Tourism Analysis, Vol. 4, pp. 83–93, 1999
Printed in the USA.

1083-5423/99 $10.00 + .00

www.cognizantcommunication.com

CORRELATES OF TOURIST VACATION BEHAVIOR: A COMBINATION OF CHAID AND LOGLINEAR LOGIT ANALYSIS

BERTINE BARGEMAN,* CHANG-HYEON JOH,† HARRY TIMMERMANS,† and PETER VAN DER WAERDEN†

*Department of Leisure Studies, Tilburg University, P.O. Box 90153, 5000 LE, Tilburg, The Netherlands
†Urban Planning Cluster, Eindhoven University of Technology, P.O. Box 513, Mail Station 20, 5600 MB, Eindhoven, The Netherlands

The aim of study is to examine the relationships between vacation choice behavior and socioeconomic variables. A sequence alignment method is used to classify respondents into homogeneous clusters, based on temporal and spatial aspects of their vacation histories. The relationship between this clustering and a set of socioeconomic variables is then examined using a combination of CHAID and loglinear analysis. The results suggest some interpretable, consistent patterns.

Vacation behavior Sequence alignment method CHAID Loglinear analysis

Perhaps the dominant approach to deriving consumer segments in tourism research is to classify tourists with respect to their vacation behavior. A multitude of approaches and examples can be found in the international literature (e.g., Cha, McCleary, & Uysal, 1995; Cohen, 1979; Etzel & Woodside, 1982; Fakey & Crompton, 1991; Gitelson & Crompton, 1984; Lang, O'Leary, & Morrison, 1997; Mazanec, 1984; Oppermann, 1997; Willenborg & Woodside, 1976; Woodside, Cook, & Mindak, 1987). Most of these segmentation attempts are based on cross-sectional data. Such attempts fail to incorporate any sequential information that might be embedded in vacation histories. Bargeman, Joh, and Timmermans (in press) therefore suggested the use of sequence alignment methods to construct vacation typologies, which do reflect the temporal and spatial aspects, underlying vacation history data. Their study resulted in eight clusters or segments, which proved to have clearly distinct vacation patterns.

Although such typologies that are based on observed vacation histories are of interest to the academic literature, they lack managerial appeal because such typologies do not allow managers to trace tourists of particular interest. Managers, therefore, are (or at least should be) interested in the results of subsequent analyses, which attempt to relate the classification of tourists according to their vacation profile to a set of socioeconomic variables. It allows them to better target their potential customers.

The identification of the relationship between (categorical) socioeconomic variables and the clusters (or segments) of a typology is a well-developed problem in statistics. Researchers can

Address correspondence to Harry Timmermans. Tel.: + 31 (0) 40 247 33 15/247 25 94; Fax: +31 (0) 40 247 58 82; E-mail: eirass@bwk.tue.nl

choose from a variety of techniques, including loglinear and logit analysis. A common problem, however, which in our opinion has not received the attention in the literature it deserves, is that the number of possible combinations of categories of the dependent and independent variables will often rapidly increase, leading to sparse observations in many cells of the multidimensional table. In the present study, we therefore used a combination of CHAID (chi-square automatic interaction detection) and loglinear modeling to find the subset of socioeconomic variables that best reflects the differences between vacation history profiles. In particular, CHAID (Kass, 1980) was used to find the categorization of the variables that accounts best for these differences, and subsequently a loglinear logit analysis was performed to estimate the effects of the socioeconomic variables on the probability of belonging to a particular cluster.

This article is organized as follows. First, in the next section, we briefly describe the methodology that was used to construct the typology of vacation behavior. Then, in the third section, we report the results of the CHAID and loglinear analysis. Concluding comments and a discussion of potential avenues for future research are given in the final section.

The Typology

Data

The data used to derive the typology consist of the Dutch "Continuous Vacation Panel," which is a panel of respondents who regularly report their vacation behavior. The panel started in 1980 and is still in the field. For our purposes, we only used those respondents who were in the panel from 1991 until 1994. The sample size is 1163, representing a random sample of the Dutch population. All respondents went on holiday at least once during this 4-year time period, either for a long or short vacation. The data set contains many different variables representing personal characteristics of the panelists and their vacation behavior. For the construction of the typology, only the following variables were used: timing of the vacation (month of departure), the duration of the vacation (number of days), and the destination of the vacation, classified into domestic and abroad.

