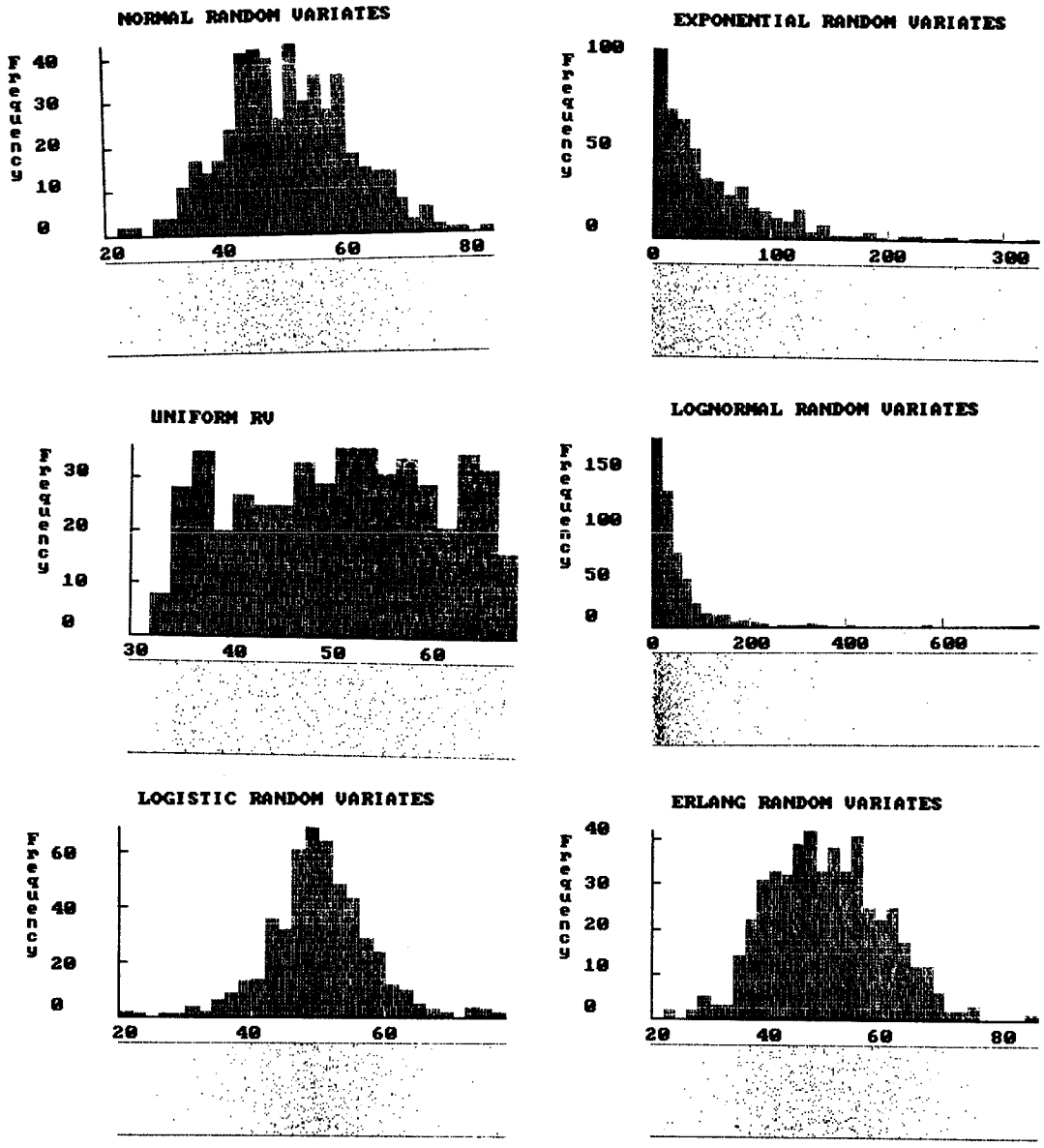


FIGURE 2 Histograms of various distributions.

Certain statistics measure the central tendency of data. The mean, median, and mode are included in TRANSTAT. The mean can often be distorted when outliers are present. It is called a non-robust estimator as a result. A possible remedy, the trimmed mean, for making the mean less sensitive to outliers is present in TRANSTAT. The trimmed mean is where a small proportion of data from each end of the sample is trimmed, and the remaining observations are used to calculate the average (*6*). The median is less sensitive to outliers and, for nonsymmetric distributions, provides a more robust measure of location. The mode presents a poor measure of location. Measures of dispersion indicate the spread or variability of the data and include the range, standard deviation, and coefficient of variation. The standard deviation provides a suitable statistic to judge the reliability of the mean as a measure of location. The standard deviation is a measure of absolute dispersion, and its units are the same as the data's. The coefficient of variation is a measure of relative dispersion. It allows populations to be compared. The peakness of the distribution is called the kurtosis. The symmetry of a distribution is measured by the coefficient of skewness. A zero value indicates a completely symmetric distribution. A distribution can be skewed to the right (positive) or left (negative). Figure 5 illustrates the summary statistics output produced from TRANSTAT. The use of four decimal places enables the analyst to get an indication of the need for rounding; smaller numbers of decimal places can be used.

