International Journal of Mathematics and
Computer Science, **17**(2022), no. 1, 195–206

A Decision Tree for Information of Foreign Tourists Traveling to Thailand

Thidaporn Supapakorn[1], Sukanya Intarapak[2], Witchanee Vuthipongse[3]

[1]Department of Statistics
Faculty of Science
Kasetsart University
Bangkok, Thailand

[2]Department of Mathematics
Faculty of Science
Srinakharinwirot University
Bangkok, Thailand

[3]Department of Tourism
Ministry of Tourism and Sports
Bangkok, Thailand

email: fscitdps@ku.ac.th, sukanyain@g.swu.ac.th, witchanee@hotmail.com

(Received Received July 2, 2021, Accepted August 6, 2021)

Abstract

The objective of this research is to find the influencing variables for classification of foreign tourists' information in Thailand. The data of 400 foreign tourists were obtained from the Ministry of Tourism and Sports. By using decision tree analysis, the results show that
1) the length of stay can classify tourists by accurately predicting the expenditure per trip accounting for 76.0 %
2) age can categorize tourists with a correct prediction of travel frequency of 63.7 %
3) age, country of residence and travel arrangement can categorize

Key words and phrases: CHAID, classification, decision tree, prediction.
AMS (MOS) Subject Classifications: 90C29, 68U35.
Corresponding Author: sukanyain@g.swu.ac.th
ISSN 1814-0432, 2022, http://ijmcs.future-in-tech.net

tourists by accurately predicting gender accounting for 63.2 %
4) the length of stay and travel arrangement can classify tourists with 61.8% accurate predictions of the country of residence.

1 Introduction

In 2019, tourists spent 1,933,368.23 million Thai Bahts in Thailand. The average exchange rate in 2019 was 31.035 Thai Bahts for 1 U.S. Dollar. So, the revenue from tourism amounted to 62.3 billion U.S. Dollars. Average spending per tourist in 2019 was 48,580 Thai Bahts (approximately 1556.5 U.S. Dollars). Interestingly, there was a clear difference between spending by Asians and Europeans on the one hand and Americans on the other hand. The average Asian person (including Chinese and Japanese visitors) spent about 2/3 of what a European spent while on holiday in Thailand. The latter as mentioned before stayed longer in the country. Asian tourists provided the largest amount of revenue for Thailand as visiting tourists. However, the contribution of visitors from European countries certainly can not be disregarded. In 2019, revenue from Chinese visitors was 543.707 billion Thai Bahts, which indicates clearly why China is the most important country for the Thai tourism industry. However, European visitors were not that far behind at 461.478 billion Thai Bahts. From older data, the average tourist stayed about nine and a half days in Thailand. The per day tourists from different countries spent quite similar amounts of money. Total expenditure per tourist was 5,238 Bahts per day in 2016. Interestingly, the money spent for accommodation, per day per person (1530 Bahts), meant that most tourists acquired accommodation in 2 and 3-star hotels, rather than in fancy resorts. Thus, the important parameters are the number of tourists per region or country, the days spent in Thailand, and the money spent per day of stay. These parameters vary somewhat widely between regions and less so within countries of one particular region [1].

A decision tree can be used to help build automated predictive models which have applications in data mining, machine learning, and statistics. Known as decision tree learning, this method takes into account observations about an item to predict that item's value. To find the influencing variables of foreign tourists' information in Thailand, the decision tree with Chi-square Automatic Interaction Detection (CHAID) method is applied to classify the travel information in this research.

2 Literature Review

Most of the first-time visitors chose to visit Thailand due to the positive word-of-mouth.

In 2016, Wongleedee [3] investigated international tourists' destination loyalty from the perspective of international tourists in Bangkok as well as to study the level of interest to revisit Bangkok in the near future. A probability random sampling of 200 international tourists was utilized. Half the sample group were males and the other half were female. A Likert-five-scale questionnaire was designed to collect the data and small in-depth interviews were also used to obtain their opinions. The results from the study revealed that the majority of respondents had a medium level of loyalty. When examined in detail, the destination loyalty indicators can be ranked according to the mean average from high to low as follows: to recommend the visit, to say positive things, to revisit in the next three years, to refer the information, and to plan to visit regularly.