Methodology

The construction of the typology of vacation behavior involved the application of the sequence alignment method to the panel data. The fundamental features of this method can be summarized as follows (Kruskal, 1983). The method represents an attempt to derive a measure of similarity between two strings of information in terms of the total amount of effort that is required to change one string into the other. Such strings typically consist of characters that represent some particular phenomenon. For example, two strings to be compared could be DAAD and AAAD, where the character D represents a domestic vacation day and the character A denotes a vacation day abroad. To calculate the amount of effort to equalize strings, weights are attached to a limited number of operations that are allowed to align two strings: "substitution," "insertion," "identity," and "deletion" operations. Let us assume that sequence $\boldsymbol{s}$ has to be equalized to $\boldsymbol{g}$. The operations are step-by-step movements (transitions) from one element of the string to another, first to equalize each of the initial segments, $\boldsymbol{s}^i$ with $\boldsymbol{g}^j$, and finally to change sequence $\boldsymbol{s}$ into $\boldsymbol{g}$. For each (i,j), these movements can be made from one of three cells $(i-1,j)$, $(i-1,j-1)$, $(i,j-1)$, called *predecessors*. Each of these moves represents an operation. The diagonal movement represents "identity" if s_i and g_j are the same at cell (i,j), and a "substitution" if s_i and g_j are not the same. In these cases, the predecessor of the cell (i,j) is the cell $(i-1,j-1)$. The horizontal movement where the predecessor is $(i,j-1)$ represents "insertion" in that the movement adds g_j to $\boldsymbol{s}^i$. Finally, the vertical movement where the predecessor is $(i-1,j)$ represents "deletion" in that the movement eliminates s_i from $\boldsymbol{s}^i$.

That is, weight $w_e(s_i,g_j)$, $w_s(s_i,g_j)$, $w_i(\phi,g_j)$, and $w_d(s_i,\phi)$ for identity (equality), substitution, insertion, and deletion operations can be assigned to each operation. The weight for the identity operation $w_e(s_i,g_j)$ is 0 in any case. The substitution operation may be thought of as the sum of insertion and deletion operations. That is, $w_s(s_i,g_j) =$

$\delta[w_i(\phi,g_j) + w_d(s_i,\phi)]$. If the substitution weight is regarded as the simple summation of the weight of two operations, then the substitution coefficient $\delta = 1$; otherwise, $\delta \neq 1$. Normally, $w_s(s_i,g_j) > w_i(\phi,g_j)$, and $w_s(s_i,g_j) > w_d(s_i,\phi)$. The computation of similarity proceeds by using these weights. Similarity is defined as the sum of operation weights assigned to change sequence s into g.

Because there are many different ways or trajectories to equalize two strings, the sequence alignment method applied by the authors was based on the *Levenshtein distance*. This is defined as the smallest number of substitutions, insertions, and deletions required to change s into g. Hence the equation for the "weighted" Levenshtein distance is:

$$d(s,g) = \frac{1}{m+n} \cdot d(s^m,g^n) \qquad (1)$$

$$d(s^i,g^j) = \min\ [d(s^{i-1},g^{j-1}) + w(s_i,g_j),\ d(s^i,g^{j-1}) + w_i(\phi,g_j),\ d(s^{i-1},g^j) + w_d(s_i,\phi)] \qquad (2)$$

where

$$w(s_i,g_j) = \begin{cases} w_e(s_i,g_j) = 0 & \text{if } s_i = g_j \\ w_e(s_i,g_j) > 0 & \text{otherwise.} \end{cases}$$

Results

To prepare the data for analysis, the vacation histories of the panelists were translated into character strings. In particular, a domestic vacation day was coded by a "D," and a vacation day abroad was coded by an "A." Thus, if a panelist spent 1 week (7 days) abroad every summer, and 2 days in the Netherlands the string representing his/her vacation history would read AAAAAA ADDAAAAAAADDAAAAAAADDAAAAAAA DD. This string would only represent the vacation trips. For the present analysis, however, strings were constructed such that they represented the specific days of the year. If a panelist was not on vacation on any particular day, an "H" was recorded. Because the panel data only record the months of the vacations and not the exact date, we recorded vacations at the start of the relevant segment of the string. Consequently, the typology will not be sensitive to any timing differences within the month.

The readily available computer program CLUSTAL W (Thompson, Higgins, and Gibson, 1994) was used to align the strings. This represented a problem in that this program was originally developed to align DNA strings, which are less complex. The program did not allow us to cluster all panelists simultaneously due to limited memory and especially computing time, as the procedure would have taken weeks to complete. We therefore developed the following strategy to derive the typology:

1. we identified a subsample of panelists, such that they span the observed vacation history patterns;
2. the sequence alignment method was then used to cluster these panelists;
3. the resulting clusters were used as seed points and the remaining panelists were assigned to these clusters.

