A Computerized Highway Link Classification System for Traffic Volume Counts

NICHOLAS J. GARBER AND FARAMARZ BAYAT-MOKHTARI

Most state transportation agencies are making a major effort to reduce their annual expenditures on traffic counts yet maintain the desired level of accuracy. These state agencies are therefore developing traffic count programs based on collecting data on a statistically selected sample of highway sections. The assumption is that if road sections can be put into groups such that each group contains sections of highways with similar characteristics, then data collected on a statistically selected sample of sections in any one group will provide traffic data representative of all sections within that group. The main variables used for this grouping are now the FHWA classification system and the average annual daily traffic (AADT) of the road. Unfortunately, the FHWA system can be subjective in cases, and in most cases the AADT of the highway section is unknown or wrong. Estimates of the coefficients of variation of the AADT of groups formed by the standard procedure have therefore tended to be high, which means that large sample sizes are needed to obtain the required accuracy. Consequently, the cost of collecting annual traffic data has not been reduced significantly. In this paper is presented a clustering technique that does not require a knowledge of the link AADTs but does require the use of certain characteristics, such as terrain, land use, and vehicle mix, which are shown to be surrogates of the AADT and can be easily obtained. The technique was used in grouping highway links in the Richmond area of Virginia, and it was found that estimates of the coefficients of variation from sample data were much lower than those recommended by the FHWA. It is concluded that the required sample sizes for annual traffic data collection are lower and this is reflected in lower costs.

Estimates of annual average daily traffic (AADT) volumes are important to the planning and operation of state highway departments. These estimates are used in planning new construction; in improving existing facilities; and, in some cases, in allocating maintenance funds. It is therefore important that any method used to obtain the estimates provide data of sufficient accuracy for the intended use. The importance of having reliable and current data on traffic volumes at hand is generally recognized, and over the years data collection programs have tended to expand. This expansion has led to large amounts of money being spent annually for the collection and analysis of traffic data. Renewed efforts are, however, now being made to reduce the annual expenditure on traffic counts yet maintain the desired level of accuracy. Most of this effort has been focused on developing statewide traffic count programs that collect traffic data on groups of statistically selected highway sections and assume that the average of each group is representative of the volume of sections that have similar traffic characteristics *(1–3)*. The first step in developing such a program requires the classification of highway sections or links into clusters such that all links within each cluster have similar traffic volume characteristics.

The primary factors commonly used for grouping highway links are the functional class of the link as given in the Highway Performance Monitoring System (HPMS) *(4)* and its AADT. The use of these factors, however, presents some problems. First, in several cases the assignment of a particular functional class may be subjective because it is sometimes difficult to differentiate between functional classes such as minor arterials and major collectors and major and minor collectors. In addition, even highways of the same HPMS functional class may not have similar traffic characteristics (e.g., seasonal variation) if they are located in different parts of a state. Second, the AADT at each link is required for that link to be properly assigned to a particular cluster. Unfortunately, in most cases the AADT is not known, and engineers then have to make assumptions based on experience. A highway link classification sytem based solely on these two factors may, therefore, give clusters or groups that include links that have different traffic characteristics. Because the accuracy of any counting program is highly dependent on developing clusters that contain only highway links with similar traffic characteristics, a suitable classification system to achieve this is essential. Such a classification system has been developed for the state of Virginia's rural highways as part of a major study to develop a statewide traffic count program.

Although the procedure was developed primarily for rural roads, it can also be used for urban roads if a set of appropriate guidelines for link identification is developed for urban roads.

DEVELOPMENT OF CLASSIFICATION SYSTEM

The classification system was developed primarily to eliminate the necessity of knowing the AADT of a link before it could be assigned to a group and to avoid sole reliance on the FHWA functional class stratification. The following steps were taken in the development of the system:

1. Link definition and identification,
2. Identification of significant variables that influence AADT, and
3. Link clustering.

Link Definition and Identification

The first step is to break down each highway in the rural area of the state into short, homogeneous sections known as highway links.