Díaz-Pérez and Bethencourt-Cejas [4] studied the segmentation of the tourism markets that had traditionally been undertaken by regression methods. CHAID (Chi-square Automatic Interaction Detection), which was more complex than other multivariate techniques, had rarely been used. This study applies the traditional methods of multivariate analysis and CHAID to the same population of tourists visiting a particular destination to compare the quality of the information obtained on tourism market segmentation. The results suggested that the analysis based on CHAID matches the nature of the problem studied better than those provided by discriminant analysis. In 2020, Díaz-Pérez et al. [5] considered the most suitable market segment(s) from an environmental and local economic development perspective in the specific context of visits to natural environments. By using the CHAID algorithm, a decision tree was constructed for means of transportation which serves as a key factor in the segmentation process. However, such a tree for visitors' resident or non-resident status cloud not be built as a first explicative variable, unless it was statistically forced. Once it was forced, the tree opened in several sub-segments, for non-residents and residents alike. Moreover, it allowed understanding of the means of transportation used by visitors according to their geographical origin as well as a set of added independent variables: accommodation establishment, length of stay, season, and other demographic variables (educational level, gender, and age).

3 Methodology

3.1 Data

The data of 400 foreign tourists in Thailand in 2019 were obtained from the Ministry of Tourism and Sports. In this study, the data included gender, age, country of residence, frequency, length of stay, expenditure per trip, travel arrangement, accommodation and quarter.

3.2 Decision Tree

A Decision Tree is a tree in which the nodes represent decisions, random transitions or terminal nodes, and the edges or branches are binary (yes/no, true/false) representing possible paths from one node to another. To use a decision tree for classification or regression, one grabs a row of data or a set of features and starts at the root and then through each subsequent decision node to the terminal node. The root or topmost node of the tree (only one root) is the decision node that splits the dataset using a variable or feature that results in the best splitting metric evaluated for each subset or class in the dataset that results from the split. The decision tree learns by recursively splitting the dataset from the root onwards according to the splitting metric at each decision node. The terminal nodes are reached when the splitting metric is at a global extremum.

A Chi-square Automatic Interaction Detection (CHAID) algorithm analysis is used to develop the decision tree models [6]. CHAID decision trees are nonparametric procedures that make no assumptions of the underlying data. This algorithm determines how continuous and/or categorical independent variables best combine to predict a binary outcome based on if-then logic by portioning each independent variable into mutually exclusive subsets based on homogeneity of the data. According to Kass [7], the CHAID algorithm operates using a series of merging, splitting, and stopping steps based on user-specified criteria [7]. CHAID analysis has several advantages over other methods of tourism market segmentation. First, Chi-square is a nonparametric statistic. Secondly, nominal and interval variables can be used as independent variables (predictors) in the model. Thirdly, continuous variables can be chosen as criterion variables, as they can be dichotomized, and finally, the criterion variable can be established according to the objectives of destination operators which increases the model's efficiency.

4 Results and Discussion

The decision tree with CHAID method is performed using IBM SPSS version 25. The descriptive statistics of foreign tourists' information in Thailand in 2019 are shown in Table 1. The findings of the study reveal that the majority of foreign tourists are Asian tourists (46.25%), between the ages of 25 and 34 years (46.00%) with their first visit traveling to Thailand (61.75%). More than half of foreign tourists (77.50%) travel to Thailand with a non-package and 87.75% of foreigner tourists stay at hotels/resorts. Moreover, the tourists stayed in Thailand 9.72 days on average with an average expenditure of 45,569.38 Thai Bahts per trip.

