New Algorithm for Grouping Observations from a Large Transportation Data Base

RUDI HAMERSLAG AND WIM H. SCHELTES

In this paper is presented a new cluster-segmentation algorithm. Its distance measure, derived by using Fisher's likelihood theory, depends on the probability density function (frequency function) of the observations. The resulting measure of similarity or dissimilarity is consistent with the likelihood theory. It shows attractive features: *(a)* curtailment of cluster-segmentation techniques; each probability density function has its own optimal measure of similarity or dissimilarity; *(b)* detection of dependencies between variables; and *(c)* all the advantages of hierarchical divisive techniques, which makes it suitable for analysis of large transportation surveys. The use of the new algorithm is illustrated by using a large data base, the Netherlands National Travel Survey. The goal of this research is to analyze mobility (expressed in daily mileage) by constructing homogeneous population groups. This example clearly demonstrates that the methodology can satisfactorily deal with numerous observations.

Policy making, decision making, designing, research, and so forth require knowledge. Experience is normally one of the major sources of knowledge. Information from surveyed data are also often used. Nowadays a wide spectrum of information about trips and the persons making them is generally available, mostly in the form of data bases. Increasing computer usage makes it possible for an increasing number of people to incorporate these data in their research.

In general, people are only able to focus on a limited amount of information. Therefore it is recommended that the data be simplifed with respect to the specified problem. Often this is done by selection and aggregation of data into groups so as to obtain a manageable number of observations.

The selection of data is linked to the object of study (also named phenomenon or entity) that is to be analyzed. The object of study, in turn, depends on the research goal. Some examples are car ownership, (daily) mileage traveled by chosen mode of transport, and number of trips (per person). Aggregation means that separate observations are put into groups depending on their characteristics ("attributes") and categories.

Aggregating observations that form the object of study into separate groups causes loss of information *(1, 2)*. Different data grouping results in different losses. Unskillful aggregation may therefore lead to erroneous clarification of observations, which induces imprecise management decisions or ineffective infrastructural design.

Characteristics, or variables, are often chosen on the basis of personal experience. Sometimes the scientific background of the researcher plays a role: economists tend to favor a person's income, whereas sociologists show preference for educational level. Each choice will lead to different data groupings, so a "multiple trueness" exists, which results in a reduced, instead of enlarged, insight into the analyzed phenomenon. Policy makers and decision makers are not likely to use such information. Therefore it is advisable to give more attention to the grouping of observations. This can be done, for example, by means of clustering and segmentation. However, several techniques exist and each one will generally lead to a, more or less, different group composition. Thus the problem of "multiple trueness" still remains, as was demonstrated at the Transportation Planning and Research Colloquium in The Netherlands *(3)*. Several reverse-clustering methods also have small disadvantages; for example,

- Limitations on the number of characteristics, which makes an a priori selection necessary;
- Limitations on the size of the data base *(4)*; to solve this, observations are intuitively aggregated; and
- Dependencies between characteristics cannot be detected.

In this paper a new method is presented that overcomes the previously mentioned objections to the traditional methods. It was derived using Fisher's likelihood theory. Proof will be given that the dissimilarity measure is determined by its underlying probability density function. This leads to a curtailment of cluster-segmentation techniques. Detection of dependencies between variables is also now possible.

TRADITIONAL CLUSTER-SEGMENTATION METHODS

The variables used to characterize data groups can be selected by means of cluster-segmentation techniques. Several algorithms exist. Only the agglomerative and divisive hierarchical methods are discussed in this paper, because they are the most commonly used. Further information about cluster-segmentation algorithms is available elsewhere *(5–8)*.

The agglomerative hierarchical techniques use the bottom-up approach ("clustering"). They represent an attempt to minimize the information loss caused by clustering observations. They are popular and also the oldest known hierarchical techniques (e.g., nearest neighbor and farthest neighbor algorithm). Clustering is done by means of a distance matrix. Thus a large number of observations (say, more than 1,000) results in an enormous matrix, which is practically impossible for most computers to solve *(4)*.

In the case of a large number of observations, divisive hierarchical techniques are preferable. They use the top-down approach ("segmentation"); the aim is to maximize the information gain that results from splitting the data base into two subordinant data bases.

R. Hamerslag, Department of Civil Engineering, Delft University of Technology, P.O. Box 5048, 2600 GA Delft, The Netherlands. W. H. Scheltes, Netherlands Institute of Transport, Polakweg 13, 2288 GG Rijswijk, The Netherlands.

The difference between groups of observations is measured by similarities and dissimilarities. One of the most favored methods is the Euclidian distance, a special case of the Minkowski metric. The Minkowski distance *(D)* between two observations ($P_{k,1}$ and $P_{k,2}$) is

$$D(1,2) = \left[\sum_{k=1}^{h} (|P_{k,1} - P_{k,2}|^{r}) \right]^{1/r}$$

where $D(1,2)$ is the dissimilarity value between first and second observation and $P_{k,j}$ is the value of the *j*th observation in the *k*th dimension. Notice that for $r = 2$ this dissimilarity measure is equivalent to the Euclidian distance in an *h*-dimensional space. For $r = 1$ the measure transforms into the so-called city-block (Manhattan) metric.