Outlier Analysis

An important aspect of the distribution fitting methodology is the determination of outliers. Outliers can be due to coding or input errors. Further, with the increasing prevalence of electronic equipment in data collection, protocol errors and instrument errors are becoming common. Outliers can also

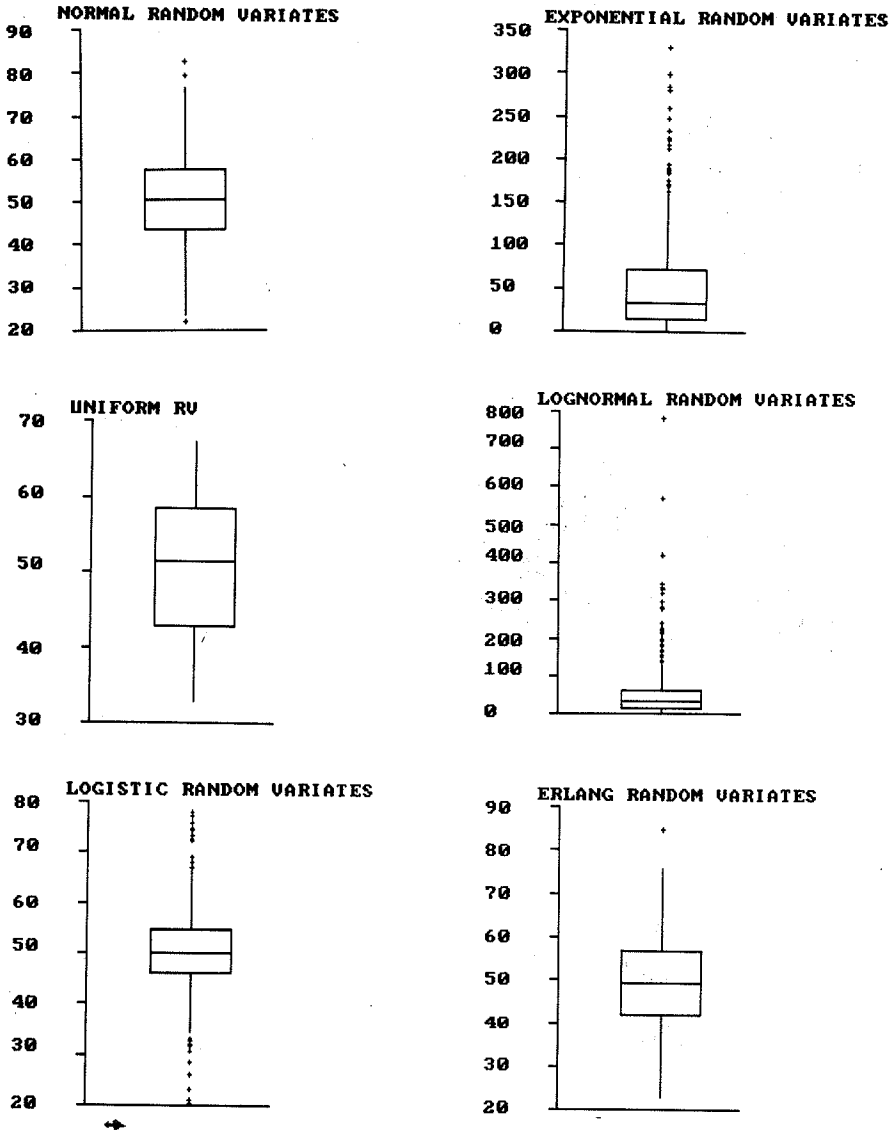


FIGURE 3 Box plots from various distributions.

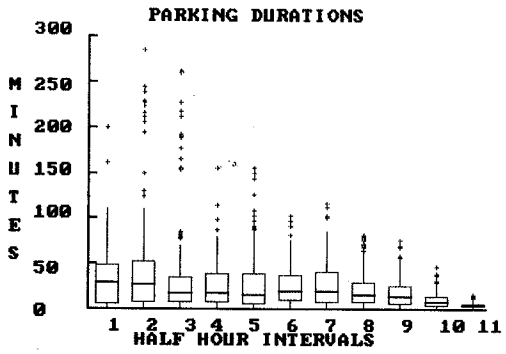


FIGURE 4 Multiple box plots of half-hourly parking duration data.

be due to measurement or copying errors, both of which can easily occur, if the data set is large. Since these values can significantly affect sample estimates, they should be identified and investigated for their validity.

The criteria used to detect potential outliers in the sample data are those of Tukey (*4*) and Barnett and Lewis (*7*). The first method is based on the test used for highlighting observations outside the fences in the box plot. Any value exceeding the distance 1.5 times the interquartile range from the median is selected. This method is distribution-free, which is most favorable since at this stage in the modeling process no distributional assumptions can be inferred. It provides a convenient and simple rule of thumb for identifying possible suspect observations. However, since it assumes symmetry, it is only relevant for samples which are approximately symmetric.

Barnett and Lewis (*7*), in their comprehensive summary of outlier tests, give little attention or hope of any non-parametric methods being used to detect outliers. This is colorfully expressed by their statement, "To deliberately abandon the model, by seeking non-parametric (or distribution free) methods in some broad aim of robustness, smacks of throwing out the bath water before the baby has even been immersed." Therefore, it is not surprising that nearly all Barnett and Lewis's (*7*) "discordancy" tests are based on knowledge of the underlying distribution. TRANSTAT also uses a "discordancy" test to identify any possible outliers for skewed

```
                TRANSTAT

          DURATIONS INTERVAL 2

NO. OF OBSERVATIONS =   90
MINIMUM OBSERVATION =    1.0000
MAXIMUM OBSERVATION =  285
SAMPLE     MEAN     =   52.8778
SAMPLE     MEDIAN   =   28.0000
STANDARD DEVIATION  =   71.1630
SKEWNESS     COEFF. =    1.7924
KURTOSIS     COEFF. =    4.9763
COEFF. OF VARIATION =    1.3458
MODE                     3.0000
```

FIGURE 5 Summary statistics produced by TRANSTAT.

distributions. It is based on the assumption that the sample comes from an exponential distribution. This can be justified by the fact that many known distributions have the exponential distribution as their limit, and many natural phenomena are distributed this way. Furthermore, while being careful to assume not all distributions are exponential, if data is significantly skewed this test is useful and performs quite well.