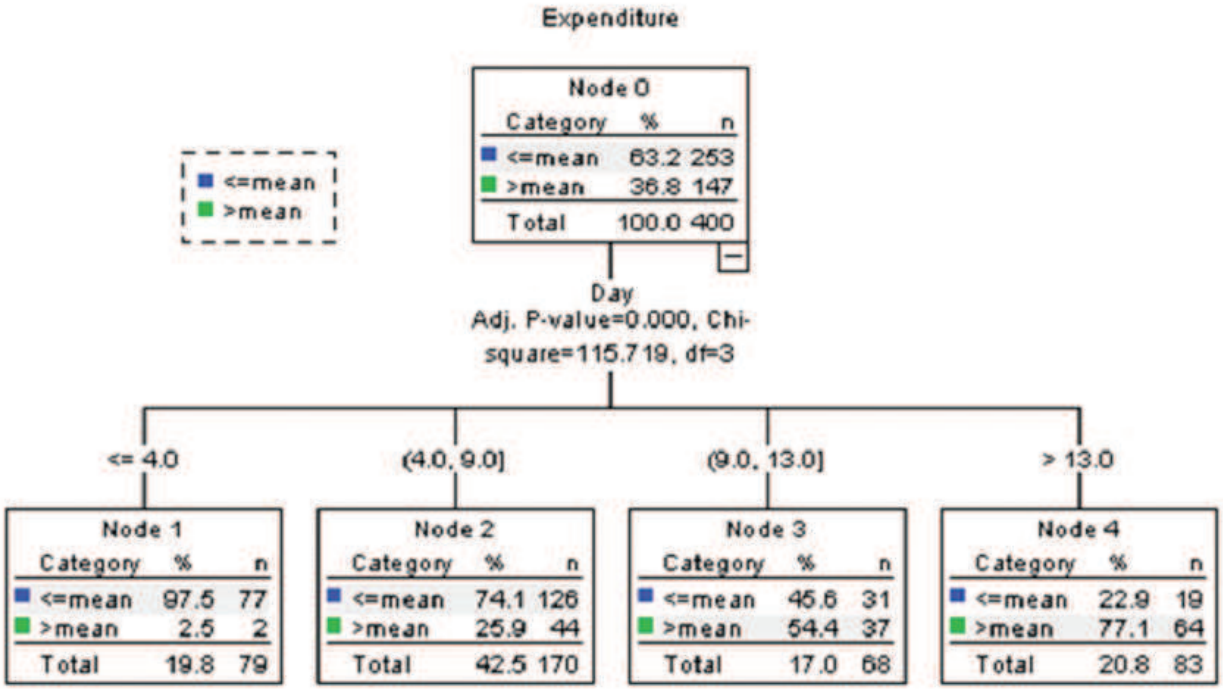

Figure 1: Decision tree of the expenditure per trip

The independent variables included gender, age, country of residence, frequency, length of stay, travel arrangement, accommodation and quarter and

Table 1: Descriptive statistics of foreign tourists' information in 2019

Variables	Frequency (%)	Mean ± Stand Deviation
Quarter		
I	102 (25.50)	
II	99 (24.75)	
III	89 (22.25)	
IV	110 (27.50)	
Country of Residence		
Asia	185 (46.25)	
Europe	106 (26.50)	
The Americas	33 (8.25)	
Oceania	26 (6.50)	
Middle East	33 (8.25)	
Africa	17 (4.25)	
Gender		
Female	199 (49.75)	
Male	201 (50.25)	
Age		
15 - 24 years	185 (46.25)	
25 - 34 years	184 (46.00)	
35 - 44 years	68 (17.00)	
45 - 54 years	33 (8.25)	
55 - 64 years	12 (3.00)	
65 years and above	1 (0.25)	
Frequency		
First Visit	247 (61.75)	
Revisit	153 (38.25)	
Length of Stay (days)		9.72 ± 7.75
Travel Arrangement		
Non package	310 (77.50)	
Package	90 (22.50)	
Accommodation		
Airbnb	9 (2.25)	
Friend's House	4 (1.00)	
Guesthouse	19 (4.75)	
Hotel/Resort	351 (87.75)	
Service Apartment	7 (1.75)	
Youth Hotel	9 (2.25)	
Other	1 (0.25)	
Expenditure per trip (Baht)		45,569.38 ± 3,247.32
$\leq$ mean	253 (63.25)	
$>$ mean	147 (36.75)	

Table 2: Percentage of classification of the expenditure per trip

Observed		Predicted	
	$\leq$ mean	$>$ mean	Percent Correct
$\leq$ mean	203	50	80.2
$>$ mean	46	101	68.7
Overall Percentage	62.3	37.8	76.0

the dependent variable is the expenditure per trip. Table 2 shows the percentage for predicting low expenditure ($\leq$ 45,569 Bahts/trip) correctly equal to 80.2%, meanwhile for predicting high expenditure ($>$ 45,569 Bahts/trip) correctly equal to 68.7%. The decision tree with CHAID method can predict the classification of the expenditure per trip with a precise maximum of 76.0% and provide the error of 24.0%. The tree diagram of CHAID method illustrated in Figure 1 shows that the biggest percentage of the expenditure per trip of less than or equal to 45,569 Bahts is on the 1st node (the low expenditure per trip and the length of stay of less than or equal to 4 days is 97.5%). On the 2nd node, the expenditure per trip of less than or equal to 45,569 Baht with the length of stay which is more than 4 days but less than or equal to 9 days is 74.1%. On the 3rd node, the expenditure per trip of more than 45,569 Bahts with the length of stay which is more than 9 days but less than or equal to 13 days is 54.4%. Meanwhile, the expenditure per trip of more than 45,569 Bahts with the length of stay which is more than 13 days is 77.1% on the 4th node.