Dissimilarities can also be qualified by complicated statistical measures, as is done, for example, in the well-known automatic interaction detection (AID) analysis. Its distance measure between groups can be written as follows:

$$D(1,2) = (N1*MEAN1^2 + N2*MEAN2^2 - N12*MEAN12^2)/S$$

where

$D(1,2)$ = dissimilarity value,
$MEANi$ = observed average value in the *i*th group,
$MEAN12$ = observed average value in the combined group,
Ni = size of the *i*th group,
$N12$ = size of the combined group ($N1 + N2$), and
S = standard deviation in the combined group.

NEW METHODOLOGY

This section deals with a recently developed grouping technique. The new dissimilarity measures, the algorithm, and the interaction of data variables are discussed.

The most interesting aspects of the method are

- It is derived by using Fisher's likelihood theory,
- Its distance measure depends on the probability density function of the observations,
- There are no restrictions with respect to the size of the data base and the number of characteristics to be analyzed, and
- Detection of dependencies between characteristics is possible by using the "likelihood ratio test."

Development of New Dissimilarity Measures

The new dissimilarity measures depend on the probability density function ("frequency function") of the observations. They are developed by using the likelihood estimation theory.

Consider a large set of observations of a certain phenomenon, for example, the daily number of trips (phenomenon) of several interviewees (observations). Suppose the probability density function *(f)* of the data base is known. When all observations (*X*1, *X*2, . . . , *Xn*) are stochastically independent, then the likelihood *(L)* follows from

$$L = f(X1) * f(X2) * f(X3) * \ldots * f(Xn)$$

and the logarithm of the likelihood *(LN L)* is

$$LN\ L = LN[f(X1)] + LN[f(X2)] + LN[f(X3)] + \ldots + LN\ [f(Xn)]$$
$$= \sum_{j=1}^{n} LN[f(Xj)]$$

The dissimilarity measure *(D)* between two groups of observations (Group 1 and Group 2) is defined as the difference in log-likelihood before and after clustering them:

$$D(1,2) = LN\ L1 + LN\ L2 - LN\ L12$$

where

$D(1,2)$ = dissimilarity value between the groups (the difference in log-likelihood before and after grouping),
Li = likelihood value of the *i*th group, and
$L12$ = likelihood value of the combined group.

New dissimilarity measures can be calculated for every data base; they optimize the gain of information that results from segmentation. Each probability density function leads to its own characteristic dissimilarity. In the Appendix the derivations are given for

- Binominal (Bernoulli) distribution, a discrete function with true/false or yes/no values (e.g., car ownership, driver's licence);
- Normal distribution, a continuous function; and
- Poisson distribution, a discrete function with nonnegative integers (e.g., daily number of trips per person).

Application of the log-likelihood difference as a new dissimilarity measure will lead to the formulas given hereafter, which are valid for multidimensional *(k)* space. The following abbreviations will be used:

$D(1,2)$ = dissimilarity value between Group 1 and Group 2,
$DIFF_{k,i}$ = auxiliary variable = $1 - MEAN_{k,i}$,
$MEAN_{k,i}$ = observed average value in the *i*th group,
$MEAN_{k,12}$ = observed overall average,
Ni = number of observations in the *i*th group,
$N12$ = total number of observations ($N1 + N2$),
$RTOT_{k,i}$ = auxiliary variable = $Ni - TOT_{k,i}$,
S_k = standard deviation, and
$TOT_{k,i}$ = total observed value in the *i*th group.

For a binominal function (with *MEAN* > 0):

$$D(1,2) = \sum_{k} [TOT_{k,1}*LN(MEAN_{k,1}) + RTOT_{k,1}*LN(DIFF_{k,1})$$
$$+ TOT_{k,2}*LN(MEAN_{k,2}) + RTOT_{k,2}*LN(DIFF_{k,2})$$
$$-TOT_{k,12}*LN(MEAN_{k,12})$$
$$- RTOT_{k,12}*LN(DIFF_{k,12})]$$

A normal distribution leads to the following dissimilarity measure:

$$D(1,2) = \sum_k [N1*(MEAN_{k,1}^2) + N2*(MEAN_{k,2}^2) - N12*(MEAN_{k,12}^2)]/2S_k^2$$

Notice that this formula largely corresponds with the measure used by Ward *(9)* and in AID analysis *(5)*.

For a Poisson distribution (with *MEAN* > 0):

$$D(1,2) = \sum_k [TOT_{k,1}*LN(MEAN_{k,1}) + TOT_{k,2}*LN(MEAN_{k,2}) - (TOT_{k,1} + TOT_{k,2})*LN(MEAN_{k,12})]$$

For the mathematical derivations of the previous formulas, see the Appendix. General information can be found elsewhere *(10, 11)*.

Grouping Algorithm

In this subsection the new dissimilarity measures in a divisive algorithm (segmentation or reverse clustering) are illustrated. With similar ease an agglomerative algorithm (clustering) could be used. Both methods have advantages and disadvantages *(12, 13)*.