The system provides information relating to the observations which may be outliers based on the above criteria. It gives the user the option of including these values in subsequent analysis. This relieves TRANSTAT of any decision regarding the inclusion of these values in the analysis but indicates that, if these values are to remain, their influence will be significant (Figure 6). This information should encourage the user to investigate these observations for their validity. Individual values can be deleted from the analysis using this option if an outlier is confirmed.

Model Guidance

A small "expert system" has been included in TRANSTAT to help users identify potential representative models. This can eliminate many unlikely models from detailed examination. An "expert system" is described by Marksjo (*8*) as being "a piece of computer software giving the illusion of being a human expert within a restricted field of competence." There has been considerable interest in expert software applications in statistics, with emphasis being given to systems where consultant (expert) and client (user) interactions are imitated (*5,9*). This new type of software can suggest actions which will achieve the goals of the analysis and provide explanations of output and of recommendations.

TRANSTAT provides a list of suggested distributions which should receive closer attention. The selection criteria are based on inferences obtained from distributional properties (*10*). These potentially should provide a reasonable fit, but formal and informal tests should be used to confirm or reject this preliminary advice. Numerous tests are performed based on the data's symmetry, kurtosis, and range to provide this initial selection of distributions. Confidence limits are generated for the skewness and kurtosis coefficients as well as for the mean and standard deviations. These are then compared with the properties of the theoretical models and recommendations given as to their similarities. The user can accept or reject the system's advice concerning suggestions. The system provides a list of most suitable distributions and ranks these by their apparent suitability. Reasons, in the form of explanations, are also given for the rejected and suggested models. This will help educate analysts. The advice provided must be interpreted as preliminary guidance, and the suggested models should be subjected to the appropriate statistical tests before any conclusions can be made. A sample of the model guidance produced by TRANSTAT is provided in Figure 7.

```
                              TRANSTAT

PRELIMINARY SCANNING OF THE DATA HAS IDENTIFIED 1
UNREPRESENTATIVE OBSERVATION(S) WHICH MAY POSSIBLY BE OUTLIERS

                    OBSERVATION(S) :

      58


THE EFFECT OF THESE  OBSERVATION(S) ON  SAMPLE STATISTICS CAN
SEEN BY THE FOLLOWING COMPARISONS:

       STATISTICS    WITH        WITHOUT UNREPRESENTATIVE VALUES

           MEAN      30.1548     30.0840
           STDVN      7.1155      6.9840
           MIN       14.0000     14.0000
           MAX       58.0000     50.0000

     DO YOU WISH TO DELETE THE ABOVE OBSERVATIONS (Y/N)?
```

FIGURE 6 Outlier detection from TRANSTAT.

```
                              TRANSTAT

                           MODEL GUIDANCE

     AFTER PRELIMINARY INVESTIGATION OF THE CHARACTERISTICS
     OF THE DATA THE FOLLOWING INFERENCES CAN BE MADE :

     THE RANGE DOES NOT INCLUDE NEGATIVE VALUES

     THE DATA IS NOT  APPROXIMATELY SYMMETRICAL

     THE KURTOSIS COEFFICIENT IS  APPROXIMATELY 9

     THE SKEWNESS COEFFICIENT IS  APPROXIMATELY 2

BASED ON THE ABOVE INFORMATION THE MODELS THAT MAY REPRESENT
THE DATA IN ORDER OF PREFERENCE ARE:

          EXPONENTIAL
          ERLANG
          LOGNORMAL

THIS IS PRELIMINARY ADVICE PLEASE ENSURE THAT THE ABOVE MODELS
ARE SUBJECTED TO THE APPROPRIATE TESTS BEFORE ANY  CONCLUSIONS
ARE MADE
```

FIGURE 7 Model guidance output from TRANSTAT.

Chi-Square Test

The previous discussion illustrated the general application of a number of techniques to the investigation of the data and the determination of the distributions likely to fit the data. These are important steps in the distribution-fitting methodology, and it is only after this investigation that data fitting should be attempted.

The statistical fit of the data to the theoretical distribution is tested in two ways. The first is a chi-square test. To test an hypothesis that enumerative data falls into k classes with probabilities $p_1, p_2 \ldots, p_k$ the chi-square test compares the actual counts $o_1, o_2, \ldots, o_k$ with the number expected in each class of k classes, $np_1, np_2 \ldots, np_k$.

The comparison is made using the test statistic,

$$\chi^2 = \Sigma (O_i - E_i)^2 / E_i \qquad \text{all } i$$

where

O_i = observed frequency in class i
E_i = model expected frequency in class i.

The null hypothesis is that the two distributions are the same; and if this is true, the test statistic has a chi-square distribution for large samples ($n > 40$).

Large value of χ^2 indicates that it is unlikely the hypothesized probabilities $p_1, p_2, \ldots, p_k$, are correct and the hypothesized distribution is incorrect. The number of degrees of freedom is the number of classes, k less 1, for each linear restriction placed on the cell counts (*11*).

