Cluster Analysis of Arizona Automatic Traffic Recorder Data

JOE FLAHERTY

Monthly factor data were used as input data for cluster analysis of 28 permanent traffic volume counters installed in Arizona. Monthly factors are the ratio of monthly average daily traffic to annual average daily traffic (AADT). Cluster analysis is a statistical procedure that reveals natural groupings in data. There are two types of clustering methods: hierarchical and nonhierarchical. Hierarchical methods use a successive series of either mergers or division. Nonhierarchical methods group objects into a collection of clusters "*K*." Monthly factor data for each location collected over 5 years were used in the cluster analysis. The group mean monthly factors of the groups that were determined and the monthly factors of each location were applied to the appropriate randomly selected daily traffic count. These counts were proxy variables for short-term 24-hr counts. Statistical analysis was used to determine the "best" method for deriving monthly factors and also provided the best estimates of AADT. From the results of this analysis, it was determined that the two primary groups derived from using four clusters were the best and the most stable of all the variations used in the analysis. The statistical analysis revealed that the results obtained from using the grouped mean monthly factors of this variation were marginally better than those from the other variations. The two distinct groups that were determined as a result of the analysis are quite stable with respect to time and provide an estimated level of precision that was greater than acceptable.

The purpose of this study is to assess in as objective a manner as possible the feasibility of implementing the procedures recommended in FHWA's *Traffic Monitoring Guide* (*1*) (TMG) for expanding short-period traffic counts to estimates of annual average daily traffic (AADT).

The TMG suggests that the best approach to use is one that omits as much subjectivity as possible and is based on sound statistical procedures. It recommends that for the purpose of developing estimates of AADT, automatic traffic recorders (ATRs) with similar patterns of monthly variation be grouped together and the means of the monthly factors of the groups be used to expand short counts to estimates of AADT.

The grouping procedure that is recommended in the TMG is a computerized statistical technique called cluster analysis. Cluster analysis is used to discern the groups. Short-term traffic counts can be simulated from daily ATR data and factored volumes using the current method can be compared with factored volumes derived from these simulated counts adjusted by the factors of the appropriate groups as determined by the cluster analysis.

The monthly factors for each of 5 years for the 28 ATRs that were installed at various locations throughout Arizona were used to conduct the cluster analysis. The monthly factors are the ratio of monthly average daily traffic to AADT.

Transportation Planning Division, Arizona Department of Transportation, Phoenix, Ariz. 85007.

CLUSTER ANALYSIS

Cluster analysis is a multivariate procedure for detecting natural groupings in data. In one respect, it is similar to discriminant analysis in which the researcher seeks to classify a set of objects into groups. The difference is, however, that unlike discriminant analysis neither the identity nor the number of groups in the data set is known. Stated another way, discriminant analysis is a classification method that pertains to a known number of groups. The operational objective of any classification method is to assign an observation to one of the groups. Cluster analysis differs from discriminant analysis in that it is a more rudimentary technique.

In cluster analysis, no prior assumptions are made concerning the number of groups or the group structure. The grouping is accomplished on the basis of similarities or distances. The necessary input is data from which similarities can be computed. In the context of this project, it is the 12 monthly factors for each ATR.

There are two basic types of clustering methods: hierarchical and nonhierarchical. Hierarchical methods use either a series of successive mergers or successive divisions. Agglomerative hierarchical methods begin with individual objects. Initially, there are as many clusters as there are objects. First, the most similar objects are grouped. These groups are then merged according to their similarities. This process continues until ultimately the similarity decreases and the groups are fused into one cluster.

Divisive hierarchical methods work, as the name suggests, in just the opposite manner. An initial group of objects is divided into two subgroups so that the objects in one group are most distant from the objects in the other. These subgroups are then divided in the same manner. This process is continued until ultimately there are as many subgroups as there are objects.

The results of both of these methods are often displayed in a dendrogram. A dendrogram is a two-dimensional diagram that depicts the mergers or divisions that have been made at sequential levels.

Nonhierarchical clustering methods group objects into a collection of clusters, *K*. *K*, the number of clusters, may be prespecified or determined by the clustering algorithm. These methods ordinarily begin with either an initial set of seed points that form the nuclei of the clusters or an initial partition of objects into groups. The beginning configuration should

be relatively free of overt bias. This can be assured by randomly selecting seed points or by randomly partitioning the objects into initial groups.