In practice the new dissimilarity measures are used as follows. The data base that is to be analyzed contains observations about several variables. For each class ("category") of a variable, the size (number of interviewees) and average observed value (object of study) are noted. The distance formula is derived from the probability density function. This is used in an internal clustering process: all classes of a variable are grouped, and the dissimilarity is calculated for each variable. The variable with the largest value is, under normal circumstances, the most discriminating one; therefore the data base is split up into its classes. This results in several subordinant data bases. Each of these will be analyzed using a similar process. The final outcome is a hierarchical list of discriminating variables.

Dependency of Group Characteristics

There is a possibility that group characteristics (variables in the data base) are mutually dependent. It is essential to know of these dependencies, especially when the use of proxies is considered or the results need to be interpreted, or both.

Notice that the previously presented dissimilarity measures represent the difference in log-likelihood before and after combining the observations. Those results can be applied directly in the formula for the likelihood ratio test statistic.

For two stochastically independent variables (*A* and *B*) the following relationship is valid because there is no "overlap":

$$PROB(A \text{ and } B)/PROB(A) * PROB(B) = 1$$

or, using logarithms:

$$LN\ PROB(A \text{ and } B) - LN\ PROB(A) - LN\ PROB(B) = 0$$

This relationship is hidden in the so-called likelihood ratio test, also known as the G^2-statistic, an easy-to-use method for analyzing dependency or independency between variables *(14, 15)*. The formula for the test statistic is

$$G^2 = -2\ LN[L(A \text{ and } B)/L(A) * L(B)]$$
$$= -2\ [LN\ L(A \text{ and } B) - LN\ L(A) - LN\ L(B)]$$

where G^2 is the test statistic, which has a χ^2 distribution and $L(i)$ is the likelihood value of the *i*th group.

Dependencies between variables are often observed. For example, personal income is, under normal circumstances, strongly related to educational level and age. With each of these variables, the possibility of creating nearly equally homogeneous groups exists. Experience, theoretical knowledge, and insight with respect to the phenomenon under analysis can be usable expedients in making a prudent choice.

ILLUSTRATING THE NEW METHODOLOGY

The new dissimilarity measures have been applied in

- Analyzing mobility *(16–18)*,
- Analyzing travel performance in home-work traffic *(19)*,
- Analyzing the mobility of elderly people *(20)*,
- Predicting the development of public transport usage,
- Analyzing differences in trip generation *(21)*,
- Analyzing car ownership *(22)*, and
- Predicting the development of car population and mobility *(23)*.

In this subsection the use of the new dissimilarity measures with a large data base is demonstrated.

Suppose the research goal is to analyze mobility of the population (expressed in daily mileage) by means of constructing homogenous population groups. A data base is available: the Netherlands National Travel Survey, which contains extensive information about households, the persons belonging to them, and the trips they make. Since 1978 about 23,000 persons have been interviewed annually. All potentially relevant characteristics (demographic, socioeconomic, etc.) were selected for analysis. This resulted in the following 14 variables (the number of distinct classes is shown in parentheses):

- Age (5)
- Car availability (3)
- Children in household (6)
- Citizenship (2)
- City size (3)
- Educational level (7)
- Employment status (5)
- Gender (2)
- Household income (6)
- Income per adult in household (6)
- Income of interviewee (6)
- Marital status (4)
- Position in household (5)
- Railway station nearby (2)

The object of study (daily mileage per person per travel mode) is assumed to have a Poisson-like distribution. Its dissimilarity formula can be found in the subsection on Development of New Dissimilarity Measures. Each travel mode can be viewed as a dimension in multidimensional space; analysis of all modes is done simultaneously.

Data from the Netherlands National Travel Survey result in the

TABLE 1 CALCULATED DISSIMILARITY VALUES (000)

Variable	No. of Classes in Variable	Loss of Information (dissimilarity) Caused by Internal Clustering of Variables (000)	
		All Classes	Final Step
Age	5	80	53
Car availability	3	201	178
Children in household	6	10	2
Citizenship	2	0	0
City size	3	8	4
Educational level	7	51	25
Employment status	5	49	49
Gender	2	74	74
Household income	6	14	8
Income per adult in household	6	13	12
Income of interviewee	6	105	74
Marital status	4	30	26
Position in household	5	101	80
Railway station nearby	2	5	4

scheme of Table 1, which indicates that car availability is the most significant characteristic; much less significant are personal income and position in household, followed by all other analyzed characteristics.

The data base is segmented into the classes of the most discriminating variable. Each subordinate data base is analyzed in a similar way. The final results are given in Table 2 and shown in Figure 1, which shows homogenous Dutch population groups with respect to daily mileage. Similar research, using data from years in the same time range, showed that these groups are fairly time stable *(17)*.

The example demonstrates clearly that this methodology can tackle large data bases without any problems. The only required input information is, for Poisson-distributed data, size and average value of each variable class.