This test was originally designed for categorical data but can be applied to continuous data models. It involves subjectively choosing class lengths and the beginning of the first cell. Different class sizes can be used, which can affect the significance of the test (*12*). The subjective choice of class intervals may be a limitation with this approach. An additional problem with this test is that low cell expectancies can seriously affect the test statistic. Even with these serious problems, the chi-square test remains the most popular test used in distribution fitting. Its popularity seems to be due to its ease in calculating the test statistic.

The size of the classes can be specified in TRANSTAT. There seems to be little indication of what is an optimal class size; in practice, many should be tried. Cochran (*13*) reports that expected frequencies should be equal. This produces unequal class lengths. Mann and Wald (*14*) suggest the following optimum number of classes (k), based on a sample size (n):

n	200	400	600	800	1000	1500	2000
k	31	41	48	54	59	70	78

The chi-square statistic is usually only calculated for classes where the expected frequencies are greater than 5. However, there seems to be little agreement among statisticians as to the minimum expected number for each class. Cochran (*13*) suggests that, "since the discrepancy between observed and a postulated distribution is often most apparent at the tails, the sensitivity of the chi-square test is likely to be decreased by an overdose of pooling at the tails." In this light, he considers that the inflexible use of a minimum expectation of 5 may be harmful. A sample of the output produced from this module of TRANSTAT is given in Figure 8.

Kolmogorov-Smirnov Test

The second test of goodness-of-fit is the Kolmogorov-Smirnov test. It is for continuous data and can be used for samples of any size (*15*). The test statistic is based on the measurement of the maximum absolute vertical difference between the two cumulative distributions (Figure 9).

The test compares $F(x)$, the hypothesized (cumulative) distribution function, with $S(x)$, the empirical distribution function. If the data follow the hypothesized distribution, then $| F(x) - S(x) |$ should be small.

```
                    TRANSTAT

              CHI SQUARE SUMMARY INFORMATION
THE CHI SQUARED GOF STATISTIC IS =  3.792273  WITH  5  DF
              FOR ALPHA=.01 CR > 15.07516
              FOR ALPHA=.05 CR > 11.01793
              FOR ALPHA=.1 CR > 9.200409
```

FIGURE 8 Chi-square output produced from TRANSTAT.

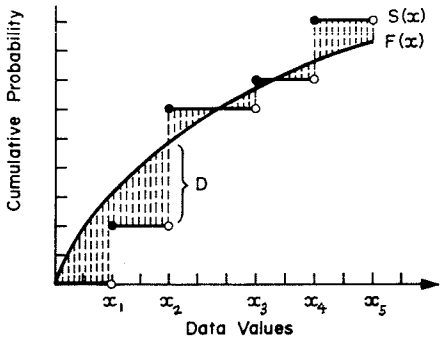


FIGURE 9 Calculation of the Kolmogorov-Smirnov test statistic.

The Kolmogorov-Smirnov test statistic for samples of size n is D = Supremum $|\ F(x) - S(x)\ |$ over all x.

The advantages of this test are that it uses all the data points, it involves no subjective grouping, and it can be used on small data sets. It therefore eliminates many of the weaknesses of the chi-square test. If the parameters of the hypothesized distribution are previously known, the critical regions are distribution-free and, for $n > 30$ are equal to $1.36/n$ at the 5% significance level.

However, if the parameters of the hypothesized distribution are estimated from the sample data, the common tabulated values are not valid; and if used, they result in a conservative test. Recognizing this major weakness, Lilliefors (*16, 17*) developed new tables of critical values for when the parameters are estimated from the sample. Adjusted critical regions were derived using Monte-Carlo techniques for the Exponential and Normal distributions. The Monte-Carlo critical values presented are approximately two-thirds that of the common tables. Monte-Carlo techniques have also been used to generate adjusted critical values for the Log-Normal distribution (*2*). Over 2000 distributions were generated for odd sample sizes up to 30, and linear interpolation was used to produce the even sample numbers. The results seem consistent with those of Lilliefors (*16*).

Durbin (*18*) showed that critical regions can only be generated exactly if the estimated parameters of the distribution are measures of location and spread. This indicates the limitation of Lilliefors's (*16*) method to the Normal, Log Normal, and Exponential distributions. However, Durbin described a half sample device where the parameters are estimated using a random sample of half the data. This procedure is asymptotically equivalent to the original test, and hence the common critical regions can be used. The only drawback with this method is that large samples (100 or more data points) are required. TRANSTAT allows the user to specify whether or not to use the half-sampling technique. If the half sample technique is chosen, critical regions are displayed for the 1%, 5%, and 10% levels of significance for all models.

Another use of D is in comparative studies. Since D is a measure of the model's goodness-of-fit, it can be used as a comparative statistic for evaluating various models. For example, a result of model fit may produce a D equal to 0.11, which can be interpreted as meaning that the two cumulative distributions differ by at most 11% and should be preferred to a model with a D value of 0.16.

The computation of the Kolmogorov-Smirnov test statistic appears to be quite lengthy and complex, but close examination indicates both of these problems can be reduced. D'Agostino and Nothier (*19*) state that this maximum distance does not necessarily occur at an observation point, and therefore only using these points in the calculation of D generally results in a conservative test. However, using mathematical analysis it can be shown that the maximum difference will occur either at an observation point of the empirical distribution or a very small distance before one of these points (*10*). Figure 10 presents some possible comparisons between the theoretical cumulative distribution ($F(x)$) and the observed one ($S(x)$). It can be seen that the maximum D value will be found at, or just before, an observation. Using this result, the computational effort is reduced to arrive at maximum: twice the number of different observation points. This is significantly more efficient than comparing these two functions at numerous small increments. The accuracy of the test is dependent upon how close to the data points the D value chosen is. This in practice should be in the order a one tenth of the measured significance.