The Systat software package for microcomputers (*2*) was used to conduct the cluster analysis of the ATR monthly factor data. The cluster module of this software package employs both hierarchical and nonhierarchical algorithms.

Hierarchical clustering was the first method applied to these data. If the input data are a rectangular matrix, a distance matrix is computed as a first step. If they are a symmetrical matrix, the input data will be used directly for computing distances. The output is a dendrogram. A dendrogram is analogous to a tree diagram. It displays the linkage of each object or group of objects as a joining of branches in a tree. The base of the tree is the linkage of all the clusters into one cluster, and the ends of the branches point to each object.

The dendrogram is displayed or printed so that the most similar objects are closest to each other in the branch ordering. Additionally, the cluster diameters (joining distances) are printed on the extreme right of the dendrogram. Thus the analyst can see the clusters that are being joined and the distances at which the joining occurs. If the ATR station numbers are input as character variables, they will appear on the extreme left of the dendrogram.

The nonhierarchical method available in Systat (*2*) is the *K*-means method. This is an iterative procedure that assigns objects to nonoverlapping clusters. The number of clusters can be prespecified. The number of prespecified clusters can be as large as the number of cases. The default number is 2. The number of iterations can also be prespecified; the default is 50.

The *K*-means algorithm produces the selected number of clusters by maximizing between, relative to within-cluster variation. It is analogous to a one-way analysis of variance with the number of groups unknown, and the largest *F*-value is sought by reassigning objects to each group. The output is tabular with summary statistics for the number of clusters. Additionally, the members of each cluster are identified and the statistics for the variables that are being clustered are included. Note that all data outputs referred to in this paper are available from the author.

These statistics are, in the aggregate, the sum of squares between clusters and the degrees of freedom, the sum of squares within clusters and the degrees of freedom, and an *F*-ratio that describes the between-cluster variability relative to the within-cluster variability. The statistics for each cluster contain the minimum, maximum, and mean values of the monthly factors. Also included is the standard deviation of the monthly factors and the joining distance of each of the cluster members.

The data for the first year were first analyzed with the joining method. The dendrogram was useful for depicting which ATR stations group together and where they group. It is difficult and cumbersome to use the dendrogram for any fruitful analysis.

The *K*-means method was applied to the same data. This method was the best of the two because the output was in a format that was more fruitful for determining the results of the analysis. The cluster members were clearly identified. The mean, minimum, and maximum, and the standard deviation of the group members were included in the output tables. The distances were displayed next to each ATR station number. It was also helpful that this method allowed the number of clusters to be prespecified and thus varied. Varying the number of clusters and accepting the "best" results is one way to circumvent the overt bias alluded to above.

The *K*-means method was applied to this data set with the number of groups varied from two to nine. The fact that the number of groups could be preselected was useful for two reasons. First, it allowed the analyst to determine how the ATRs were related to each other. Second, it allowed the analyst to see how strongly they were related to each other as the number of groups was increased.

As a result of following the above procedure, a few things became apparent. The first of these was two ATR stations that apparently were not related to each other or to any other ATR station in the other groups. Second, for the stations within the groups, it appeared that the similarity of the pattern of the monthly factors was more a function of geography and topography than functional classification of the highway on which the ATR station was situated. Additionally, the population of the surrounding area did not appear to provide much of an explanation as to why the ATRs grouped as they did.

As the number of groups was increased, beginning with six groups, the two largest groups, based on the number of members, remained relatively constant, but members of the smaller groups were being transferred to new groups. The implication of this is that less and less information was contained in these groups, and they were more a function of white noise.

The same approach to cluster analysis was applied to the other four years of ATR monthly factor data. The number of groups that were prespecified was again varied from two to nine. The joining method was used to cross check the results of the *K*-means method. The results of these analyses were by and large consistent with the results obtained from the analysis of the first-year ATR monthly factor data. The two ATR stations that were outliers in the first year's analysis were also outliers in the other years. Again, there were two primary groups in which the same ATR stations consistently grouped with each other. Also, as with the first-year data, increasing the number of groups led to the "unstable" stations forming new groups.

It was then decided that the two ATR stations that were consistent outliers be excluded from the cluster analysis because they obviously had monthly factors that were vastly different from the other ATR stations and from each other.

