


Fall 2014

Application Of Bayesian Networks In Consumer Service Industry

Yuan Gao
Purdue University

Follow this and additional works at: https://docs.lib.purdue.edu/open_access_theses

 Part of the [Industrial Engineering Commons](#), [Operational Research Commons](#), [Recreation, Parks and Tourism Administration Commons](#), and the [Statistics and Probability Commons](#)

Recommended Citation

Gao, Yuan, "Application Of Bayesian Networks In Consumer Service Industry" (2014). *Open Access Theses*. 325.
https://docs.lib.purdue.edu/open_access_theses/325

This document has been made available through Purdue e-Pubs, a service of the Purdue University Libraries. Please contact epubs@purdue.edu for additional information.

**PURDUE UNIVERSITY
GRADUATE SCHOOL
Thesis/Dissertation Acceptance**

This is to certify that the thesis/dissertation prepared

By Yuan Gao

Entitled
APPLICATION OF BAYESIAN NETWORKS IN CONSUMER SERVICE INDUSTRY

For the degree of Master of Science in Industrial Engineering

Is approved by the final examining committee:

Vincent G. Duffy

Mark Lehto

Thomas KuczekAbhi

To the best of my knowledge and as understood by the student in the Thesis/Dissertation Agreement, Publication Delay, and Certification/Disclaimer (Graduate School Form 32), this thesis/dissertation adheres to the provisions of Purdue University's "Policy on Integrity in Research" and the use of copyrighted material.

Vincent G. Duffy

Approved by Major Professor(s): _____

Approved by: Abhijit Deshmukh

10/27/2014

Head of the Department Graduate Program

Date

APPLICATION OF BAYESIAN NETWORKS IN CONSUMER SERVICE
INDUSTRY

A Thesis

Submitted to the Faculty

of

Purdue University

by

Yuan Gao

In Partial Fulfillment of the

Requirements for the Degree

of

Master of Science in Industrial Engineering

December 2014

Purdue University

West Lafayette, Indiana

ACKNOWLEDGEMENTS

I would like to thank my advisor, Prof. Vincent G. Duffy for the valuable guidance, advice and encouragement he has provided to me throughout the past 2 years of my Master's program. His help and support was invaluable and will remain a great asset for my future work. I would also like to thank the other committee members Prof. Thomas Kuczek and Prof. Mark Lehto, for taking time to review my thesis, participate my final defense and for their great inputs.

Thanks to my research colleagues Onan Dimerel, Le Zhang, Vivia Chao and Byeongho Lee. It was an honor working with them. They inspired and motivated me with their experience sharing and intelligence. Also thanks to Ching-Wei Cheng for his great help and professional knowledge in statistics.

My sincere gratitude goes to the Hawaii Tourism Authority's Tourism Research Division for collecting and publishing the tourism research and report data on their official website. These data allowed me to undertake this research in an area that I have a personal interest in.

Finally, I want to thank the School of Industrial Engineering at Purdue University and the faculty and staff who gave me the chance to obtain the knowledge needed in carrying out this study.

TABLE OF CONTENTS

	Page
LIST OF TABLES	vii
LIST OF FIGURES	ix
ABSTRACT	xii
CHAPTER 1. INTRODUCTION.....	1
1.1 Overview	1
1.2 Organization of the Document.....	2
1.3 Background	3
1.3.1 Big Data Challenge	3
1.3.2 Service Industry Challenge	4
1.4 Purpose of the Study	6
1.5 Assumptions and Hypothesis	7
CHAPTER 2. LITERATURE AND CONTRIBUTIONS.....	9
2.1 Theoretical Contributions.....	9
2.1.1 Bayesian Network Modeling.....	9
2.1.1.1 Literature and Gap	9
2.1.1.2 Contribution: Using Aggregate Data	11
2.1.2 Consumer Behavior Analysis	11
2.1.2.1 Literature and Gap	11
2.1.2.2 Contribution: Using Unbiased Data.....	12
2.2 Practical Contributions.....	12
2.2.1 Literature and Gap	12
2.2.1.1 I/O Model of Consumer Service Process	12
2.2.1.2 Features of the Travel and Tourism Industry	16