Numerous plots can aid the interpretation of D. First, the plot of both the model and the empirical data illustrates the differences between them (Figure 11a). A second plot shows the difference between the theoretical model and the observed data points over the data's range (Figure 11b). More specifically, Figure 11b presents the D value in terms of the independent variable and provides the 5% confidence intervals. The plot provides an immediate view of the quality of the fit on the distribution to the data.

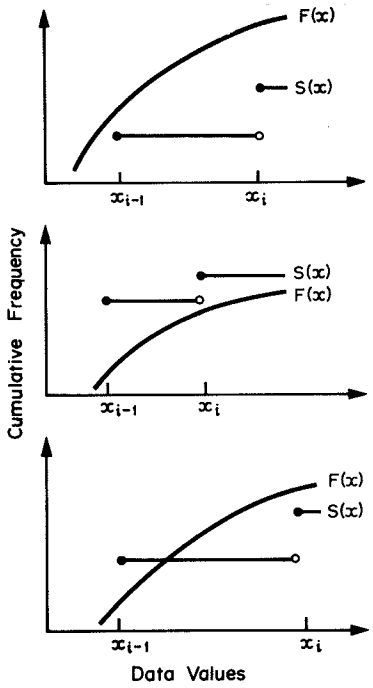
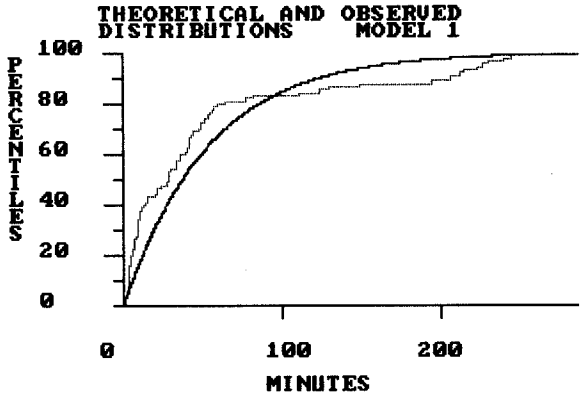
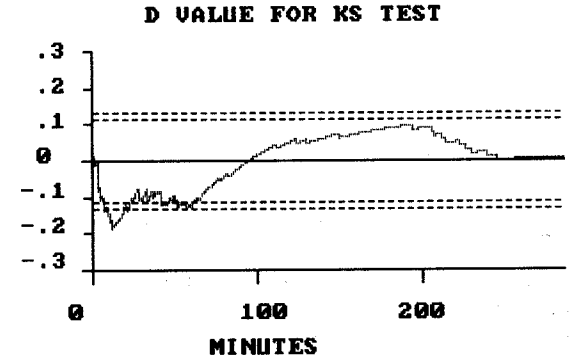


FIGURE 10 Calculation of D.



(a)



(b)

FIGURE 11 Kolmogorov-Smirnov analysis plots produced from TRANSTAT: (a) plot of empirical and theoretical cumulative distributions and (b) difference between observed and theoretical distributions.

PARKING DURATION ANALYSIS USING TRANSTAT

To illustrate the distribution-fitting methodology, TRANSTAT is applied to the investigation of parking durations. A parking study was performed on a Saturday morning in July 1986 at the Mountain Gate Shopping Center lower carpark, Ferntree Gully Road, Ferntree Gully, Australia. An input/output technique was used to collect the data. Observers recorded the time in minutes and the registration number of vehicles entering or leaving the facility. This carpark has three entrances and exits, a good number for such a survey. The carpark has approximately 300 stalls with a variety of shops adjacent to it, including a supermarket, newsagents, and a restaurant. The data collected was summarized into half-hourly intervals of durations by arrival times. This enabled the temporal variation in parking duration curves to be investigated. The first period was for the half-hour from 8:00 a.m. to 8:30 a.m., with the last period being from 1:00 p.m. to 1:30 p.m. Short-term parkers mainly consisted of short shopping trips (*e.g.*, out-of-hours banking at the Commonwealth Handy Bank Automatic Teller and staff being dropped off). High durations were mainly from the staff of the shops and from the supermarket patrons.

After the collection of the data, the first step in the distribution fitting methodology is to begin to build knowledge about the data. This can start by looking at the summary statistics. The summary statistics (Table 1) of the duration data by half-hour intervals provides valuable insight into the data's structure. The durations are presented visually in the multiple box plot in Figure 4. It can be seen that there is a general decrease in the average duration as the study progresses, and that there are many large durations in the first few periods. The large durations were vehicles of the staff of the businesses in the shopping center. They were therefore deleted from analysis as they were considered to be outliers. The criterion used for identifying long-term parkers was durations longer than 180 minutes. The mean duration time of vehicles, corrected for outliers, decreased steadily by the interval of arrival (Figure 12). A simple linear regression equation was fitted to this data with the following relationship established:

$$\text{Mean Duration} = 46.6601 - 3.4481*(\text{Interval no.})$$

The fit was reasonable with an R-squared value of 78.44%. This result is very useful, as the model type that is chosen to represent durations by intervals has the mean as a parameter. The estimation of this parameter can be made with associated confidence intervals using this relationship.