These two ATR stations, one near the primary entrance to Grand Canyon National Park, and the other on the primary route to Puerto Penasco (Rocky Point), Mexico, have tremendous variation in their monthly factors. Station 17, near the Grand Canyon, has monthly factors of an average of 1.66 in July and an average of 0.45 in December. Station 26 on the route to Rocky Point is somewhat the reverse with average monthly factors of 0.70 in August and an average of 1.32 in December.

Because of their location and the vast swings in their monthly factors, it was obvious that they each had unique patterns of variation primarily influenced by recreational activity. These stations could be considered as one group each.

These stations and their monthly factors were deleted from each year's ATR monthly factor data set. The data sets were

reanalyzed in the same manner as described earlier. The results of running this analysis again were consistent with those of the prior analysis except that there were no consistent outliers.

The stations in the two primary groups were consistent with the prior analysis. Again, the functional classification of the highway on which the ATRs were located was not a determinant of how the stations clustered. It again appeared that the underlying reasons for the groupings were geography and topography.

One important consideration and objective of deriving ATR groupings has to be the stability of the groupings over time. To check for the stability of the groups, the data were summarized in a table to determine which stations were grouping together in each year and over the 5-year period.

The information in the table was then transferred to maps for ease in determining the geographic and topographic distributions for each year and for the 5-year period. The map for the entire 5-year period confirmed the basic stability of the two primary groupings over the 5-year period. There were, however, some aberrations with respect to these groupings. It appeared that ATR Station 28 did not have the same pattern of stability over the 5-year period that could be expected with respect to its location. It is located in Tucson, which is an urbanized area with a relatively low elevation. On further investigation, it became apparent that the reason it did not conform to expectations with respect to consistent group membership was that in 1 month the traffic at the location was abnormally low. This abnormally low volume was attributable to traffic restrictions that were imposed because of construction.

If anything, these findings point out the need for cogent analysis of the results of the cluster analysis output. Cluster analysis is a powerful analytical tool for discerning "natural" groupings of data. However, the analyst must have an understanding of the data that are being analyzed. In this instance it is imperative that the analyst be, or have available as a resource, someone who is knowledgeable about traffic conditions in proximity to the ATR locations and the state as a whole.

The crucial point to be made in this context is that no analysis, in a strict sense, is completely objective. The conclusions reached as a result of the analysis, must be reasonable and justifiable. Thus it is reasonable to expect that ATR Station 28 would, over the 5-year period, be in Group 1. If it is not, then why is it not? From the discussion above, it is clear that in 1 month of a year the traffic volumes were so divergent from the norm that it led to an unexpected result.

In addition to the individual plots of monthly volumes relative to annual volumes for each ATR in each of the 5 years another graphic tool of analysis was employed in this study. The data set contained the 12 monthly factors for each of the 5 years for each of the ATRs. The data were smoothed by a distance-weighted least-squares algorithm. As the name implies, this algorithm fits a line through a set of points by least-squares regression. Every point on the smoothed line is a function of a weighted quadratic multiple regression on all the points. This procedure produces a true locally weighted curve through the points. This algorithm permits the surface to flex locally to better fit the data.

This plotting procedure is ordinarily used to determine the shape of the function needed to regress one variable on another when the analyst is uncertain of the functional form. They are used here as a post hoc indicator of functional form to clarify the results obtained from cluster analysis. Examples of these plots are shown in Figures 1 through 4. They are representative of Groups 1 and 2, inconsistent and recreational, respectively. The figures show patterns that on a station-by-station basis are consistent with the results of the cluster analysis.

STATISTICAL VALIDATION

The determination of the groups is not the culmination of the analysis. In some respects it is only the beginning. The use of group monthly factors to adjust short period counts must be validated. This validation should be based on a comparison of the present method to the alternative method under development. This comparison can be made by synthesizing short counts and applying both the mean monthly factors derived from the cluster analysis and the monthly factors of the individual ATRs. The "known" AADTs from the ATRs serve as a benchmark for the validation of the factoring approaches.

Randomly selected Monday, Tuesday, Wednesday, and Thursday volumes from the ATRs were used as proxy variables for short-term traffic counts. Data bases were created for each of the 5 years. These simulated short counts can then be adjusted by the monthly factors developed from the cluster analyses and compared with simulated daily volumes adjusted by the monthly factors of each ATR and the unadjusted simulated short period counts.