	Page
2.2.2	Contribution: Holistic Model, Applicable Recommendations . 18
CHAPTER 3.	METHODOLOGY 20
3.1	Multivariate Analysis..... 20
3.2	Machine Learning..... 22
3.2.1	Typology of Machine Learning..... 22
3.2.2	Choice of Techniques and Methods..... 24
3.3	Theoretical Basis: Bayesian Networks 25
3.3.1	Bayes' Theorem..... 25
3.3.2	Introduction of Bayesian networks 26
3.3.2.1	Example of A Simple Bayesian Model 28
3.3.2.2	Features of Bayesian Network 32
3.3.2.3	Bayesian Network and Artificial Neural Network..... 34
3.3.3	Information Theory and Statistics Theory 38
3.3.3.1	Mutual Information 38
3.3.3.2	Pearson Correlation Coefficient 38
CHAPTER 4.	STUDY DESIGN 40
4.1	Design of Analysis..... 40
4.1.1	Hawaii Tourism Industry 40
4.1.1.1	Slow-Down and Decline..... 41
4.1.2	Data 44
4.1.2.1	Data Preprocessing for Analysis 46
4.2	Modeling Tool: BayesiaLab 48
4.2.1	Introduction of BayesiaLab and Features 50
4.2.1.1	Algorithm..... 51
4.2.1.2	Limitations with Trial Version 54
CHAPTER 5.	RESEARCH APPROACH AND RESULTS 55
5.1	Step One: Initial Predictor Ranking and Screening 55
5.1.1	Ranking Criteria and Method 56
5.1.1.1	K-Means Clustering 56

	Page
5.1.2	Step One Results..... 59
5.2	Step Two: Unsupervised Learning..... 61
5.2.1	Unsupervised Learning and Supervised Learning 61
5.2.1.1	Unsupervised Learning Algorithm 63
5.2.2	Step Two Analysis 63
5.2.2.1	Posterior Probability Inference 66
5.2.3	Predictor Ranking for Each MMA..... 69
5.2.4	Step Two Results..... 71
5.3	Step Three: Supervised Learning 72
5.3.1	Supervised Learning Algorithm 73
5.3.2	Step Three Analysis and Results 73
5.3.2.1	U.S. West..... 74
5.3.2.2	U.S. East..... 77
5.3.2.3	Japan 80
5.3.2.4	Canada 83
5.4	Validation..... 86
5.4.1	Cross Validation within Original Data Set 87
5.4.1.1	Validation of Unsupervised Model..... 90
5.4.1.2	Validation of Supervised Model..... 97
5.4.2	Validation with An Additional Data Set..... 113
5.4.2.1	Validation of Unsupervised Model with Additional Data Set 113
5.4.3	Validation Summary..... 116
5.5	Results Summary 117
5.5.1	MMA 118
5.5.2	Travelling Season 119
5.5.3	Choice of Accommodation 120
5.5.4	Purpose of Travel..... 122
5.5.5	Repeat Visitor 124
5.5.6	Travel Style 124

	Page
CHAPTER 6. CONCLUSION	126
6.1 Hypothesis Validation	126
6.2 Conclusions	127
CHAPTER 7. DISCUSSION	129
7.1 Visitor Origin	130
7.2 Purpose of Travel	133
7.3 Consideration of Cross Validation	135
7.3.1 Sample Size	135
7.3.2 Consideration of Outlier	136
7.4 Limitations and Future Work.....	136
REFERENCES	141
APPENDICES	
Appendix A Permission for the Use of Data.....	148
Appendix B Purdue IRB Approval.....	155

LIST OF TABLES

Table	Page
Table 3.1 Marginal and Joint Probabilistic distribution table of Lung Cancer and Smoker	29
Table 3.2 Similarities and Differences between Bayesian Network and Artificial Neural Network.....	36
Table 4.1 Example of Monthly Visitor Highlight Raw Data.....	46
Table 4.2 Example of Converting Absolute Values to Percentage	47
Table 4.3 Complete List of Variables and Definition	48
Table 5.1 Ranked Predictors by MI with Each Outcome	60
Table 5.2 Reduced Predictor Set after Step 1	61
Table 5.3 Top 5 Factors by MI for Each Outcome and MMA.....	71
Table 5.4 Data Set for Each MMA After Step 2	72
Table 5.5 Posterior Probability distribution of U.S. West_Supervised	76
Table 5.6 Posterior Probability distribution of U.S. East_Supervised	79
Table 5.7 Posterior Probability distribution of Japan_Supervised.....	82
Table 5.8 Posterior Probability distribution of Canada_Supervised.....	85
Table 5.10 Predictors Ranking by Mutual Information with Each Outcome and Average_Training Data Set	91