Table 3: Percentage of classification of the frequency

Observed		Predicted	
	First Visit	Revisit	Percent Correct
First Visit	194	53	78.5
Revisit	92	61	39.9
Overall Percentage	71.5	28.5	63.7

When the independent variables are gender, age, country of residence, expenditure per trip, length of stay, travel arrangement, accommodation and quarter and the dependent variable is the frequency of traveling to Thailand. The decision tree with CHAID method can forecast the classification of the frequency of traveling to Thailand with an exact maximum of 63.7% and provide the error of 36.3%; whereas, the percentage for forecasting the first visit accurately equal to 78.5% and for forecasting the revisit accurately

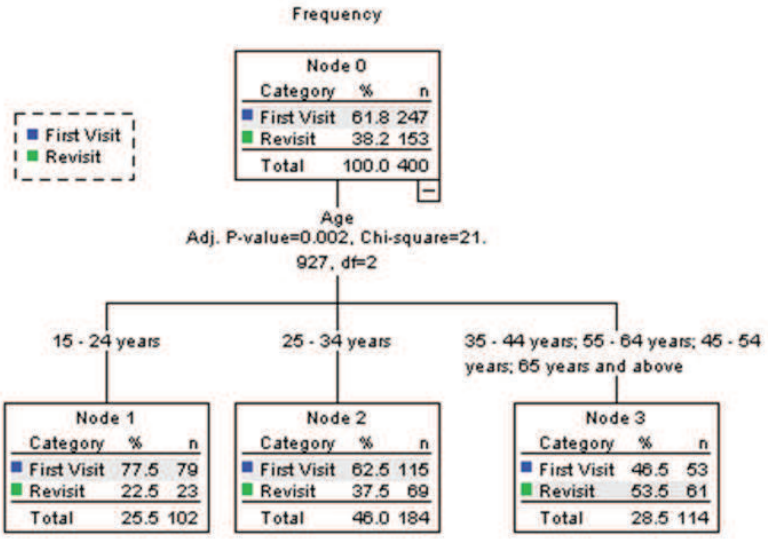

Figure 2: Decision tree of the frequency

equal to 39.9% (Table 3). In Figure 2, the tree diagram presents that the percentage of the first visit and age of between 15 and 24 years is 77.5% on the 1st node. The first visit with age which is between 25 and 34 years old is 62.5% on the 2nd node; while the revisit with age, which is more than 34 years old is 53.5% on the 3rd node.

Table 4: Percentage of classification of gender

Observed		Predicted	
	Female	Male	Percent Correct
Female	106	93	53.3
Male	54	147	73.1
Overall Percentage	40.0	60.0	63.2

Whereas the dependent variable is gender and the independent variables include age, country of residence, frequency, expenditure per trip, length of stay, travel arrangement, accommodation and quarter, Table 4 shows the decision tree with CHAID method can predict the classification of gender with an exact maximum of 63.2% and provide the error of 36.8%. The