The data in Table 3 make it possible to analyze dependencies. Two variables are independent when

$$G^2 > -2[LN\ L(A \text{ and } B) - LN\ L(A) - LN\ L(B)]$$

The value of $LN\ L(A \text{ and } B) - LN\ L(A)$ is given in the last column of Table 3 and shows the additional information gained in the second segmentation step. The first column gives the value of $LN\ L(B)$, the increase in information had the segmentation process been started with that variable. When used, the G^2-statistic will demonstrate that (in this case) only a few variables are independent of car availability.

Researching the stability of these variables with respect to geographic area or time, or both, might give additional insight; and experience, theoretical knowledge, and the like with respect to the phenomenon under analysis can be usable expedients in making a prudent selection.

CLOSING REMARKS

A new methodology for clustering and segmentation has been presented. Its main advantages are that

TABLE 2 HOMOGENEOUS POPULATION GROUPS IN THE NETHERLANDS WITH RESPECT TO TRAVEL PERFORMANCE (1980)

Group	Size of Group	Daily Kilometrage							
		Car Driver	Car Passenger	Train	Bus, Subway, Streetcar	Bicycle	Moped	Pedestrian	Total (includes other modes)
Car not available									
Under 18 years[a]	7,407		7.4	1.2	2.2	6.3	1.5	0.8	19.7
Housewife in non-car-owning household	1,027	0.1	4.4	2.2	2.0	2.2	0.3	1.1	12.6
Nonhousewife in non-car-owning household	2,580	0.9	4.8	3.1	2.5	3.0	0.5	1.1	16.5
Housewife in car-owning household	2,022	0.3	13.0	0.8	1.2	1.6	0.1	1.0	18.2
Nonhousewife in car-owning household	1,196	1.2	9.1	3.5	2.8	3.7	1.7	0.9	23.7
Subtotal	14,232	0.5	7.7	2.0	2.1	3.7	0.8	1.0	18.2
Car Sometimes Available									
Housewife	2,188	6.9	12.5	0.5	0.7	1.7	0.1	0.8	23.4
Nonhousewife	487	14.0	9.5	5.5	2.2	3.4	0.6	0.9	37.6
Subtotal	2,675	8.0	12.0	1.3	1.0	2.0	0.1	0.8	25.7
Car Available									
No Income	469	16.7	9.6	0.2	0.8	1.0	0.0	0.7	29.4
Under DFL 8,000 (net)	175	22.0	6.1	0.8	0.4	1.0	0.1	0.7	31.8
DFL 8,000-17,000 (net)	846	23.6	4.4	0.5	0.4	1.2	0.0	0.6	30.9
DFL 17,000-24,000 (net)	1,705	27.0	4.9	0.8	0.9	1.2	0.1	0.6	36.3
DFL 24,000-38,000 (net)	1,730	33.0	4.6	1.4	0.9	1.5	0.0	0.7	42.9
Over DFL 38,000 (net)	925	42.9	3.5	2.4	0.6	1.4	0.0	0.8	52.4
Subtotal	5,850	29.6	4.8	1.2	0.7	1.3	0.0	0.7	38.9
Total	22,757	12.3	7.3	1.6	1.5	2.5	0.4	0.9	26.9

Note: 1 km ≈ 0.62 mi.

[a] Minimum age for car drivers in The Netherlands is 18 years.