N. J. Garber, Virginia Highway and Transportation Research Council and Department of Civil Engineering, University of Virginia, Charlottesville, Va. 22901. F. Bayat-Mokhtari, Virginia Highway and Transportation Research Council, Box 3817, University Station, Charlottesville, Va. 22903.

The main requirement for a link is that each point on it have the same traffic characteristics, such as AADT and daily, weekly, and seasonal variations in traffic volume. The following three basic guidelines were used to identify a link:

1. Freeways and Interstates—Any section of the highway between any two consecutive interchanges is considered a link. This satisfies the main requirement because traffic volume changes cannot occur between consecutive interchanges on these highways.
2. Arterials—Any section of road between any two consecutive major intersections or between any two consecutive intersections if the length is 2 mi or greater is considered a link. This is based on a survey that indicated that the minimum travel distance on such facilities is usually greater than 2 mi.
3. Collectors—Any section of road between any two consecutive major intersections is taken as a link, but with the condition that a new link should start whenever there is a change in the physical appearance of the highway. For example, a new link will begin at a section where the highway changes from a two-lane undivided highway to a three-lane highway with the middle lane used for turning movements.

By using these guidelines and the road inventory mileage record file for each highway district, each rural highway was divided into a number of links.

Each link was identified by its maintenance jurisdiction (i.e., state, county, or incorporated area), route number, county in which the link is located, and a sequence number identifying all of the links belonging to the same highway and located within the same county. In addition, the length of each link is given together with its starting point and end point.

Identification of Significant Variables

Because the main objective was to develop a clustering system that does not initially require the AADT of each link, but at the same time will produce groups that consist of highway links with similar AADTs, it was necessary to identify those variables that have a significant effect on AADT so that they could be used as surrogates of AADT in the clustering system.

A detailed search of the literature indicated that the following candidate variables have some impact on AADT and other traffic characteristics:

1. Locational characteristics
   - Urban versus rural
   - Terrain
   - Area land use
2. Design characteristics
   - Number of lanes
   - Access control
   - Lane and roadway width
3. FHWA functional classification
4. Traffic composition
   - Percentage of passenger cars
   - Percentage of out-of-state passenger cars
   - Percentage of trucks with three or more axles
5. Posted speed limit

To identify which of these variables have a significant effect on AADT, statistical tests were carried out using AADT data for 1977 through 1980 obtained at 112 permanent count stations in Kansas, Maryland, and Virginia. Before any statistical test was carried out, however, it was decided to combine all of the variables under Item 4 into a single variable, defined as "functional use," by considering the individual variables in the following manner:

PC_{ij} = percentage of passenger vehicles in total traffic on link i in state j;
POC_{ij} = percentage of out-of-state passenger vehicles in passenger vehicle traffic on link i in state j;
PHT_{ij} = trucks with three or more axles as a percentage of total traffic on link i in state j;
$\overline{PC_j}$, $\overline{POC_j}$, $\overline{PHT_j}$ = average of PC_{ij}, POC_{ij}, and PHT_{ij}, respectively (i.e., state average for PC, POC, and PHT); and
SPC_j, $SPOC_j$, $SPHT_j$ = standard deviations of PC_{ij}, POC_{ij}, and PHT_{ij}, respectively.

Limits of the mean for a state plus or minus one standard deviation were used to determine whether a particular variable for a given link was high, average, or low with respect to that of the state in which the link is located. For example,

if $PC_{ij} > \overline{PC_j} + SPC_j$, the link PC is high;
if $PC_{ij} < \overline{PC_j} - SPC_j$, the link PC is low; and
if $\overline{PC_j} - SPC_j < PC_{ij} < \overline{PC_j} + SPC_j$, the link PC is average.

Similar limits were defined for POC_{ij} and PHT_{ij}. Table 1 is a matrix of the predominant combinations into which a given link fell and the five types of functional uses that were obtained. These are referred to in this study as recreational, local service, long-distance service, industrial, and local commercial links.