A total of eight clusters was identified. Table 1 shows some vacation-related characteristics of these clusters. Clusters I and VII were the two largest clusters, with 20.9% ($N = 243$) and 21.6% ($N = 251$) of the respondents, respectively.

Cluster I consisted of respondents who were primarily oriented to their own country. The average frequency of their vacations was 1.3, which is relatively low. The average duration of their vacations was also relatively low, with 8.8 days. On the other hand, the degree of spatial repetition, defined as the degree to which respondents visit the same destination at successive occasions, was the highest among all eight clusters.

The vacation pattern of Cluster II was very stable. On average, respondents belonging to this cluster spent 8 days of domestic vacation, and 20 days abroad, suggesting a foreign orientation. The average number of vacation days per trip was 11.6, which made this cluster the one with most vacation days. This cluster also scored highest in terms of frequency: the average number of vacation periods was 2.5. The degree of spatial repetition was about average with a score of 0.59.

The vacation behavior of Cluster III was more mixed. The average number of domestic vacation days was the highest, but the average number of

Table 1
Size and Characteristics of the Clusters

Cluster	No. of Respondents	Frequency (per year)	Duration (days)	Spatial Repetition	Dominant Orientation
I	243	low (1.3)	8.8	0.71	domestic
II	129	high (2.5)	11.6	0.59	abroad
III	52	low (1.7)	9.6	0.64	mixed
IV	69	medium (2.2)	10.6	0.61	abroad
V	110	medium (2.3)	10.0	0.63	abroad
VI	103	medium (1.9)	11.3	0.62	abroad
VII	251	low (1.7)	9.8	0.58	abroad
VIII	206	low (1.1)	7.3	0.56	mixed

vacation days spent abroad was also relatively high. On average, this cluster spent 11 days of vacation in the Netherlands, and 5 days abroad. The average number of vacation days per trip was 9.6. The average number of vacations per year was 1.7. Furthermore, the degree of spatial repetition was 0.64.

Clusters IV and V were both focused on foreign vacations, with a typical number of about 18 days per year. The average number of vacation days per trip for both clusters was 10, and the average number of vacations per year was 2.2 and 2.3 for Clusters IV and V, respectively. The degree of spatial repetition of both clusters was also very similar: 0.61 against 0.63. The main difference between the clusters concerned the actual sequence of the vacations. The vacations of Cluster V were spread more over time than those of Cluster IV.

Cluster VI was very similar to Cluster II. Table 1 shows that the average number of vacation days for the two clusters was roughly the same. The difference between the two clusters was primarily one of frequency and spatial repetition. The frequency of Cluster VI was 1.9 (against 2.5 for Cluster II); the degree of spatial repetition was 0.62 (against 0.59 for Cluster II).

Cluster VII was also relatively similar to Clusters II and VI, except that the frequency and duration were significantly lower. The average number of annual vacation trips for this cluster was only 1.7; the average duration of a vacation trip was only 9.8 days. This cluster also had a lower degree of spatial repetition, the relevant index being only 0.58.

Finally, Cluster VIII could be qualified as the relative inactive. The average number of vacation days per trip was the lowest of all clusters (7.3 days). The average frequency was also the lowest of all clusters (1.1 trips per year). Perhaps to compensate for the low frequency, this cluster exhibited the highest degree of variety-seeking behavior as indicated by a spatial repetition index of 0.56 only.

Correlational Analysis

Analysis

In the subsequent analysis, we were especially interested in the question whether correlates of the eight clusters could be found. Six predictor variables were used in this analysis. Table 2 lists the information that was available for each panelist: household composition, social class, net annual household income, number of hours of work per week, age, and level of urbanization.