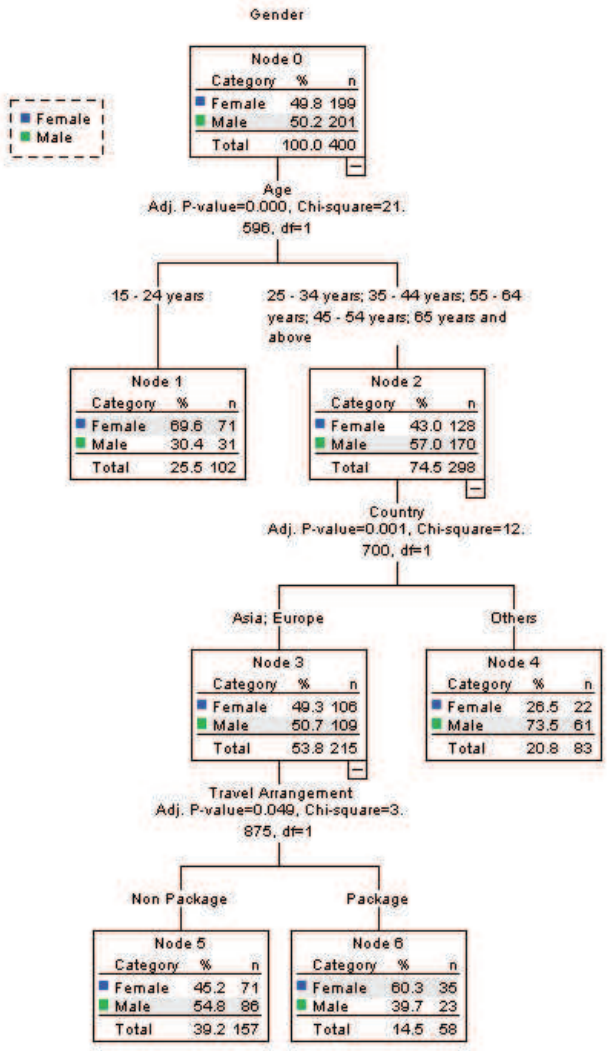

Figure 3: Decision tree of gender

percentage of predicting male accurately equals 73.1% and that of female accurately equals 53.3%. In Figure 3, the tree diagram of CHAID method shows that the percentage of female and age of between 15 and 24 years old is 69.6% (the 1st node). On the 4th node, male from the Americas, Oceania, Middle East or Africa is 73.5%. On the 5th node, males from Asia or Europe with a non-package of travel is 54.8%. Moreover, females from Asia or Europe

with a package of travel is 60.3% (the 6th node).

Table 5: Percentage of classification of the country of residence

Observed		Predicted		
	Asia	Europe	Others	Percent Correct
Asia	153	11	21	82.7
Europe	18	62	26	58.5
Others	36	41	32	29.4
Overall Percentage	51.7	28.5	19.8	61.8

While the country of residence is the dependent variable and the independent variables are age, gender, frequency, expenditure per trip, length of stay, travel arrangement, accommodation and quarter, the percentage for forecasting Asia correctly equal to 82.7%, the percentage for forecasting Europe correctly equal to 58.5%, and the percentage for forecasting others correctly equal to 29.4%; whereas the decision tree with CHAID method can predict the classification of the country of residence with a precise maximum of 61.8% and provide the error of 38.2% (Table 5). The tree diagram in Figure 4 presents that the percentage of Asian tourists is 66.9% on the 4th node (the non-package of travel and the length of stay less than 7 days). On the 5th node, Asian tourists with the non-package of travel and the length of stay less than 7 days are 96.0%. On the 3rd node, Europe tourists with the length of stay more than 11 days are 54.4%. Moreover, other tourists with the length of stay more than 7 days but less than or equal to 11 days are 40.5% (the 2nd node).

The recommendations of the findings from the results using the decision tree for travel information are most foreign tourists who stay in Thailand more than 9 days will have the expenditure of travel more than 45,569 Bahts per trip. Most foreign tourists over 34 years old will travel to Thailand repeatedly. As for foreign tourists aged between 15-34 years, they will travel to Thailand for the first time. Therefore, the Ministry of Tourism and Sports should have an appropriate marketing to foreign tourists of each group, such as a tourism promotion in popular cities (Bangkok, Chiang Mai or Phuket) for foreign tourists who are between 15-34 years old because most of these people come to Thailand in the first visit. In addition, they should focus on promoting tourism in other cities for foreign tourists who are 34 years old and above because most of these people have traveled to Thailand at least once. Most foreign tourists who are 24 years old and above come from Asia and Europe and travel in Thailand with a travel agency and are female tourists.