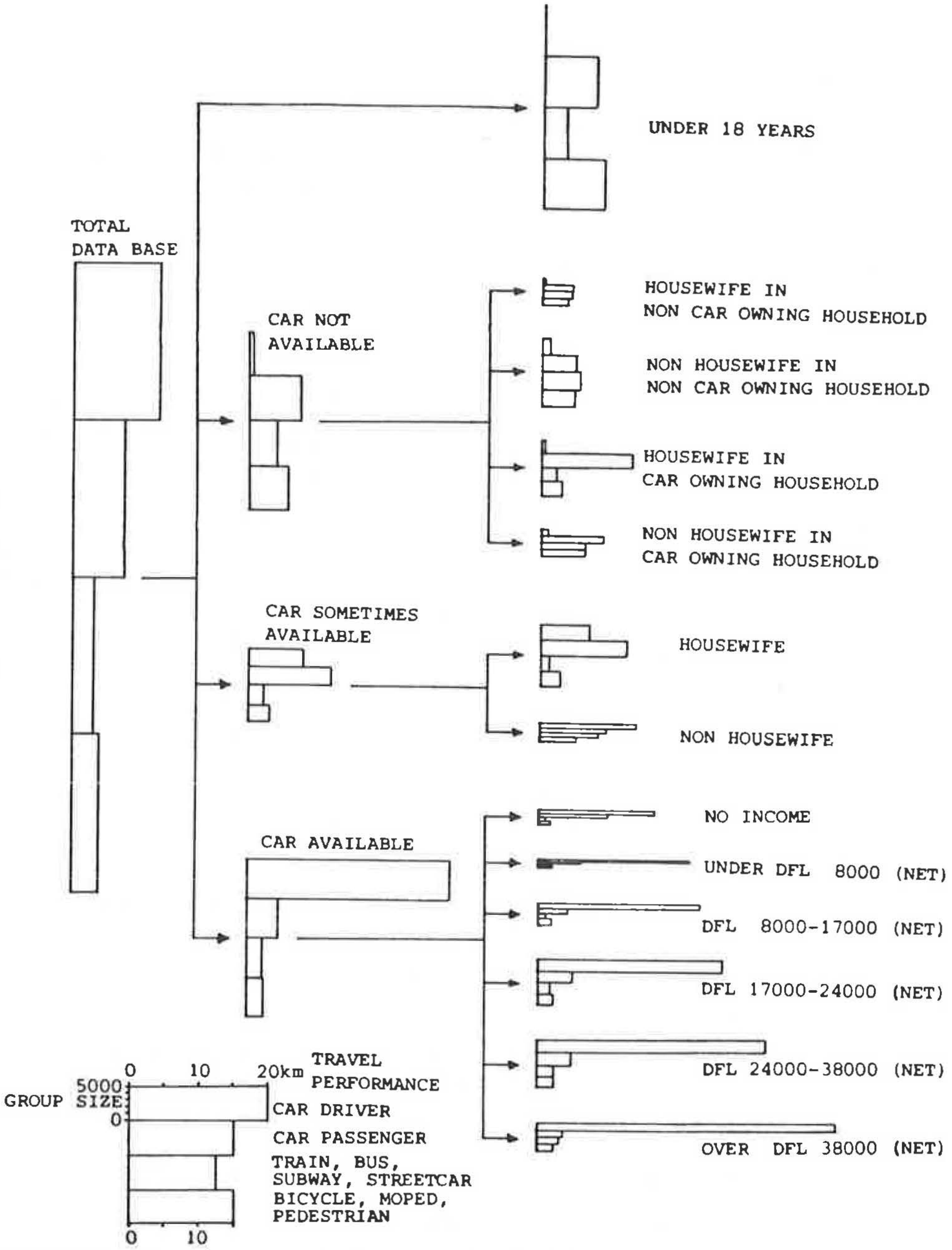

FIGURE 1 Homogenous population groups in The Netherlands with respect to travel performance (1980).

- There are no restrictions on the number of observations. There is also no limit on the quantity of data base variables. An a priori grouping of observations is therefore never necessary. This advantage is present because the technique belongs to the divisive hierarchical algorithms. Agglomerative hierarchical algorithms use distance matrices to calculate the differences between each pair of observations. However, the number of elements in such a matrix is limited by the memory of the computer used.
- The new methodology is consistent with the likelihood theory. It is therefore easier to justify its use than that of other cluster-segmentation methods because each probability density function will have its own specific measure of similarity or dissimilarity.
- Its dissimilarities can be multidimensional; for example, a measure based on daily mileage by several modes of transport.
- Dependencies between variables can be detected by using the likelihood ratio test.

The new method has been applied to various kinds of transportation research. Not only travel performance and trip generation, but also analysis of home-work trips and car ownership were objects of study. The results were in general accordance with expectations.

ACKNOWLEDGMENTS

The authors wish to express their thanks to DHV consulting engineers for providing the necessary facilities during the development of the derived cluster-segmentation algorithm. Also

TABLE 3 ANALYZING DEPENDENCIES

		Car Availabilty Categories[a]				LN L(A and B) –
Variable	LN L(B)	A	B	C	D	LN L(A)
Car availabilty	201					
Personal income	105	0	9	2	12	23
Position in household	101	0	19	3	8	30
Age	80	0	15	2	3	20
Gender	74	1	12	3	7	23
City size	8	1	5	1	4	11
Railway station nearby	5	1	3	0	2	6
Citizenship	0	0	0	0	0	0

Note: Only the most extreme results are given. A = under 18 years, B = car not available, C = car sometimes available, and D = car available.

acknowledged are Steve Ashworth, Caroline Balhuizen, and Eric Kroes of NVI; Ben Immers of Delft University; and Ryuichi Kitamura of the University of California at Davis for their kind assistance during the preparation of this paper.