A problem in analyzing the influence of the predictor variables on the probability of belonging to a particular cluster concerns the very large number of possible combinations. The multidimensional cross-table that is made up by these six variables plus the eight clusters has $8 \times 5*3*8*4*10*5 = 192{,}000$ cells. Because we only have 1163 respondents, it is obvious that we face the problem of sparse cells. We therefore decided to use a combination of CHAID and loglinear logit analysis. CHAID is used as a means of prescreening the multidimensional table, and results in a subset of predictor variables and a new categorization of each predictor variable. These results are then used in subsequent loglinear logit analysis to estimate the effects of the resulting subset of predictor

Table 2
Original Set of Predictor Variables and Their Categories

Variable	Categories
Household composition	1. single 2. household with young children (<6 years of age) 3. household with older children (6–12 years of age) 4. household with old children (13–17 years of age) 5. household with adults only (≥18 years of age) (2–4: based on age of youngest child)
Social class	1. high 2. medium 3. low
Net annual household income	1. <Nlg 10,000 2. Nlg 10,000–20,000 3. Nlg 20,000–24,000 4. Nlg 24,000–31,000 5. Nlg 31,000–36,000 6. Nlg 36,000–44,000 7. Nlg 44,000–55,000 8. ≥Nlg 55,000
Number of hours of work per week	1. 0 hours 2. 1–24 hours 3. 25–39 hours 4. ≥40 hours
Age	1. 0–18 years old 2. 6–14 years old 3. 15–18 years old 4. 19–24 years old 5. 25–29 years old 6. 30–39 years old 7. 40–49 years old 8. 50–64 years old 9. 65–74 years old 10. ≥75 years old
Urbanization	1–5 scale, ranging from highly urban to nonurban

variables on the probability that a respondent belongs to a particular cluster.

CHAID (Kass, 1980) is derived from the technique of automatic interaction detection (AID), which was originally developed by Sonquist and Morgan (1964). AID is appropriate for an interval scaled dependent variable and qualitative or categorized independent or predictor variables. The purpose of the technique is to find an optimal breakdown of the data. To that effect, the data are split into two subsets such that the between-subset-sum-of-squares is maximized. This process is repeated until some stop criterion is met. CHAID can be considered as an extension for nominal or categorized dependent variables and is thus appropriate for the present analysis. The technique provides optimal splits although not necessarily bisections, by maximizing the significance of the chi-square statistic at each split. The result is a partitioning of the data into mutually exclusive, exhaustive subsets that best describe the categorized dependent variable.

The technique starts by finding the best partitioning for each predictor variable. This is achieved by finding the contingency table of the predictor and the dependent variable with the highest significance level of the chi-square statistic. Dependent upon the nature of the predictor variable, categories may be merged. In case of nominal variables, any merging is allowed. In case of ordinal data, merging is allowed as long as the ordinal nature of the data is not violated. Next, the significance levels of the chi-square statistic of the various predictor variables are compared and the best pre-

dictor is chosen. The data are then split according to the merged categories of the selected predictor variable. Each resulting cluster is further analyzed in the same way. This branching process stops when the cross-tables are not significant anymore or when the clusters become too small. The final result of the CHAID analysis is the reduction of a given $r \times c_j$ contingency table, with $r \geq 2$ categories of the dependent variable and $c_j \geq 2$ categories of the predictor variables j to the most significant $r \times d_{j'}$ table ($1 \leq d_{j'} \leq c_j$, $j' \leq j$).

It will be evident that this procedure represents a combinatorial problem. The total number of possible ways to reduce the multidimensional table rapidly increases with an increasing number of predictor variables and/or categories. CHAID therefore is not based on complete enumeration, but uses an alternative approach that cannot guarantee the optimal solution, but gives satisfactory results. It first finds the contingency subtable of the pair of categories of the predictor with the lowest significance. If this significance is below a user-defined threshold value, the two categories are merged. This process is repeated until no further merging can be achieved. Next, each resulting category consisting of three or more of the original categories is tested. If the most significant split of the compound category exceeds a certain threshold value, the split is implemented. These steps are repeated until no further improvement is achieved.

Results

The results of the CHAID analysis are displayed in Figure 1. The following operational decisions were made. A variable was not split if it was not significant at the 5% level. A level of 10% was used for testing the Bonferroni significance. Figure 1 shows that household composition was the most important predictor variable in the sense that it caused the first split. The merge suggests that three segments with different vacation behavior can be identified: households without children, households with young children, and households with older/old children.

The relationship between the predictor variables and the classification of panelists can be further elaborated by inspecting cross-tables that express the relative distributions of the predictor variables across the eight clusters. Table 3 summarizes the relationship between the predictor variables and these clusters. It shows that households without children were dominant in, for example, Clusters V and VIII. As expected, households with young children spent their vacation domestically. If they went abroad, the duration of their vacation tended to be relatively short. Households with older/old children tended to make more vacations abroad and these vacations tended to involve more days. Moreover, they went on vacation more often.

Households without children were further split into a nonretired (younger than 65 years) and retired (65 years or older) age segment. The latter predominantly showed a mixed pattern with a (very) low frequency of going on holiday. If older people went abroad, they stayed there for a relatively short time. Households without children that belonged to the other age segment predominantly belonged to Clusters VII and VIII.