Table	Page
Table 5.12 Predictors Ranking by Mutual Information with Each Outcome and Average _Testing Data Set	95
Table 5.14 Top 5 Factors by MI for Each Outcome and MMA_Training Data Set	98
Table 5.15 List of Factors for Supervised Learning_Training Data Set	99
Table 5.16 Posterior Probability distribution of U.S. West_Supervised_Training Data Set	100
Table 5.17 Posterior Probability distribution of U.S. East_Supervised_Training Data Set	101
Table 5.18 Posterior Probability distribution of Japan_Supervised_Training Data Set	102
Table 5.19 Posterior Probability distribution of Canada_Supervised_Training Data Set	103
Table 5.20 Target Node/State Influences of U.S. West_Testing Data Set	104
Table 5.21 Target Node/State Influences of U.S. East_Testing Data Set	105
Table 5.22 Target Node/State Influences of Japan_Testing Data Set.....	106
Table 5.23 Target Node/State Influences of Canada_Testing Data Set.....	107
Table 5.24 Supervised Model Validation of U.S. West	109
Table 5.25 Supervised Model Validation of U.S. East	109
Table 5.26 Supervised Model Validation of Japan.....	111
Table 5.27 Supervised Model Validation of Canada.....	112
Table 5.32 Complete List of Variables and Definition	118

LIST OF FIGURES

Figure	Page
Figure 4.1 Traditional I/O Model	13
Figure 4.2 Service I/O Model	14
Figure 4.3 Consumer Service Industry Framework	15
Figure 4.4 Airline seat occupancy of flights to Hawaii from Shanghai	17
Figure 5.1 Supervised Learning Algorithms in Two Phases	23
Figure 5.2 Example of a Two-Node Bayesian Network	29
Figure 5.3 An extended example of Bayesian network.....	31
Figure 5.4 An example of a multilayer feed-forward neural netowrk.....	33
Figure 6.1 Hawaii Tourist Arrivals by Air 1951-2005.....	43
Figure 6.2 Hawaii Visitor Expenditures 1951-2005.....	44
Figure 7.1 Analysis Procedures.....	55
Figure 7.2 Comparison of Different Discretization Binning Selections.....	59
Figure 7.3 Unsupervised Learning Model.....	64
Figure 7.4 Node Force Mapping	66
Figure 7.5 Posterior Probability distribution of MMA given Highest Arrivals	67
Figure 7.6 Posterior Probability distribution of MMA given Highest Avg_Stay....	67
Figure 7.7 Posterior Probability distribution of MMA given Highest Exp_pp/D ...	68
Figure 7.8 Visitor Characteristics Posterior Distribution: Japan v.s. US West	69

Figure	Page
Figure 7.9 Supervised Model for U.S. West	77
Figure 7.10 Supervised Model for U.S. East.....	80
Figure 7.11 Supervised Model for Japan.....	83
Figure 7.12 Supervised Model for Canada	86
Figure 7.13 3-Year Trend_U.S. West	89
Figure 7.14 Unsupervised Learning Model_Training Data Set	92
Figure 7.15 Posterior Probability distribution of MMA given Highest Arrivals_ Training Data Set.....	93
Figure 7.16 Posterior Probability distribution of MMA given Highest Avg_Stay_Training Data Set	93
Figure 7.17 Posterior Probability distribution of MMA given Highest Exp_pp/D_Training Data Set.....	93
Figure 7.18 Visitor Characteristics Posterior Distribution: Japan v.s. US West_Training Data Set	94
Figure 7.19 Arrivals of 4 Regions_Testing Data Set.....	96
Figure 7.20 Average Lengths of Stay (day) of 4 Regions_Testing Data Set	96
Figure 7.21 Daily Expenditure per Person (\$) of 4 Regions_Testing Data Set...	97
Figure 7.22 Visitors Chracteristics: Japan v.s. U.S. West_Testing Data Set.....	97
Figure 7.23 Arrivals of 4 Regions_Additional Data Set.....	114
Figure 7.24 Average Lengths of Stay (day) of 4 Regions_Additional Data Set	114
Figure 7.25 Daily Expenditure per Person (\$) of 4 Regions_Additional Data Set.....	115

Figure	Page
Figure 7.26 Visitors Characteristics: Japan v.s. U.S. West_Additional Data Set	116
Figure 7.27 Choices of Accommodation by Percentage.....	121
Figure 7.28 MMA Likelihood Given High Stay_F&R% and POT_Vst%	122
Figure 7.29 Purpose of Travel by Percentage	123
Figure 9.1 Quote from Survey Feedback.....	139
Appendix Figure	
Figure A.1 Permission of Data Use by Hawaii Tourism Authority_1	153
Figure A.2 Permission of Data Use by Hawaii Tourism Authority_2	154
Figure B.3 IRB Approval for Conducting Survey with Service Providers_1	155
Figure B.4 IRB Approval for Conducting Survey with Service Provider_2	156

ABSTRACT

Gao, Yuan. M.S.I.E., Purdue University. December 2014. Application of Bayesian Networks in Consumer Service Industry. Major Professor: Vincent G. Duffy.