REFERENCES

1. H. Kassoff and H. D. Deutschmann. "Trip Generation: A Critical Appraisal." In *Highway Research Record 297*, HRB, National Research Council, Washington, D.C., 1969, pp. 15–30.
2. M. G. Richards. *The Use of Disaggregate Models in Travel Demand Analysis.* BGC Consulting Engineers, The Netherlands, 1975.
3. *Proc., Transportation Planning and Research Colloquium* (P. H. L. Bovy, ed.). Zandvoort, The Netherlands, 1983.
4. D. Wishart. *Clustan User Manual (Cluster Analysis Package).* Edinburgh University, Scotland. 1978.
5. B. Everitt. *Cluster Analysis.* Heinemann Editions, London, England, 1974.
6. J. A. Hartigan. *Clustering Algorithms.* Wiley Series in Probability and Mathematical Statistics, John Wiley & Sons, Inc., New York, 1975.
7. H. Späth. *Cluster Analysis Algorithms.* Univeristy of London, England, 1980.
8. D. Steinhausen and K. Langer. *Clusteranalyse, ein Einfuhrung in Methoden und Verfahren der Automatischen Klassifikation.* Federal Republic of Germany, 1977.
9. J. M. Ward. Application of a Hierarchical Grouping Procedure to a Problem of Grouping People. *Educational and Psychological Measurement*, No. 23, 1963.
10. R. Hamerslag. "Afstandsmaat voor de Samenstelling van Probleemgerichte Homogene Bevolkingsgroepen." In *Colloquium Vervoersplanologisch Speurwerk 1980* (P. H. L. Bovy et al., eds.), The Netherlands, 1980.
11. R. Hamerslag and W. H. Scheltes. "Het Gebruik van Cluster/Segmentatiemethoden in de Verkeerskunde." In *Verkeerskundige Werkdagen 1985* (T. De Wit, ed.), The Netherlands, 1985.
12. R. Dubes and A. K. Jain. *Clustering Methodologies in Exploratory Data Analysis.* MSF Grant ENG76-11936A01.1981.
13. B. Everitt. Unresolved Problems in Cluster Analysis. *Biometrics*, No. 35, 1979.
14. Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland. *Discrete Multivariate Analysis, Theory and Practice.* MIT Press, Cambridge, Mass., 1975.
15. M. G. Kendall and W. R. Buckland. *A Dictionary of Statistical Terms.* Longman, Edinburgh, Scotland. 1971.
16. R. Hamerslag. *Population Groups with a Homogeneous Travel Performance by Car, Public Transport and Bicycle.* Department of Civil Engineering, Delft University of Technology, The Netherlands, 1982.
17. R. Hamerslag, L. H. Immers, and J. M. Jager. "The Development of Mobility in The Netherlands." In *Tenth Transportation Planning and Research Colloquium.* (P. H. L. Bovy, ed.), Zandvoort, The Netherlands, 1983.
18. *Maatschappelijke Ontwikkelingen en Mobiliteit.* Projectbureau IVVS. Netherlands Institute of Transport, Rijswijk, The Netherlands, 1983.
19. J. M. Jager. *Onderzoek naar de Factoren die de Hoeveelheid Woon-Werkverkeer Bepalen.* Department of Civil Engineering, Delft University of Technology, The Netherlands, 1984.
20. J. J. Hooning et al. *Een Verkenning naar het Mobiliteitspatroon en de Mobileitsbehoefte van ouderen.* DHV Consulting Engineers, The Netherlands, 1985.
21. J. Termorshuizen, A. W. Dersjant, and T. Grijpma. *Studie in Ruimte en Tijd van Verplaatsingsproductie (Casestudie).* Municipality of The Hague, The Netherlands, 1983.
22. J. M. Jager and W. H. Scheltes. "Analysis of Auto-Ownership by Using a Divisive Hierarchical Technique." In *Transportation Research Record 1037*, TRB, National Research Council, Washington, D.C., 1985, pp. 66–72.
23. R. Hamerslag and W. H. Scheltes. Impact of Fuel Price and Income Changes on Transportation. *Proc., Transportation Research Forum Annual Meeting*, Amelia Island, Fla., 1985, pp. 184–193.

APPENDIX

The derivation of the distance measure was as follows. Starting with the probability density function of the data base, the log-likelihood *(LN L)* is

$$LN\,L = LN[f(X1)] + LN[f(X2)] + LN[f(X3)] + \ldots LN[f(Xn)]$$

$$= \sum_{j=1}^{n} LN[f(Xj)]$$

where

f = probability density function,
L = likelihood value of the cluster, and
Xj = the jth observation in the cluster.

The dissimilarity measure between two observation clusters is defined as the difference in log-likelihood before and after clustering:

$$D(1,2) = (LN\,L1 + LN\,L2) - LN\,L12$$

where

$D(1,2)$ = dissimilarity measure between Group 1 and Group 2,
Li = likelihood value of the ith group, and
$L12$ = likelihood value of the combined group.

This appendix will demonstrate the derivation of dissimilarity formulas for some widely used density functions. Included are

- Binominal distribution,
- Normal distribution, and
- Poisson distribution.

Suppose there are two groups of data ($G1$ and $G2$) of known size ($N1$ and $N2$, respectively). All observations are stochastically independent. Let $X_{k,i,p}$ be the pth observation (pth person) of the ith group in the kth dimension of this multidimensional space. The mean value (*MEAN*) for the ith group in the kth dimension and the total value (*TOTAL*) follow from

$$MEAN_{k,i} = \sum_{p=1}^{Ni} (X_{k,i,p}/Ni)$$

and

$$TOTAL_{k,i} = \sum_{p=1}^{Ni} X_{k,i,p}$$

where

$MEAN_{k,i}$ = mean value for the ith group in the kth dimension,
Ni = number of observations in the ith group,
$TOTAL_{k,i}$ = total value of the ith group in the kth dimension, and
$X_{k,i,p}$ = the pth observation of the ith group in the kth dimension.