The households with older/old children were differentiated by the CHAID analysis according to their social class. The results were as expected. Households of medium and low social class tended to spend their vacation in the Netherlands or were relatively inactive as illustrated by the relatively high percentage for Cluster VIII. In contrast, households of high social class belonged to the clusters characterized by frequent foreign vacations of longer duration.

The younger age households without children were broken down into three segments. As Table 3 indicates, all three segments predominantly spent their vacations of low frequency and short duration abroad. However, some small differences could be observed. For instance, respondents with higher income tended to spend vacations of longer duration and higher frequency compared to respondents with medium income (see the percentages for Clusters II, IV, and V). The latter group was relatively more characterized by a mixed or domestic pattern with a (very) low frequency. The lowest income was relatively well presented in Clusters II, IV, and VI.

The next step in CHAID concerned the splitting of the households with older/old children of medium and low social class into a nonworking and a working segment. As Table 3 shows, respon-

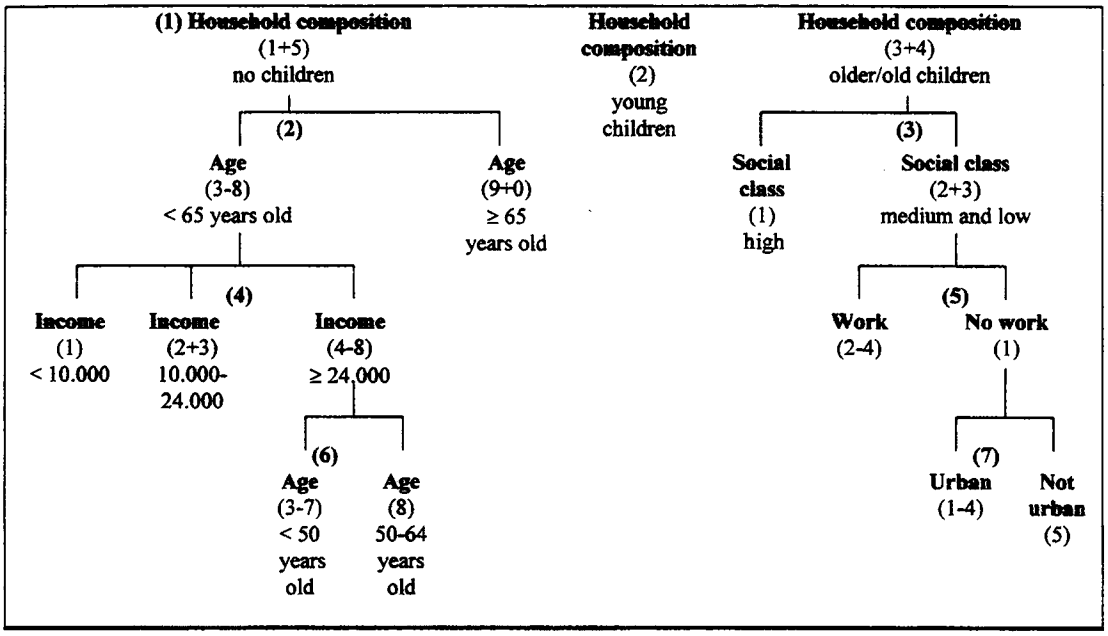

Figure 1. Results of CHAID.