TABLE 1 FUNCTIONAL USE CLASSIFICATION

PC_{ij} (%)	POC_{ij} (%)	PHT_{ij} (%)	Functional Use
High	High	Low	Recreational
High	Low	Low	Local service
High	Average	Low	Long-distance service
Average or low	Low	High	Industrial
Average or low	Low	Average or low	Commercial

- Recreational links—Links in this category have a relatively high volume of out-of-state passenger cars, which may easily be affected by seasonal factors. The exception to this is links located in the vicinity of state boundaries. In general, seasonal characteristics have a significant impact on traffic volume on these links.
- Local service links—These links are used mainly by residents of the area for commuter trips and exhibit relatively little variation in traffic volume throughout the year.
- Long-distance service links—The traffic characteristics of these links are similar to those of the local service links, but they contain a larger portion of long-distance commuter trips.
- Industrial links—These have a relatively high percentage of

heavy trucks and are typically on highways connecting major industrial cities.

- Local commercial links—These links have an average percentage of trucks with three or more axles but a relatively high percentage of pickups. Business trips are likely to be predominant on these links.

The set of candidate variables was further reduced by discarding the locational variable (urban versus rural), because all links considered were located in rural areas, and by discarding access control, because this is somewhat related to the FHWA functional classification system.

The remaining candidate variables were then tested for significant effect on AADT using an analysis of variance (ANOVA). The significant variables identified were

- FHWA functional class,
- Functional use as described in this paper,
- Land use of the county in which the link is located,
- Population of the county in which the link is located, and
- Type of terrain.

These variables were therefore used in the clustering technique described in the following subsection.

Link Clustering

McQueen's K-means method was used as the clustering technique *(5)*. This technique provides for the assignment of a set of m data units to a number of clusters such that data elements within any given cluster are "similar to" or "near" each other. The first K data units are initially selected as K clusters of one member each. The distances between all paired combinations of the K clusters are then computed. If the smallest distance is less than a predetermined minimum (c), the two associated clusters are merged. The centroid of the new cluster is determined and the process is repeated until the distance between the centroids of any two clusters is greater than c. The distances between each of the remaining (m–K) data units and the centroids of each of the clusters already formed are computed and each data unit is assigned to the cluster with the nearest centroid (i.e., minimum d) if this distance is less than c. After each assignment, the centroid of the gaining cluster is computed. If the distance to the nearest centroid of any data unit is greater than a refining parameter (R) where $R \geq c$, then that data unit is taken as a separate cluster.

In this study the squared Euclidean distance in n dimensional space, defined by Equation 1, was used as the basis for representing "similarity" or "nearness" among the data units.

$$d^2_{ij} = \sum_{h=1}^{n} (x_{ih} - s_{jh})^2 \qquad (1)$$

where

d^2_{ij} = squared Euclidean distance,
x_{ih} = value of variable h for case i,
x_{jh} = value of variable h for case j,
d^2_{ij} = squared Euclidean distance between case i and j, and
n = number of variables

The number of clusters obtained is dependent on the values of c and R. In this study R was taken as equal to c. The links were initially clustered for a minimum value of 0.01 for c, which gave the largest number of clusters. The number of clusters was then gradually decreased by gradually increasing the value of c.

It can be seen that the variables identified as significant are mainly nominal variables (e.g., terrain and land use) and cannot, for this reason, be used directly in Equation 1. It was therefore necessary to convert these nominal variables to interval variables. The procedure adopted to achieve this was to represent each category of the nominal variable by 1 percent of the average AADT of the continuous count stations in that category. This provides for a common measure of all variables, and at the same time employs the relative impact of each category of each of those variables on the AADT. As an example the computation carried out for terrain is shown in Figure 1. Based on these computations,

	Type of Terrain		
	Flat	Rolling	Mountainous
Sample Size	37	51	24
$\bar{x}$ AADT (1979)	3035	4831	5071
$\bar{x}$ AADT (1980)	2976	4674	5179
$\bar{\bar{x}}$	3006	4753	5095
Code	30	48	51

FIGURE 1 Sample computation for converting nominal variables to interval variables (from continuous count stations in Maryland, Kansas, and Virginia).

the following codes were used to represent the x_{ih}'s of the significant variables.