The standard deviation of the durations also decreased in proportion to the mean with a constant linear relationship evident. A linear regression model was fitted providing a very good relationship (R squared of 93.74%). The following relationship was established using Least Squares Regression:

$$\text{Mean} = -9.8457 + 1.4784*(\text{Standard deviation})$$

Data for all the intervals were positively skewed ranging from 1.2741 to 2.5641. These high values of skewness indicate that the normal, logistic, and rectangular distributions would not be suitable for modeling this data. The kurtosis coefficients were between 3.9 and 9.00, with those around 9 indicating the suit-

TABLE 1 SUMMARY STATISTICS FOR PARKING DURATION DATA BY INTERVAL

INTERVAL NO.	N	MAX	MIN	MEAN	MEDIAN	STANDARD DEV.	SKEW. COEFF.	KURTOS. COEFF.	COEFF VAR.
1	48	200	1	35.19	30	40.00	2.21	8.82	1.14
1/A	47	160	1	31.68	29	32.13	1.73	6.98	1.01
2	90	285	1	52.88	28	71.16	1.79	4.98	1.35
2/A	79	149	1	28.92	20	30.71	1.84	6.74	1.06
3	153	262	1	37.05	17	53.94	2.56	9.00	1.46
3/A	143	165	1	24.90	16	27.92	2.67	11.94	1.12
4	160	155	1	25.14	17	24.01	1.94	8.58	0.96
5	181	155	1	26.35	15	30.14	2.21	8.56	1.14
6	219	102	1	25.31	19	21.52	1.27	4.31	0.85
7	193	116	1	27.35	20	24.17	1.28	4.41	0.88
8	171	81	1	21.98	16	19.01	1.29	3.94	0.87
9	104	75	1	18.08	14	16.42	1.44	4.76	0.91
10	63	47	1	11.48	8	10.28	1.47	4.74	0.90
11	38	15	1	4.90	4	3.40	1.54	4.96	0.70
ALL	1420	285	1	26.96	16	34.19	3.46	18.98	1.27
ALL/A	1398	165	1	23.96	15	24.42	2.10	9.19	1.02

NOTE: /A = data with extreme low values removed.

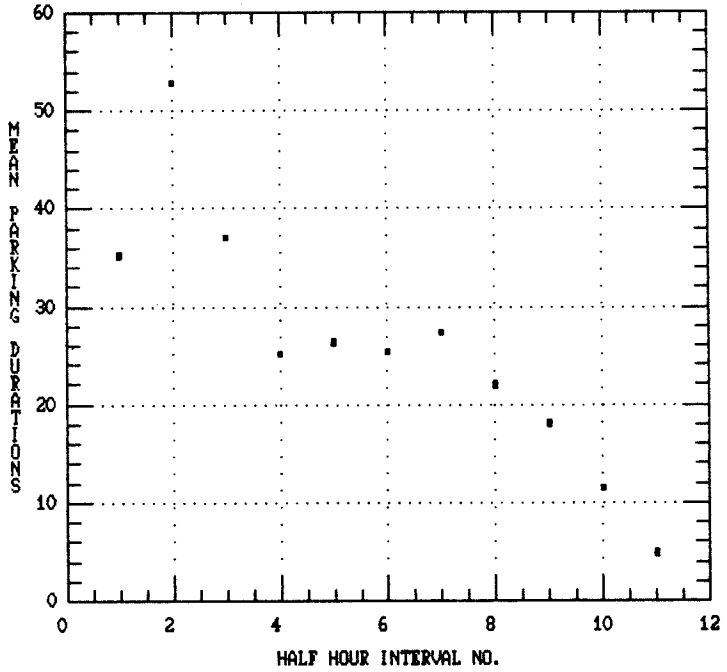


FIGURE 12 Plot of mean duration by arrival interval.

ability of the exponential distribution. A histogram and jitter plot of the data for half-hour interval number 5 is presented in Figure 13. The data for the whole morning can be conveniently summarized in a multiple box plot (Figure 4). The positive skewness of the data by intervals is evident from the small distance from the median and the minimum value compared to the maximum value. The steady decay of duration times by interval is obvious from the reducing medians and maximum values. The high number of outliers, particularly in the early intervals, highlights the long-term (staff) parkers. The general decrease in frequency with increasing duration is consistent with an exponential or Erlang distribution.

From the exploratory data investigation, it was considered that only the Exponential, Erlang, and Log Normal models could possibly represent the data to an acceptable level. Observations over 180 minutes were deleted from the analysis, as their properties were of little interest. The Kolmogorov-Smirnov test was used for analyzing the goodness-of-fit of the distributions and the half-sample parameters were used when the sample size was over 150. This enabled critical regions to be established for these data sets. For the intervals where there was a small number of observations, the Monte-Carlo generated values were used.

The results of each of the models performance are presented in Thompson (*10*). Overall, only one data set (half-hour interval no. 11) failed to be satisfactorily represented by any of the models. To summarize the model fit, a ranking procedure was adopted. A score of 3 points was given for a model fit with no significant difference between the model and the data at the 10% level. A score of 2 was assigned to

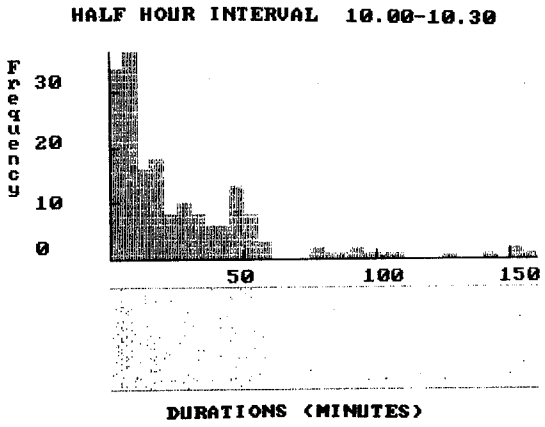


FIGURE 13 Histogram of parking durations half-hour interval no. 5.

a fit where a significant difference was detected at the 10% level. A score of 1 was assigned to a fit where a significant difference was detected at the 5% level. A zero score was assigned if a significant difference was detected at the 1% level. Using these rankings, the results are summarized in Table 2.