Table 3
Relationship Between Predictor Variables and Cluster Membership

Predictor Variable/Cluster	Category/Percentage															
Household composition	1+5	2	3+4	1+5	1+5	3+4	3+4	1+5	1+5	1+5	3+4	3+4	1+5	1+5	3+4	3+4
Age				3–8	9+0			3–8	3–8	3–8			3–8	3–8		
Social class						1	2+3				2+3	2+3			2+3	2+3
Net annual household income								1	2+3	4–8			4–8	4–8		
Number of hours work per week											1	2–4			1	1
Age													3–7	8		
Urbanization															1–4	5
Cluster																
I domestic, low frequency	13.0	32.8	26.2	13.8	10.0	15.6	34.2	0.0	18.2	13.2	37.7	29.3	12.1	14.8	39.5	33.3
II abroad, high frequency	8.5	7.2	18.1	8.8	7.7	28.6	10.2	21.4	2.0	10.2	10.5	9.8	9.2	11.9	14.8	0.0
III mixed, low frequency	6.2	4.3	1.7	3.3	16.2	2.0	1.5	7.1	4.0	2.9	2.6	0.0	1.4	5.2	3.7	0.0
IV abroad, medium frequency, long duration	6.8	6.8	3.8	7.5	4.6	2.0	5.1	14.3	3.0	8.5	0.9	11.0	6.3	11.9	0.0	3.0
V abroad, medium frequency, shorter duration	12.6	6.0	6.4	13.4	10.0	7.5	5.6	7.1	5.1	16.1	7.9	2.4	13.5	20.0	7.4	9.1
VI abroad, medium frequency, longer duration	6.7	8.1	13.1	8.1	1.5	17.0	10.2	14.3	7.1	8.2	5.3	17.1	9.7	5.9	4.9	6.1
VII abroad, low frequency, short duration	24.8	15.7	20.1	26.2	20.0	21.1	19.4	21.4	32.3	24.6	21.9	15.9	30.4	15.6	23.5	18.2
VIII mixed, very low frequency	21.4	19.1	10.5	18.9	30.0	6.1	13.8	14.3	28.3	16.4	13.2	14.6	17.4	14.8	6.2	30.3
N	585	235	343	455	130	147	196	14	99	342	114	82	207	135	81	33

Explanation: For example, in the second column 1+5 means category 1 (single) and category 5 (household with adults only). Likewise, in the fifth column 3–8 means all categories 3–8 (15–64 years old). See Figure 1 for definition of categories.

dents in the nonworking segment were mainly domestically oriented or were going abroad for a short vacation, while people in the working segment tended to spend a longer foreign vacation. Moreover, respondents of the working segment went on holiday more often.

The percentages in the sixth row of Table 3 express the difference between those younger than 50 years of age and those between 50 and 64, of households without children in the highest income segment. Both predominantly spent their vacation abroad, but there were some differences in their vacation patterns in terms of frequency and duration.

Finally, the households of medium and low social class with old/older children without paid work were further split according to degree of urbanization. As Table 3 indicates, both belong primarily to Clusters I and VII. However, the vacation pattern of respondents living in nonurban areas was more inactive as expressed by the also high percentage for Cluster VIII.

Given these results, the categories of the predictor variables were merged as indicated in Table 4. In order to base the analysis on sufficient observations per cell, urbanization was not included in the subsequent analysis. Given the ordinal nature of the age variable, categories 1 and 2 were treated as belonging to the younger age segment, and merged with categories 3–8. These five predictor variables were included in the loglinear analysis to estimate the effects of these variables on the probability of belonging to a particular cluster. Because the aim of the analysis was to analyze these probabilities, the loglinear model was specified such that all interaction effects between the predictor variables were included in the model specification (see Everitt, 1977).

Table 5 suggests that the probability of belonging to Clusters I, II, VI, and VII was highest for households with older/old children. Remember that Clusters I and VII were clusters with a low vacation frequency, whereas Clusters II and VI had a medium to high frequency. Thus, it seems that these results pick up two different tendencies. On the one hand, people with older/old children belong to a cluster of relative inactive vacationers. On the other hand, these people have the time (and perhaps the money) to go on vacation more often.

The examination of Table 5 demonstrates that the social class variable had only three significant effects: people of a high social class had a significantly higher probability of belonging to Clusters II, V, and VI. This finding suggests that people of a higher social class tend to be more involved in foreign travel, and can afford to go on vacation more often. Furthermore, the fact that the vacation pattern of Cluster V was more spread over the year suggests that people of a higher social class are relatively less confined to specific time periods.

The effects for annual income indicate that the probability of belonging to Cluster I is highest for the high income segment. Clusters II, III, IV, and

Table 4
List of Most Important Predictor Variables and Their Categories

Variable	Categories
Household composition	1. household without children 2. household with young children (<6 years of age) 3. household with older/old children (6–17 years of age) (2–3: based on age of youngest child)
Social class	1. high 2. medium and low
Net annual household income	1. <Nlg 10,000 2. Nlg 10,000–24,000 3. ≥Nlg 24,000
Number of hours of work per week	1. 0 hours 2. ≥1 hours
Age	1. 0–64 years old 2. ≥65 years old