- Terrain
  - Flat, 30
  - Rolling, 51
  - Mountainous, 24
- Population of county in which link is located
  - <10,000, 20
  - 10,001 to 20,000, 42
  - 20,001 to 30,000, 38
  - 30,001 to 40,000, 46
  - 40,001 to 50,000, 93
  - 50,001 to 100,000, 68
  - >100,000, 74
- Land use
  - Agricultural, 23
  - Industrial, 63
  - Service, 30
  - Mining, 40
- Functional use
  - Recreational, 31
  - Local service, 59
  - Long-distance service, 37
  - Local commercial, 25
  - Industrial, 10
- FHWA classification
  - Interstates, 93
  - Principal arterials, 42
  - Minor arterials, 31
  - Collectors, 19

TABLE 2 WEIGHTING FACTORS FOR SIGNIFICANT VARIABLES

Variable Type	*F* to Remove Based on AADT Data for 1977	1978	1979	1980	Average *F*	Weighting Factor
Functional use	11.00	11.86	10.08	11.9	11.38	12
FHWA functional classification	13.03	11.86	12.24	9.89	11.01	11
Population of county	4.4	5.09	5.2	4.68	4.84	5
Terrain	2.97	2.98	2.38	2.69	2.76	3
Land use	0.33	0.70	0.74	0.68	0.71	1

These values may be considered representative and can be used by any state because they were computed from data obtained at continuous count stations located in different parts of the country. On the other hand, specific values for a particular state may be computed if there are an adequate number of continuous count stations located in the state.

Another problem that had to be overcome before the technique could be used is related to the differences in the order of magnitude and dispersion among the values of the n variables. It was noted that, if the order of magnitude of the range of values that a particular variable takes is much larger than the range of values for other variables, the value of the "Euclidean distance" between data units will be based on that variable. To overcome this problem, the values of the variables were standardized using Equation 2.

$$Z_{ih} = (x_{ih} - \bar{x}_h)/SD(x_h) \quad (2)$$

where

Z_{ih} = standard value for variable h for case i,
x_{ih} = mean value of variable h for case i,
$\bar{x}_h$ = mean value of variable h, and
$SD(x_h)$ = standard deviation of variable h.

It was also noted that the variables used do not all have the same degree of impact on the AADT of a given link. Therefore a weighting factor had to be included for each variable. A stepwise regression analysis was used to determine the relative influence of each variable on the AADT by assigning the average value of "F to remove" of each variable as the weighting factor for that variable (Table 2). These factors may be used, or may be determined for a given state using the same procedure if data are available at an adequate number of permanent count stations.

The squared Euclidean distance used in the McQueen's K-means technique is therefore

$$d^2_{ij} = \sum_{h=1}^{m} W_h(Z_{ih} - Z_{jh})^2 \quad (3)$$

where W_h is the weighting factor for variable h and m is the number of variables.

A FORTRAN computer program was written for executing the whole procedure based on the flowchart shown in Figure 2.

RESULTS

The methodology was tested using the Interstate, arterial, and collector roads in the Richmond district. At total of 363 highway links were obtained using the guidelines presented earlier. Figure 3 is a plot of the number of clusters versus c and indicates that the rate of increase of the number of clusters is relatively high for values of c less than 2.5 and low for values of c greater than 6.5. These values correspond to nine and four clusters, respectively, and suggest that a reasonable number of clusters is between four and nine.

Because the coefficient of variation of the AADTs within any given cluster is an indication of how successful the clustering procedure is, data on average daily traffic were collected on a sample of links in each cluster for a system of eight clusters, and the coefficients of variation were estimated for each cluster. The results obtained, given in Table 3, indicate that all of the coefficients of variation were lower than the recommended FHWA values.