Symbolic names used in this appendix are given in Table A-1

Binominal Distribution

The binominal (Bernoulli) distribution is only defined for the values true/false (or 1/0, yes/no, etc.). Its mathematical form is

$$\begin{aligned} f(X) &= MEAN && \text{for } X = 1 \\ &= 1 - MEAN && \text{for } X = 0 \end{aligned}$$

Its likelihood (*L*) of *N* observations is

$$L = MEAN^{TOTAL} * (1 - MEAN)^{N - TOTAL}$$

and the log likelihood (*LN*) is

$$\begin{aligned} LN\ L &= LN(MEAN^{\ TOTAL}) + LN[(1 - MEAN)^{N - TOTAL}] \\ &= TOTAL{*}LN(MEAN) + (N - TOTAL){*}LN(1 - MEAN) \end{aligned}$$

For Group $G1$ this results in

$$\begin{aligned} LN\ L(G1) = \sum_{k} [&TOTAL_{k,1}{*}LN(MEAN_{k,1}) \\ &+ (N1 - TOTAL_{k,1}){*}LN(1 - MEAN_{k,1})] \end{aligned}$$

where

L = likelihood,
$MEAN$ = average observed value,
$N1$ = size of Cluster $G1$, and
$TOTAL$ = total observed value in cluster.

Groups $G2$ and $G12$ lead to similar formulas. Calculation of the dissimilarity (D) follows from

$$\begin{aligned} D(1,2) &= LN\ L(G1) + LN\ L(G2) - LN\ L(G12) \\ &= \sum_{k} [TOTAL_{k,1}{*}LN(MEAN_{k,1}) \\ &\quad + TOTAL_{k,2}{*}LN(MEAN_{k,2}) \\ &\quad - TOTAL_{k,12}{*}LN(MEAN_{k,12}) \\ &\quad + (N1 - TOTAL_{k,1}){*}LN(1 - MEAN_{k,1}) \\ &\quad + (N2 - TOTAL_{k,2}){*}LN(1 - MEAN_{k,2}) \\ &\quad - (N12 - TOTAL_{k,12}){*}LN(1 - MEAN_{k,12})] \end{aligned}$$

where

$MEAN$ = average observed value,
Ni = size of the ith cluster, and
$TOTAL$ = total observed value.

Normal Distribution with Constant Variance

In general, a normal distribution will be chosen when the object of study contains both positive and negative observations. It can also be used in case its mean value differs significantly from zero. The probability density function is

$$f(X) = [S(2\pi)^{1/2}]^{-1} * \exp\{-(1/2)[(X - MEAN)/S]^2\}$$

where

f = probability density function,
$MEAN$ = average observed value,
S = standard deviation,
X = stochastic variable, and
π = mathematical constant (about 3.14)

TABLE A-1 SYMBOLIC NAMES

Group Identification	Size of Group	Observations	Mean Value	Total Value
$G1$	$N1$	$X_{k,01,1} \cdots X_{k,01,N1}$	$MEAN_{k,01}$	$TOTAL_{k,01}$
$G2$	$N2$	$X_{k,02,1} \cdots X_{k,02,N2}$	$MEAN_{k,02}$	$TOTAL_{k,02}$
$G12$ ($G1$ and $G2$)	$N12 = N1 + N2$	$X_{k,12,1} \cdots X_{k,12,N12}$	$MEAN_{k,12}$	$\mathrm{TOTAL}_{k,12}$

The calculation of likelihood *(L)* and log-likelihood *(LN L)* results in

$$\begin{aligned} L = {} & [S(2\pi)^{1/2}]^{-1} * (\exp\{-(1/2)[(X1 - MEAN)/S]^2\}) \\ & * [S(2\pi)^{1/2}]^{-1} * (\exp\{-(1/2)[(X2 - MEAN)/S]^2\}) \\ & * [S(2\pi)^{1/2}]^{-1} * (\exp\{-(1/2)[(X3 - MEAN)/S]^2\}) \\ & * \ldots \end{aligned}$$

and

$$\begin{aligned} LN\, L = {} & LN([S(2\pi)^{1/2}]^{-1} * \exp\{-(1/2)[(X1 \\ & - MEAN)/S]^2\}) \\ & + LN([S(2\pi)^{1/2}]^{-1} * \exp\{-(1/2)[(X2 - MEAN)/S]^2\}) \\ & + LN([S(2\pi)^{1/2}]^{-1} * \exp\{-(1/2)[(X3 - MEAN)/S]^2\}) \\ & + \ldots \end{aligned}$$

where

L = likelihood,
$MEAN$ = average observed value,
S = standard deviation,
Xj = the *j*th observed value of a stochastic variable, and
π = pi, mathematical constant (about 3.14)

For group *G*1 (size *N*1 and observations $X_{k,1,1}$ through $X_{k,1,N1}$) this leads to

$$\begin{aligned} LN\, L(G1) = {} & N1*LN[S(2\pi)^{1/2}]^{-1} \\ & + \left\{ -(1/2)*\sum_k \sum_{p=1}^{N1} [(X_{k,1,p} - MEAN_{k,1})/S]^2 \right\} \end{aligned}$$