The daily volumes for each of the ATR stations were adjusted by monthly factors developed in four different ways for comparison purposes. The four monthly factors and the way they were developed are the monthly factors of each ATR and the group monthly factors for three groups, four groups, and five groups, as determined from the respective cluster analysis.

The hard-copy output of the statistical analysis contained the number of cases, the minimum value, the maximum value, the range, the mean value, the standard deviation, the standard error, and the coefficient of variation for each type of the simulated volumes in the data files. They represent, respectively, the AADT, the unadjusted daily volume, the estimated AADT adjusted by the ATR station's own monthly factors, and the estimated AADT using the appropriate group monthly factors derived from the cluster analysis for three, four, and five clusters.

The monthly factors that were used to calculate the simulated estimated volumes were analyzed in a manner similar to the analysis described immediately above. Data files containing the monthly factors of each ATR and the appropriate group monthly factors from the cluster analyses were used to obtain the same statistics as with the simulated volumes. It was of course unnecessary to use a random sample because this comparative analysis was based solely on the monthly factors. In addition to comparing the summary statistics for each ATR, statistical comparisons were made on each group as determined by the cluster analyses.

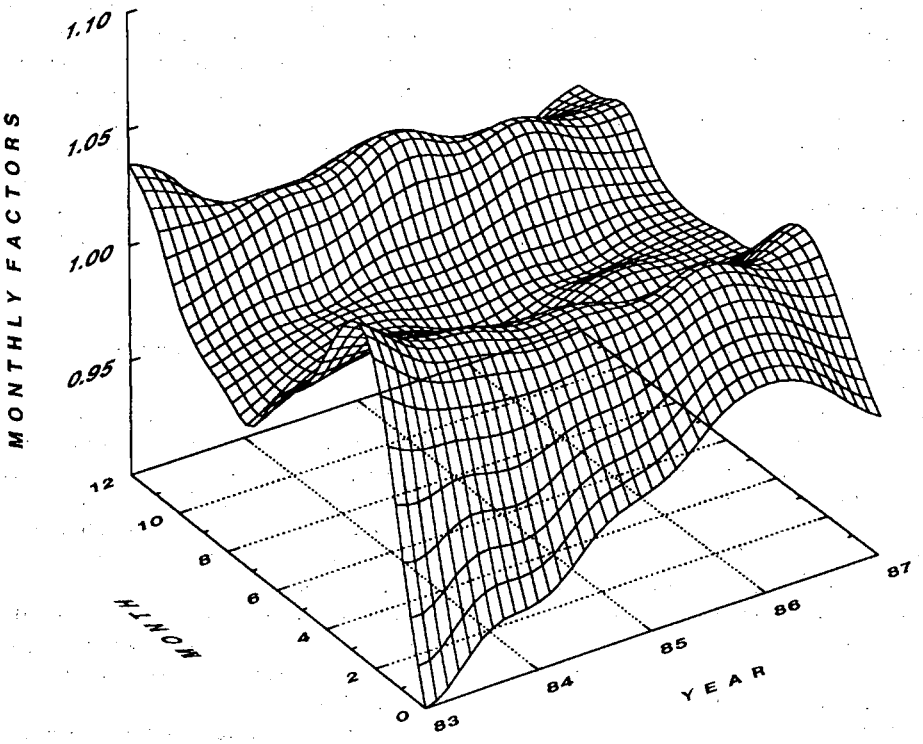

FIGURE 1 ATR Station 14 in Phoenix.