CONCLUSION

It has been demonstrated that the clustering procedure presented in this paper can be used to form groups of highway links with similar traffic characteristics, without the necessity of first assigning a value for the AADT of each link. The coefficients of variation of the average daily traffic obtained for a test run in the Richmond district show that coefficients well below those recom-

TABLE 3 ADT ESTIMATED COEFFICIENTS OF VARIATION FOR EACH CLUSTER OF A CLUSTERING SYSTEM OF EIGHT CLUSTERS IN THE RICHMOND DISTRICT

Cluster No.	Predominant Type of Highway	No. of Links Sampled	ADT	COV
1	Interstate	5	8,601	0.13
2	Principal and minor arterials	8	6,500	0.16
3	Interstate	9	13,910	0.20
4	Major collectors	6	6,630	0.14
5	Principal and minor arterials	7	12,737	0.16
6	Major collectors	10	3,584	0.18
7	Interstate	8	33,799	0.38
8	Minor arterials	8	1,708	0.18

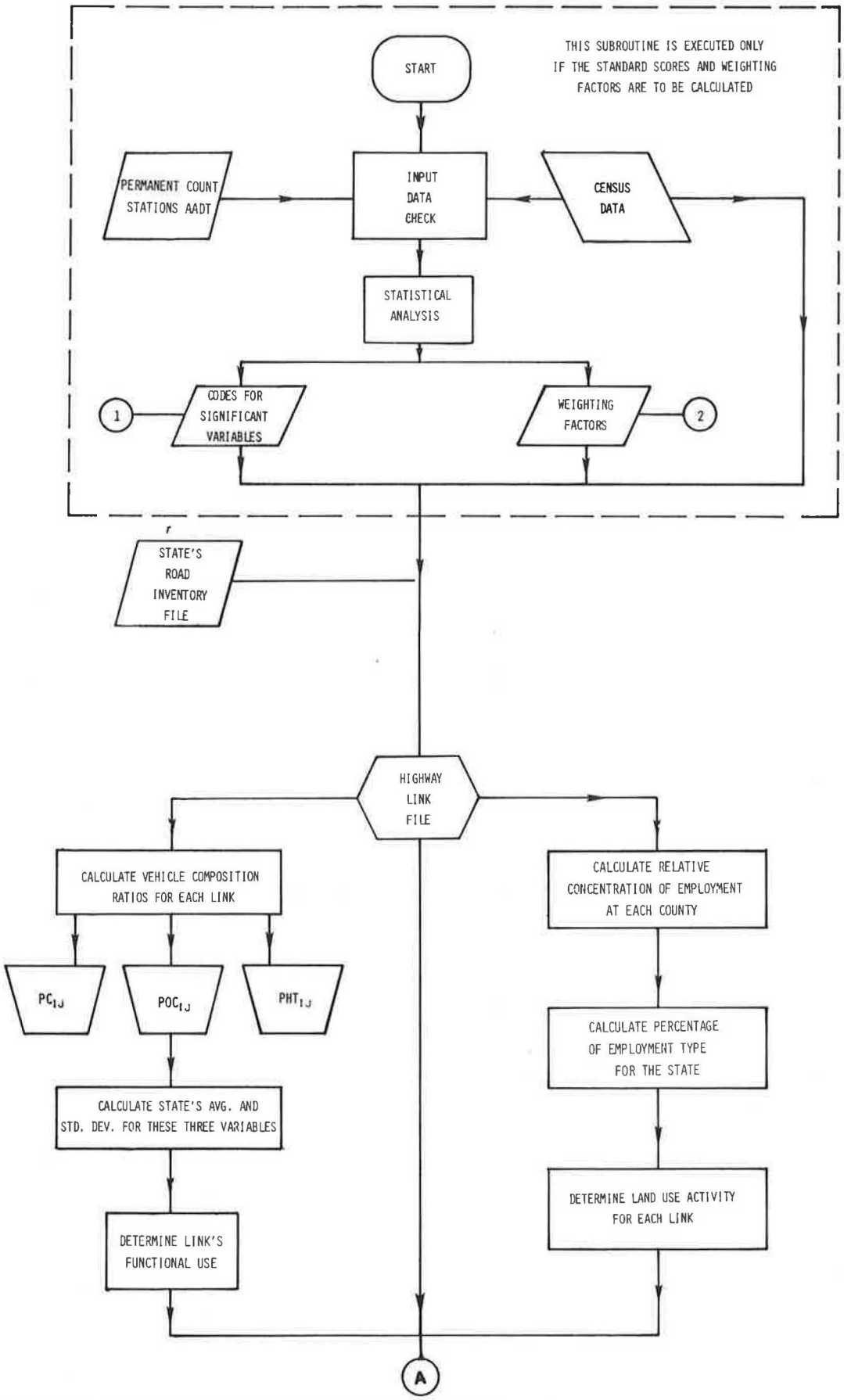

FIGURE 2 Schematic representation of proposed highway classification algorithm.