Similar formulas are found for Groups *G*1 and *G*12. The dissimilarity measure *(D)* can be calculated from

$$D(1,2) = LN\, L(G1) + LN\, L(G2) - LN\, L(G12)$$

which results in

$$\begin{aligned} D = {} & N1*LN[S(2\pi)^{1/2}]^{-1} \\ & -(1/2)*\sum_k \left\{ \sum_{p=1}^{N1} [(X_{k,1,p} - MEAN_{k,1})/S]^2 \right\} \\ & + N2*LN[S(2\pi)^{1/2}]^{-1} \\ & -(1/2)*\sum_k \left\{ \sum_{p=1}^{N2} [(X_{k,2,p} - MEAN_{k,2})/S]^2 \right\} \\ & + N12*LN[S(2\pi)^{1/2}]^{-1} \\ & -(1/2)*\sum_k \left\{ \sum_{p=1}^{N12} [(X_{k,12,p} - MEAN_{k,12})/S]^2 \right\} \end{aligned}$$

Simplified,

$$\begin{aligned} D = {} & -\sum_k \left\{ \sum_p [MEAN_{k,1}{}^2 - 2*(MEAN_{k,1}*N1) * MEAN_{k,1}] \right\} /2S^2 \\ & -\sum_k \left\{ \sum_p [MEAN_{k,2}{}^2 - 2*(MEAN_{k,2}*N2) * MEAN_{k,2}] \right\} /2S^2 \\ & +\sum_k \left\{ \sum_p [MEAN_{k,2}{}^2 - 2*(MEAN_{k,12}*N12) * MEAN_{k,12}] \right\} /2S^2 \end{aligned}$$

Finally the formula results in

$$\begin{aligned} D = {} & \sum_k (N1*MEAN_{k,1}{}^2 + N2*MEAN_{k,2}{}^2 - N12 \\ & * MEAN_{k,12}{}^2)/2S^2 \end{aligned}$$

Poisson Distribution

A Poisson distribution is characterized by exclusively nonnegative integer observations. For large mean observed values, the function approaches a normal distribution. The mathematical form of the Poisson probability density function *(f)* is

$$f(X) = [MEAN^X * exp(-MEAN)]/X!$$

where

f = probability density function,
$MEAN$ = average observed value, and
X = observation.

Likelihood *(L)* and log-likelihood *(LN L)* follow from

$$\begin{aligned} L = {} & [MEAN^{X1} * \exp(-MEAN)]/X1! * [MEAN^{X2} \\ & * \exp(-MEAN)]/X2! * \ldots \end{aligned}$$

and

$$\begin{aligned} LN\, L = {} & X1 * LN(MEAN) - MEAN - LN(X1!) \\ & + X2 * LN(MEAN) - MEAN - LN(X2!) + \ldots \end{aligned}$$

where

L = likelihood,
$MEAN$ = average observed value, and
Xj = the *j*th observed value of a stochastic variable.

For Group $G1$ (size $N1$ and observations $X_{k,1,1}$ through $X_{k,1,p}$) this results in

$$LN\ L(G1) = \sum_{k} \sum_{p=1}^{N1} [X_{k,1,p} * LN(MEAN_{k,1})$$
$$-MEAN_{k,1} - LN(X_{k,1,p}\ !)]$$

Similar formulas are found for Groups $G2$ and $G12$. The dissimilarity (D) follows from

$$D(1,2) = LN\ L(G1) + LN\ L(G2) - LN\ L(G12)$$

This leads to

$$D = \sum_{k} \sum_{p=1}^{N1} [X_{k,1,p} * LN(MEAN_{k,1})$$
$$-MEAN_{k,1} - LN(X_{k,1,p}\ !)]$$
$$+ \sum_{k} \sum_{p=1}^{N2} [X_{k,2,p} * LN(MEAN_{k,2})$$
$$-MEAN_{k,1} - LN(X_{k,2,p}\ !)]$$
$$- \sum_{k} \sum_{p=1}^{N12} [X_{k,12,p} * LN(MEAN_{k,12})$$
$$-MEAN_{k,12} - LN(X_{k,12,p}\ !)]$$

because

$$\sum_{p=1}^{N1} LN(X_{k,1,p}!) + \sum_{p=1}^{N2} LN(X_{k,2,p}!) = \sum_{p=1}^{N12} LN(X_{k,12,p}!)$$

and

$$N1*MEAN_{k,1} + N2*MEAN_{k,2} = N12*MEAN_{k,12}$$

Thus the formula can be simplified to

$$D = \sum_{k} [TOTAL_{k,1}*LN(MEAN_{k,1}) + TOTAL_{k,2}*LN(MEAN_{k,2})$$
$$-TOTAL_{k,12}*LN(MEAN_{k,12})]$$

Notice that this formula is also usable where all observations are zero ($MEAN = 0$) because

$$LIM\ TOTAL_{k,i} * LN(MEAN_{k,i}/Ni) =$$

$$LIM\ TOTAL_{k,i} * LN(\mathrm{TOTAL}_{k,i}/Ni) = 0 \quad \text{for } TOTAL_{k,i} \downarrow 0$$