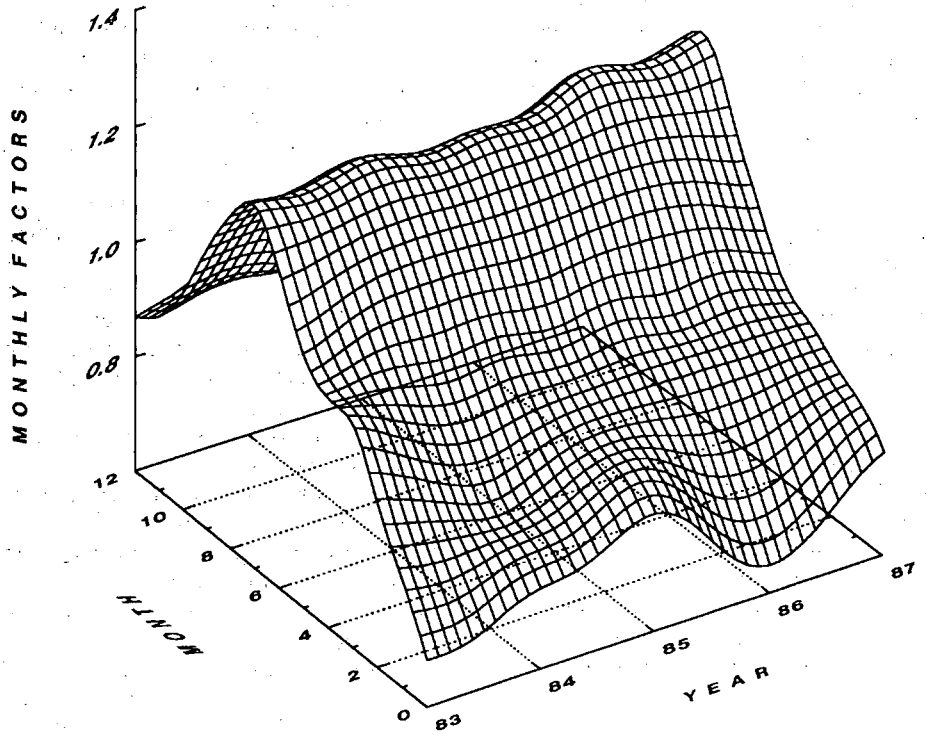

FIGURE 2 ATR Station 6 near Show Low.

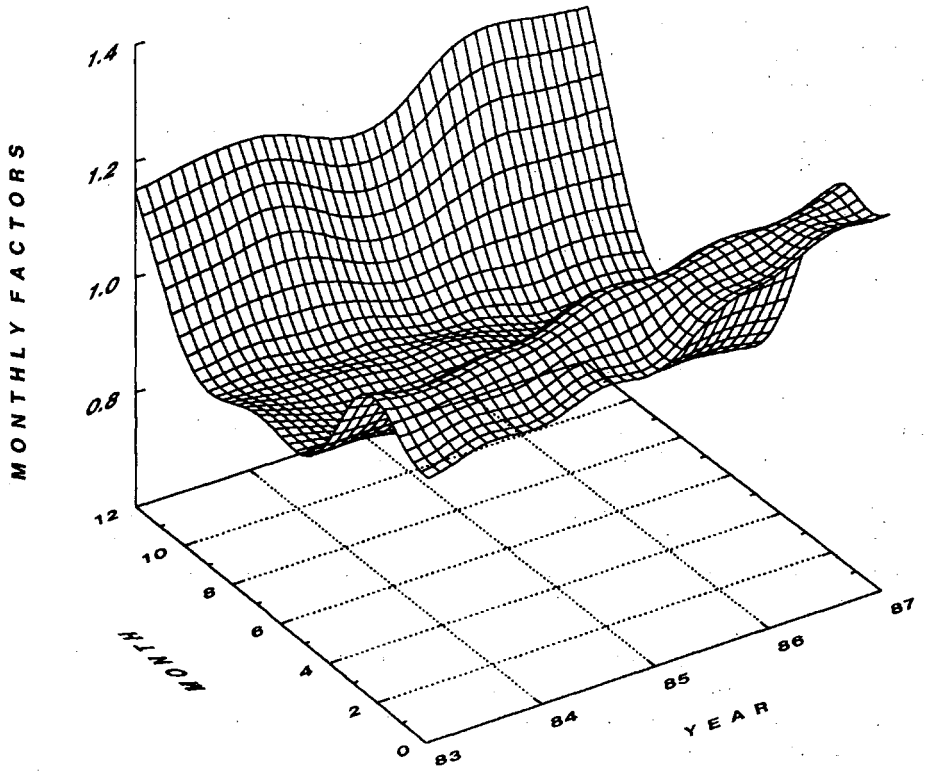

FIGURE 3 ATR Station 25 in Yuma.