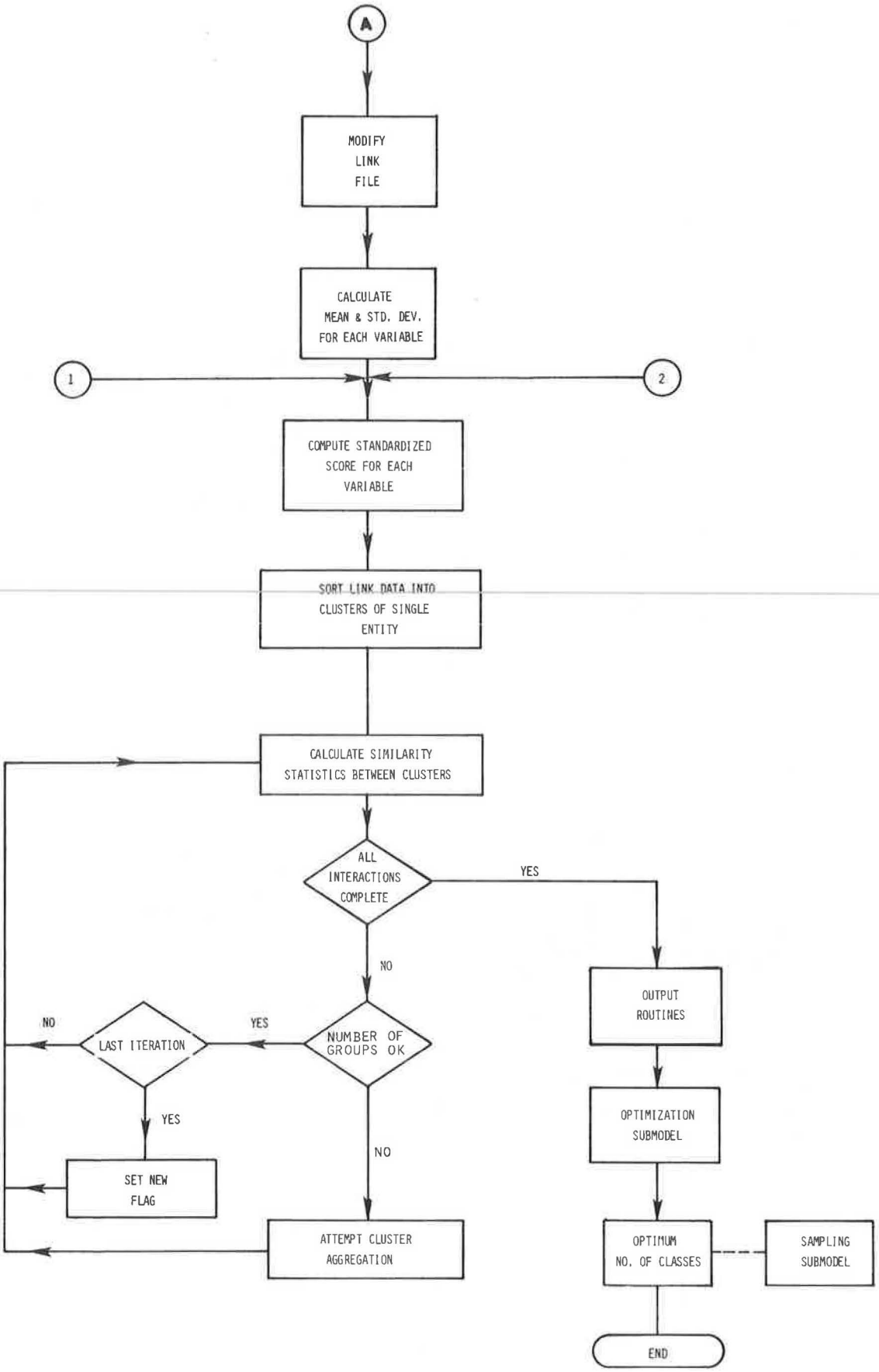

FIGURE 2 *continued*

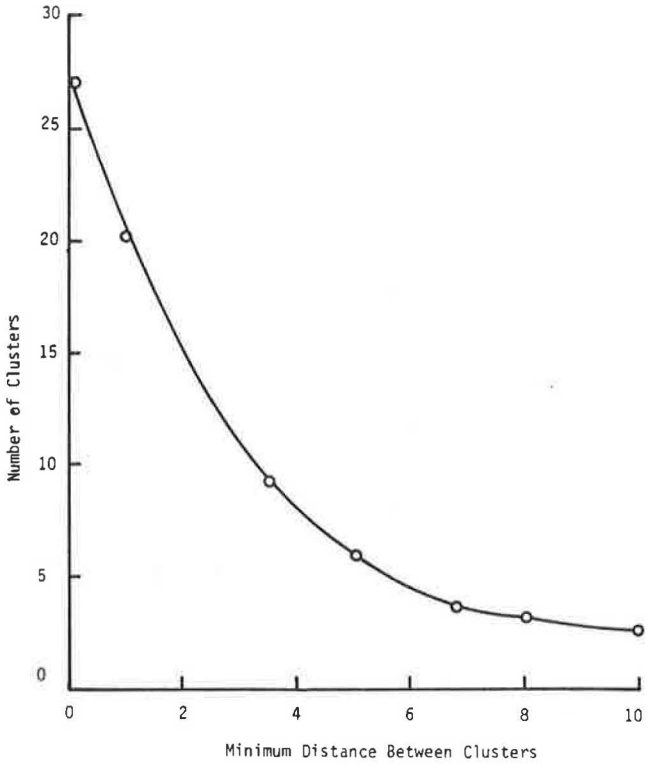

FIGURE 3 Number of groups versus *c* for Richmond district.

mended by the FHWA can be obtained. The procedure requires that all highways being considered for grouping be divided into homogeneous links such that the traffic volume characteristics on any one link do not vary along that link. Input data for each link include the population of the county in which the link is located, the terrain of the area, the land use, and vehicle type. This clustering procedure will develop groups of highway links such that volume data collected at a statistically selected sample of links within a given group will give a good indication of the average AADT of the links in that group. Because the coefficients of variation of the AADTs in each group will be low, when obtained by this procedure, the number of links required to be sampled for a given level of accuracy will be relatively small and will therefore result in cost savings.

REFERENCES

1. *Guide to Urban Traffic Volume Counting*. FHWA, U.S. Department of Transportation, Sept. 1981.
2. *Development of a Statewide Traffic Counting Program Based on Highway Performance Monitoring System*. FHWA, U.S. Department of Transportation, March 1984.
3. P. Garwyn. *Accuracy of Annual Traffic Flow Estimates from Automatic Counts*. TRRL Supplementary Report 515. Transport and Road Research Laboratory, Crowthorne, Berkshire, England, 1979.
4. *Highway Performance Monitoring System*. FHWA, U.S. Department of Transportation, Sept. 1980.
5. W. J. Dixon et al. *BMDP Biomedical Computer Program P*. University of California Press, Los Angeles, 1979.