The best overall model was the Exponential, with the Log-normal distribution also performing well. The Erlang model rarely fitted the data well. Figure 14 illustrates the fit of both the Exponential and the Log Normal models to the duration data observed in half-hour interval number 8. An alternative ranking procedure based on the smallest D value for each model produced similar results. They are presented in Table 3.

The closeness in performance of the Exponential and Log Normal models is highlighted in Tables 2 and 3. Overall, the Exponential distribution appears to accurately represent the parking durations of shoppers by half-hourly intervals. The associated parameters are a function of the mean of the data. This provides a convenient method for the generation of different distributions with varying means, since the mean/half-hourly interval relationship allows for the temporal variation to be included.

The more common method of looking at the distribution of parking durations is to group the data for the entire study period. The distribution of parking durations for all the data is therefore presented in the box plot in Figure 15. It illustrates the high degree of skewness and large number of outliers that are present.

The summary statistics of the durations for all the data (Table 1) exhibit many of the same characteristics as the individual half-hour interval data: high skewness and kurtosis coefficients. This general trend is further highlighted in the frequency and jitter plots presented in Figure 16. A very high coefficient of kurtosis of 18.7883 was calculated.

The distribution of the parking duration for the entire study period was also investigated. It showed no distribution could be accepted at the 10% significance level. However, after excluding the outliers from the data, the Log Normal model just failed to be accepted at the 10% level, with a D value of 0.0477. Poorer fits were found when every distribution was applied to the data sets with the long-term parkers included.

CONCLUSIONS

This paper has outlined a methodology for traffic distribution fitting. The methodology incorporates the steps of data investigation, model selection, parameter estimation, and goodness-of-fit testing. The methodology has been formulated in

TABLE 2 RANKING OF SIGNIFICANCE OF ERLANG, EXPONENTIAL, AND LOG NORMAL DISTRIBUTIONS TO DATA

	Half-Hour Interval Number											Rank
Model	1	2	3	4	5	6	7	8	9	10	11	Sum
Exponential	2	1	3	3	2	0	3	3	3	3	0	23
Erlang	3	0	3	3	0	0	0	0	0	0	0	9
Log normal	0	0	3	2	3	3	2	3	1	3	0	20

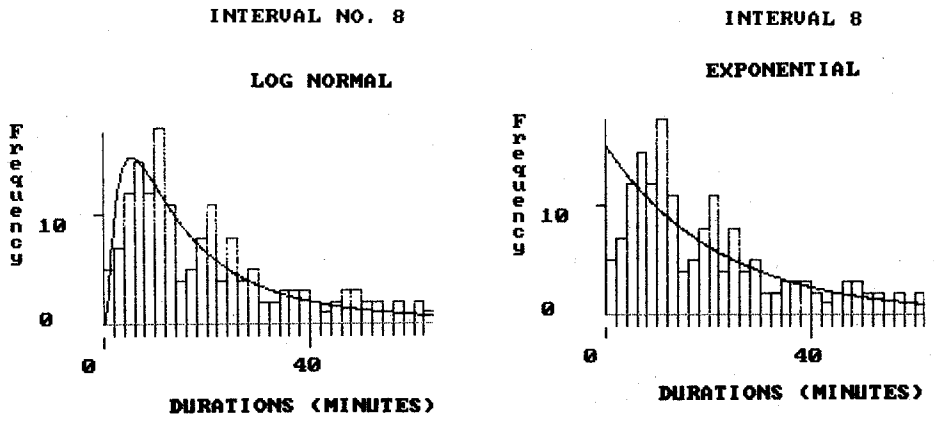


FIGURE 14 Exponential and Log Normal models of interval 8.

TABLE 3 RANKINGS OF DISTRIBUTION FITS BASED ON D

Model	Half-Hour Interval Number											Rank Sum
	1	2	3	4	5	6	7	8	9	10	11	
Exponential	2	2	3	2	2	3	3	2	3	2	1	25
Erlang	3	1	2	3	1	1	1	1	1	1	2	17
Log normal	1	3	1	1	3	2	2	3	2	3	3	24

NOTE: 3 = smallest D, 2 = 2nd smallest D, 1 = largest D.

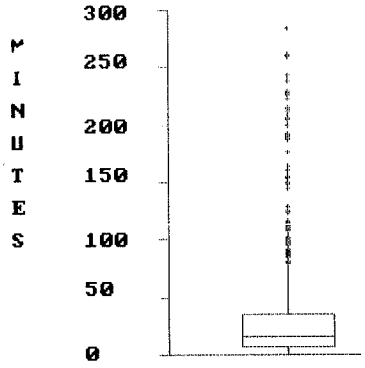


FIGURE 15 Box plot of durations for all intervals.