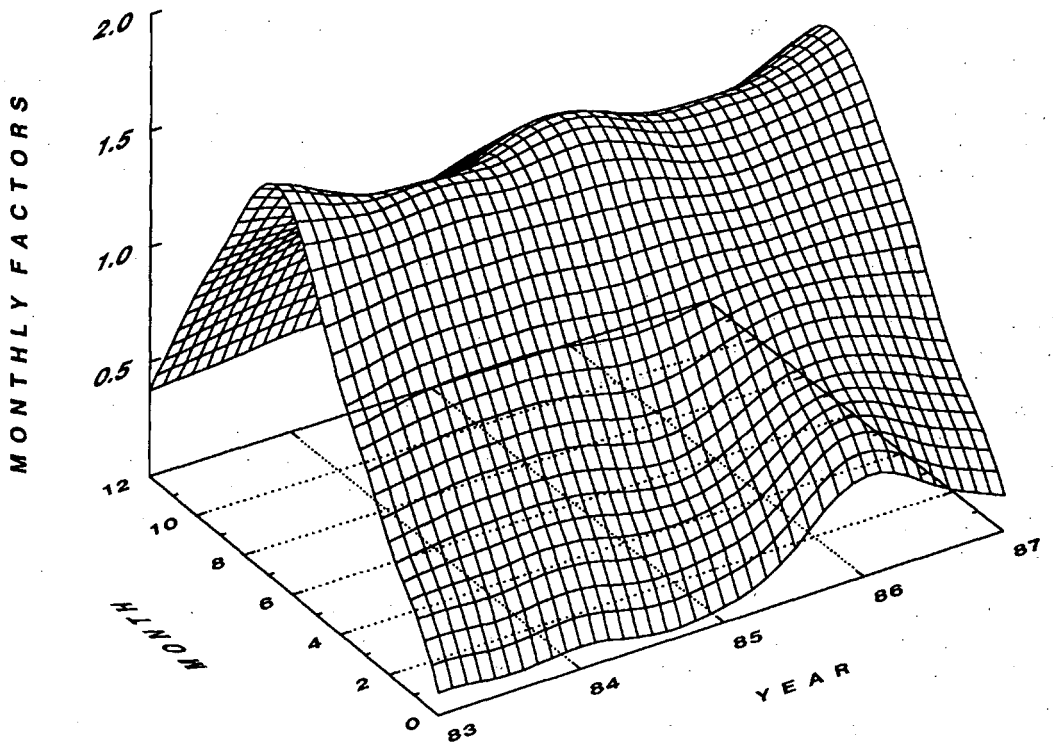

FIGURE 4 ATR Station 17 near Valle.

These summary statistics provide an objective basis on which the efficacy of factoring and a comparison of the various methods of factoring can be made. The different factored simulated short-term counts can be compared on the basis of standard deviations, coefficients of variation, and so on.

These comparisons were made on a station-by-station basis and on a group basis. In virtually every run for all the stations for each year there was an improvement, on the basis of the summary statistics, using simulated factored short counts to estimate AADT rather than unadjusted simulated short counts. The standard deviations were lower and the coefficients of variation were lower.

Although the goal of this research is to develop as objective a method as possible of estimating AADT from short-term counts, it must be remembered that no method is completely objective. The results obtained from computerized statistical analysis cannot be just blindly accepted. They must be interpreted and they must make sense. They are an aid to, not a substitute for, informed decision making. In the context of this analysis, the various statistical analyses had to be consistent and make sense both on an annual basis and over the 5-year period.

From the cluster analysis the procedure using four prespecified groups was the most consistent, particularly with respect to Groups 1 and 2. There was a significant decrease in the joining distance in going to four prespecified groups from three prespecified groups. There was little or no change in the joining distances when going to five prespecified groups from four.

The summary statistics for the monthly factors for each ATR and for Groups 3 through 5 when sorted and analyzed by the various groups also showed that, particularly with respect to Groups 1 and 2, the four-cluster variation was best. The standard deviations, standard errors, and coefficients of variation were consistently the lowest for this variation compared with the other variations. By and large the same results were obtained when the data were analyzed by the ATR stations within the various groups. These results were consistent over the 5-year period.

In all the variations of the monthly factor data that were analyzed, the ATR stations in Groups 3 through 5 were inconsistent—especially with respect to the number of stations in each group and membership of the group. These various group member stations did not group together consistently and bounced from group to group from year to year.

The ATR stations that consistently were in Groups 1 and 2 with the four groupings as determined from the cluster analysis show a consistent pattern of variation in each year and over the 5-year period. The inconsistent ATR stations when graphed over the 5-year period show the pattern of variation characteristics of Group 1 in some years and Group 2 in others. The ATR stations at Valley and Why have characteristics that are different from any other ATR stations and each other.