Table 5
Results of the Loglinear Logit Analysis

Variable	Category		Estimated Effect	z-Value
Household composition	without children	on Cluster I	−1.1184	−4.23
	young children	on Cluster I	−0.3825	−1.40
	without children	on Cluster II	−1.2149	−4.08
	young children	on Cluster II	−1.6058	−4.48
	without children	on Cluster III	0.1279	0.24
	young children	on Cluster III	0.2972	0.53
	without children	on Cluster IV	0.2391	0.61
	young children	on Cluster IV	−0.0644	−0.15
	without children	on Cluster V	0.2601	0.79
	young children	on Cluster V	−0.7356	−1.79
	without children	on Cluster VI	−1.0888	−3.52
	young children	on Cluster VI	−1.1365	−3.20
	without children	on Cluster VII	−0.2215	−0.86
	young children	on Cluster VII	−0.8739	−2.88
Social class	high	on Cluster I	−0.0765	−0.36
	high	on Cluster II	1.0325	4.19
	high	on Cluster III	−0.2454	−0.70
	high	on Cluster IV	0.4139	1.41
	high	on Cluster V	0.5507	2.20
	high	on Cluster VI	0.5466	2.11
	high	on Cluster VII	0.3017	1.48
Annual income	<Nlg 10,000	on Cluster I	−1.0351	−0.88
	Nlg 10,000–24,000	on Cluster I	−0.1848	−0.74
	<Nlg 10,000	on Cluster II	0.9110	1.06
	Nlg 10,000–24,000	on Cluster II	−1.0211	−2.56
	<Nlg 10,000	on Cluster III	0.2355	0.20
	Nlg 10,000–24,000	on Cluster III	−0.1241	−0.33
	<Nlg 10,000	on Cluster IV	0.4242	0.44
	Nlg 10,000–24,000	on Cluster IV	−1.3274	−2.61
	<Nlg 10,000	on Cluster V	−0.7091	−0.60
	Nlg 10,000–24,000	on Cluster V	−1.0225	−2.75
	<Nlg 10,000	on Cluster VI	0.6858	0.72
	Nlg 10,000–24,000	on Cluster VI	−0.5328	−1.45
	<Nlg 10,000	on Cluster VII	−0.3819	−0.45
	Nlg 10,000–24,000	on Cluster VII	−0.3025	−1.24
Work	0 hours	on Cluster I	0.3577	1.70
	0 hours	on Cluster II	0.2194	0.87
	0 hours	on Cluster III	0.4586	1.16
	0 hours	on Cluster IV	0.2326	0.77
	0 hours	on Cluster V	−0.0076	−0.03
	0 hours	on Cluster VI	−0.3214	−1.22
	0 hours	on Cluster VII	0.2931	1.40
Age	0–64 years old	on Cluster I	0.9883	2.60
	0–64 years old	on Cluster II	0.7177	1.65
	0–64 years old	on Cluster III	−0.9104	−2.07
	0–64 years old	on Cluster IV	1.0473	2.03
	0–64 years old	on Cluster V	0.7616	1.91
	0–64 years old	on Cluster VI	1.9201	2.49
	0–64 years old	on Cluster VII	0.9045	2.87

VII are dominated by the low income segment. Clusters V and VII show a strongly relationship with the high income segment.

The predictor variable "number of hours of work per week" differentiates between people with and without paid work. The estimated effects show that this variable was not significantly related to the classification into vacation types. The probability of belonging to the clusters was not significantly influenced by the work/nonwork variable, although nonworkers tended to be more inactive or spent their vacation domestically.

Finally, as far as age is concerned, Table 5 suggests that almost all listed effects were statistically significant. Older people tended to have a mixed vacation pattern as indicated by the higher positive effect of the people in the 65+ age cohort for Cluster III. The probability of this age cohort belonging to any one of the other clusters is lower.

Conclusions and Discussion

The present article has reported the results of a combined CHAID/loglinear analysis to examine the relationship between a set of socioeconomic variables and the probability of belonging to a particular vacation cluster. These clusters were distinguished on the basis of temporal and spatial characteristics of vacation histories, using a sequence alignment method. The combined CHAID/loglinear approach was suggested to circumvent the sparse cell problem, which results from the large number of cells in the multidimensional contingency tables, as implied by the large number of categories for the dependent and predictor variables. First, CHAID was used to find the subset of predictor variables and a new compound categorization of predictor variables that is more strongly related to the clustering of the respondents in terms of their vacation history data. Next, the resulting subset of predictor variables and the associated compound categorization were used as independent variables in a loglinear logit analysis to predict the probability that a particular respondent belongs to a particular cluster.

Our experiences with this methodology suggest that it represents a viable approach to address problems of this complexity. CHAID can be used at the prescreening stage to reduce the complexity of the problem to acceptable proportions. Loglinear models are very flexible in estimating the strength of the relationships. In our case, we decided to incorporate in the final analysis almost all variables that were significant in the preprocessing stage. One might argue that the selected number of variables is still too large. In that case, one simply needs to select only those variables that entered the splitting process earlier, or alternatively use more stringent significance criteria.