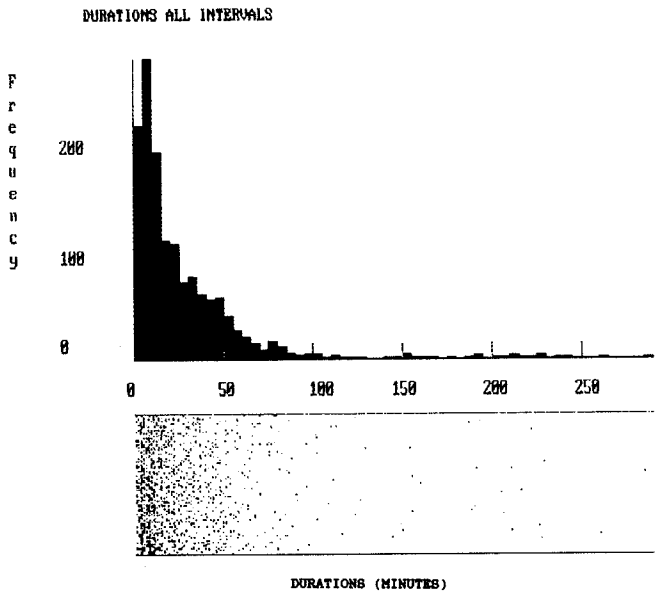


FIGURE 16 Histogram of all durations.

terms of a statistical package (TRANSTAT). The package has incorporated recent developments in exploratory data analysis, outlier determination, and expert systems. It is a comprehensive probability distribution-fitting package which enables probability density functions to be easily fitted to sample data. Full statistical details are provided enabling subsequent probability analysis. The package offers the traffic analyst an interactive, fast, and accurate package to aid in distributional analysis. TRANSTAT has been used for checking distributional assumptions in classical statistical analysis, preparing input to simulation models, and performing probability and analysis.

The distribution fitting methodology was applied to the investigation of the distribution of parking times. The distribution of parking times was shown to be an exponential distribution. However, more importantly, the study demonstrated that the distribution of parking times varied throughout the day. The mean parking time was found to decrease by approximately 3.5 minutes/half hour, from a maximum of 46 minutes in the first half-hour to a minimum of 9 minutes. Other applications of TRANSTAT look at the speed of vehicles in parking lots and headway distributions (*10*).

REFERENCES

1. M. A. P. Taylor and W. Young. *Traffic Analysis: New Technology and New Solutions.* Hargreen Publishing Co., Melbourne, Australia, 1987.
2. R. G. Thompson and W. Young. New Developments in Distribution Fitting. Presented at the International Symposium on Simulation and Mathematical Modeling, Melbourne, Australia, 1987.
3. S. Dvorats and A. Musett. *Basic in Action.* Butterworths Publishing Co., Sydney, Australia, 1984.
4. J. W. Tukey. *Exploratory Data Analysis.* Addison Wesley Publishing Co., Boston, Mass., 1977.
5. J. M. Chambers, W. S. Cleveland, B. Kliener, and P. A. Tukey. *Graphical Methods for Data Analysis.* Duxbury Press, Boston, Mass., 1983.
6. D. F. Andrews, P. J. Ickel, F. R. Hampel, P. J. Huber, W. H. Rogers, and J. W. Tukey. *Robust Estimates of Location.* Princeton University Press, Princeton, N.J., 1972.
7. V. Barnett and T. Lewis. *Outliers in Statistical Data.* John Wiley & Sons, New York, N.Y., 1978.
8. B. S. Marksjo. Expert Systems for Local Government Planners. In *Microcomputers for Local Government Planning and Management* (P. W. Newton and M. A. P. Taylor, eds.), Hargreen Publishing Co., Melbourne, Australia, 1985.
9. G. J. Hahn. More Intelligent Statistical SOFTWARE and Statistical Expert Systems: Future Directions. *The American Statistician,* 39(1), pp. 1–16.
10. R. Thompson. *TRANSTAT: A Statistical Package for Traffic Planners.* Master of Engineering Science thesis. Monash University, Australia, 1987.
11. Mendenhall and Scheaffer. *Mathematical Statistics with Applications.* Duxbury Press, North Scituate, Mass., 1973.
12. E. M. Gumbel. On the Reliability of the Classical Chi-Square Test. *Annals of Mathematical Statistics,* 14, 1943, pp. 253–263.
13. W. G. Cochran. The Chi Square Test of Goodness of Fit. *Annals of Mathematical Statistics,* 23, 1952, pp. 315–345.
14. H. B. Mann and A. Wald. On the Choice of the Number of Class Intervals in the Application of the Chi Square Test. *Annals of Mathematical Statistics,* 13, 1942, pp. 306–317.
15. J. R. Benjamin and C. A. Cornell. *Probability, Statistics and Decision for Civil Engineers.* McGraw Hill, New York, N.Y., 1970.
16. H. W. Lilliefors. On the Kolmogorov-Smirnov Test for Normality with Mean and Variance Unknown. *American Statistical Association Journal,* 1967, pp. 399–402.

17. H. W. Lilliefors. On the Kolmogorov-Smirnov Test for the Exponential Distribution with Mean Unknown. *American Statistical Journal,* 1969, pp. 387–389.
18. J. Durbin. *Distribution Theory for Tests Based on the Sample Distribution Function.* Society of Industrial and Applied Mathematics, Philadelphia, Penn., 1973
19. R. B. D'Agostino and G. E. Noethier. On the Evaluation of the Kolmogorov Statistic. *The American Statistician,* 27(2), 1973, pp. 81–82.

Publication of this paper sponsored by Committee on Traffic Flow Theory and Characteristics.