When an analysis of variance was conducted on the random samples of short counts factored by the four different methods, the null hypotheses that the means of each were the same was easily rejected because the samples were different once the short counts were factored. The mean-square error provided the most useful information, particularly for assessing the efficiency of the estimators. The mean-square error for the within group was the lowest for the estimates obtained by using the group monthly factors derived from the four clusters. Virtually every measure led to the conclusion that the groupings determined by the K-means = 4 method provided the best groupings both on an annual basis and over the 5 years.

As mentioned earlier, the stations that were inconsistent with regard to group membership have similarities in their patterns of monthly factors with the stations in either Group 1 or 2, but they vary from year to year. This observation, however, raises the question, What about those ATRs that are inconsistent with respect to how they group? They do not consistently fall in Group 1 or 2. They do not consistently group with each other. When some of them do group together, there are not enough of them to provide a statistically valid sample.

Looking at the ATR data on an annual basis was not very fruitful. The inconsistencies in the monthly factor data for these ATRs that were the cause of the instability of their grouping over time could not be circumvented by this backdoor approach.

Because of the inconsistency and instability of these ATRs, which is largely caused by the apparent erraticism of the variation in monthly traffic volumes, it is better to exclude these ATRs from the groupings than to attempt to force them to fit into one of the two groups or into one of their own.

One way to perhaps circumvent this problem would be to conduct short-term counts at those times of the year when seasonal adjustments are not necessary. If the ratio of monthly ADT to AADT is approximately 1 then it would not be necessary to adjust a short-term count conducted in that time period for seasonal variation. A data base that contained the monthly factors for those ATR stations that were inconsistent over the 5 years was created.

The individual monthly factors for each of these ATRs were averaged over the 5-year period. The results of this procedure were not encouraging to say the least. None of these ATR monthly factors was on average approximately 1 for any month.

On the basis of the analyses just described, it appears clear that in Arizona there are two distinct, clearly defined, consistent groups whose mean monthly factors can be applied to short-term (24-hr) counts conducted in their respective domains to arrive at a reliable estimate of AADT. The results of the cluster analysis were not quite in conformance with the results expected. The expected results were that almost every ATR station would fall into a clearly defined group. This type of result clearly was not the case in this study.

Two possible explanations for this difference from expected results may be (a) the number of years of data that were analyzed in this study and (b) Arizona's skewed population distribution and topographical divergence.

The two groups have one group in which the ATRs are situated at relatively low elevations and in or near the Phoenix and Tucson metropolitan areas. The second group consists of ATRs situated at relatively high elevations with relatively high volumes in the summer and relatively low volumes in the winter. They are Groups 1 and 2, respectively. Most of the ATR stations that were inconsistent have the characteristics of the two aforementioned groups but they vary from year to year as to which group they resemble. These ATRs are mostly in the western half of the state and along I-40 from Flagstaff

and west. Two ATRs are clearly recreational: ATR 17 near the Grand Canyon and ATR 26 on the primary route to Puerto Penasco, Mexico.

The two distinct groups that were determined as a result of the analysis are quite stable with respect to time and provide a greater-than-acceptable estimated level of precision. It is anticipated that the cluster analysis will be conducted on an annual basis and that the results will be incorporated into the traffic counting program.

Using the group monthly factors will facilitate the assignment of short-term count segments of the state highway system because approximately 75 percent of them are in the domain of the ATRs in Group 1 or 2. The remaining 25 percent will have to be assigned to specific ATRs for adjustment purposes. Seasonal counts will be needed to make the assignments of these count sections to specific ATRs and to delimit the domains of Groups 1 and 2.

ACKNOWLEDGMENTS

This study was supported by FHWA and Arizona Department of Transportation (ADOT). The author acknowledges the helpful comments offered by Ed Green and Dale Buskirk of ADOT.

REFERENCES

1. *Traffic Monitoring Guide*, FHWA, U.S. Department of Transportation, 1985.
2. *Systat Software Package for Microcomputers*. Systat, Inc., Evanston, Ill., 1990.

Publication of this paper sponsored by Committee on Vehicle Counting, Classification, and Weigh-in-Motion Systems.