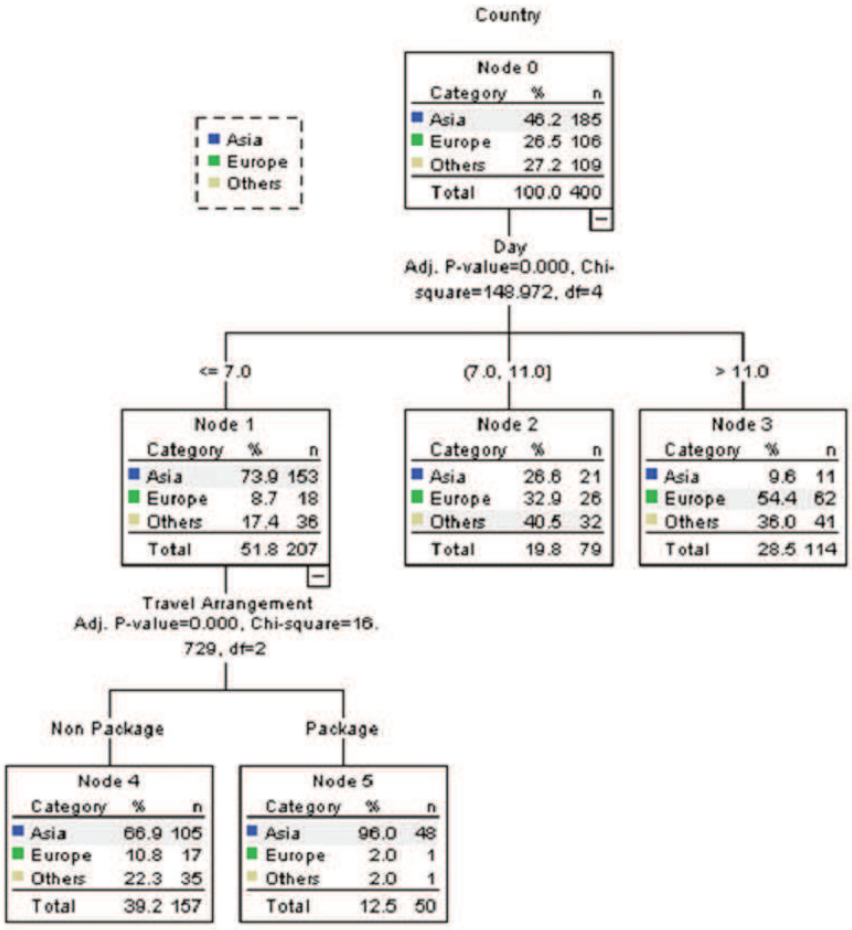

Figure 4: Decision tree of the country of residence

Therefore, the travel agencies should design tourism programs to fit those tourists. Moreover, the travel agencies should promote to tourists from Asia and Europe. Furthermore, most foreign tourists who stay in Thailand less than or equal to 7 days and travel with a travel agency are Asian tourists. Therefore, the travel agencies should design travel programs in Thailand with less than 7 days for them.

5 Conclusion

This research provided valuable information for tourism management since the decision tree correctly classifies more than 60% of travel information for foreign tourists traveling to Thailand. The findings of the study can serve as guidelines for modifying tourism management and marketing policies, strategies and plans, used by relevant parties such as the government and travel agencies in order to attract more foreign tourists in the future.

Acknowledgment. This research was supported by the Program Management Unit for Human Resources & Institutional Development (B05F630034). In addition, we would like to give credit to the Ministry of Tourism and Sports for providing the data.

References

[1] Tourism Statistics Thailand, Revenue from foreign Tourists visiting Thailand, 2020, https://www.thaiwebsites.com/tourism.asp.

[2] H. M. Htun, S. Padungyoscharoen, S. San, Influences of motivation toward revisit intention, destination loyalty and positive word-of-mouth, Apheit journal, **4,** no. 2, (2015), 115–130.

[3] K. Wongleedee, An examination of international tourists' destination loyalty: a case study of international tourist in Bangkok, International Journal of Management and Applied Science, **2,** no. 11, (2016), 151–154.

[4] F. M. Díaz-Pérez, M. Bethencourt-Cejas, CHAID algorithm as an appropriate analytical method for tourism market segmentation, Journal of Destination Marketing & Management, **5,** no. 3, (2016), 275–282.

[5] F. M. Díaz-Pérez, C. G. García-Gonzalez, A. Fyall, The use of the CHAID algorithm for determining tourism segmentation: A purposeful outcome, Heliyon, **6,** (2020), 1–11.

[6] G. V. Kass, Significance Testing in Automatic Interaction Detection, Ph.D. thesis. University of Weiwaterstrand, South Africa, 1975.

[7] G. V. Kass, An exploratory technique for investigating large quantities of categorical data, Applied Statistics, **29,** no. 2, (1980), 119–127.