In substantive terms, our findings suggest that vacation history profiles are systematically related to a set of socioeconomic variables. The significant effects of the loglinear logit analysis are all easily interpretable and consistent with a priori expectations. Overall, the findings suggest that the "medium" clusters are also mixed in terms of the underlying variables as they relate to a smaller number of significant variables. For example, very few effects for Clusters III and IV were significant. It means that these clusters reveal vacation history patterns that are shared by many different socioeconomic groups. On the other hand, the results of the analyses indicate that especially the cluster with domestic vacations and the cluster with the highest vacation frequency are strongly related to particular socioeconomic profiles. Such results provide valuable clues for marketers to position their products and identify relevant user segments.

As with any analysis of this kind, the results are open to different interpretations. Moreover, as many of the predictor variables are interrelated, it is difficult to disentangle the specific contribution of any single variable. Furthermore, psychological variables such as motivation, routine-like behavior, level of involvement, and lifestyle variables cannot be captured as part of this kind of methodology. Therefore, to supplement the present findings, a series of qualitative interviews will be conducted in the next stage of the overall research project. The authors hope to report on the results of the additional qualitative analysis in future publications.

Acknowledgments

The research project was funded by the Cooperation Centre Tilburg and Eindhoven Universities

(SOBU). The authors also wish to thank the Netherlands Research Institute for Recreation and Tourism (NRIT) and the Netherlands Board of Tourism (NBT) for making available the CVO data for analysis. The constructive comments of two anonymous reviewers are highly appreciated. Any errors remain the responsibility of the authors.

References

Bargeman, B., Joh, C. H., & Timmermans, H. (in press). A typology of tourist vacation behavior using a sequence alignment method. *Annals of Tourism Research.*

Cha, S., McCleary, K. W., & Uysal, M. (1995). Travel motivations of Japanese overseas travelers: A factor-cluster segmentation approach. *Journal of Travel Research, 34*(1), 33–39.

Cohen, E. (1979). A phenomenology of tourist experiences. *Sociology, 13*, 179–201.

Etzel, M. J., & Woodside, A. (1982). Segmentation vacation markets: The case of the distant and near-home travelers. *Journal of Travel Research, 20*(4), 10–14.

Everitt, B. S. (1977). *The analysis of contingency tables.* London: Chapman and Hall.

Fakeye, P. C., & Crompton, J. L. (1991). Image differences between prospective, first-time, and repeat visitors to the Lower Rio Grade Valley. *Journal of Travel Research, 30*(2), 10–15.

Gitelson, R. J., & Crompton, J. L. (1984). Insights into the repeat vacation phenomenon. *Annals of Tourism Research, 11*(2), 199–217.

Kass, G. V. (1980). An exploratory technique for investigating large quantities of categorical data. *Applied Statistics, 29*(2), 119–127.

Kruskal, J. B. (1983). An overview of sequence comparison. In D. Sankoff & J. B. Kruskal (Eds.), *Time warps, string edits, and macromolecules: The theory and practice of sequence comparison* (pp. 1–44). London: Addison-Wesley.

Lang, C-T., O'Leary, J. T., & Morrison, A. M. (1997). Distinguishing the destination choices of pleasure travelers from Taiwan. *Journal of Travel & Tourism Marketing, 6*(1), 21–40.

Mazanec, J. A. (1984). How to detect travel market segments: A clustering approach. *Journal of Travel Research, 23*(1), 17–23.

Oppermann, M. (1997). First-time and repeat visitors to New Zealand. *Tourism Management, 18*(3), 177–181.

Sonquist, J. N., & Morgan, J. A. (1964). *The detection of interaction effects* (Monograph 35). Survey Research Centre, Institute for Social Research, University of Michigan.

Thompson, J. D., Higgins, D. G., & Gibson, J. T. (1994). CLUSTAL W: Improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Research, 22*(22), 4673–4680.

Willenborg, J. F., & Woodside, A. G. (1976). Segmentation of vacation attraction market: Some recursive models. In *Proceedings of the 7th Annual Conference of the Travel Research Association* (pp. 247–252). Salt Lake City: Bureau of Economic and Business Research, University of Utah.

Woodside, A. G., Cook, V. J., Jr., & Mindak, W. A. (1987). Profiling the heavy traveler segment. *Journal of Travel Research, 25*(4), 9–14.