The purpose of the present study is to explore the application of Bayesian networks in the consumer service industry to model causal relationships within complex risk factor structures using aggregate data. An analysis of the Hawaii tourism market was conducted to find out how visitor characteristics affect their behavior and experience as consumers during the trips, and influence the tourism market outcomes represented by measurable factors. Two hypotheses were proposed regarding the use of aggregate data and the influence of visitor origin, and were verified through the analysis. The source data came from the Hawaii Tourism Authority's official website, including monthly tourists highlight reports over a period of 36 months. The analysis verified the hypotheses that visitor origin, as a symbol of cultural background, plays an important role in their behavior, preferences, decisions and experience in consuming. The results were validated both statistically and against literature and expert opinion. In the increasingly segmented tourism market, such findings can help tourism service providers improve consumer satisfaction and loyalty with assistance in policy-making, investment decision-making, resource planning, and strategic marketing.

CHAPTER 1. INTRODUCTION

1.1 Overview

Complex systems widely exist in business, industry and society nowadays. According to Maglio et al (Maglio, Vargo, Caswell, & Spohrer, 2009), the service system is a configuration of people, technologies, and other resources that interact with other service systems to create mutual values. It is a highly interactive and knowledge-based sector where the maximum output relies on a comprehensive understanding of how the factors in the networks influence each other.

Bayesian network, also known as Bayesian belief network, is a graphical model representing conditional probabilistic dependencies (Han, Kamber, & Pei, 2012c). Backed by information theory and learning algorithms, Bayesian network has seen extensive applications in data mining, especially for complicated systems involving association and causal relationships yet to be unveiled. Conventionally, the network topology is built up based on a set of individual data or expert knowledge (Conrady & Jouffe, 2013b), but these are not always feasible to obtain in reality.

The travel and tourism industry is a service sector involving a wide range of elements that interact with each other. Through an application of Bayesian network in the tourism market of Hawaii, this study will demonstrate how to model a multi-factor system based on existing aggregate data and how to interpret the model. The analysis provided qualitative and quantitative representation of how pairs of variables interact in an omni-directional network by examining the posterior probability distribution given prior condition settings.

1.2 Organization of the Document

The rest of this document consists of chapters two through seven. Chapter 2 (Literature and contributions) provides a literature review of the existing research work and gaps which this study is proposed to fill. Chapter 3 (Methodology) introduces the theoretical background this study has stemmed from, including Bayesian networks and information theory. Chapter 4 (Study Design) describes in detail the problem settings of the study, the source and preparation of the data, and the modeling software BayesiaLab. In Chapter 5 (Research Approach and Results), the analysis procedures were introduced step by step, with the results accompanying to explain how the research was conducted and why so. At the end of this chapter, the key findings were summarized and validated statistically using cross validation and an additional data set. Chapter 6 (Conclusion) concludes the results and verifies the initial hypotheses. Finally, Chapter 7 (Discussion) further interprets the relationships unveiled by the Bayesian network models, and validates the results against literature and expert opinions. This

chapter also went through the limitations in the analysis and validation process, as well as the future work.

1.3 Background

1.3.1 Big Data Challenge

In the digital age, across a wide variety of fields, data are being collected and accumulated at a dramatic pace (Fayyad, Piatetsky-shapiro, & Smyth, 1996). From the daily life of ordinary people to business, scientific, politics, military sectors, massive data are generated, logged and stored every second. This explosive growth of available data volume is a result of the computerization of our society and the fast development of powerful data collection and storage tools (Han, Kamber, & Pei, 2012a). However, it is a challenge to best utilize and correctly interpret these data to draw out valuable information. As Dr. William Cook said during an interview (Cook & IIE Annual Conference & Expo, 2014): “How to best utilize the ever-increasing amounts of available data” is the most pressing challenge in the field of industrial and systems engineering today.

There are several contributors to this challenge being so tremendous. First, the volume and the speed of accumulation creates dauntingly gigantic database impossible for manual analysis. Second, in some cases, immediate data feeding and analysis is needed to project the fast-changing trends (for example: the stock market). Third, many systems are so complicated that no individual expert has the knowledge to resolve them by him/herself.

The traditional method of turning data into knowledge relies on manual analysis and interpretation, and the classical approach to data analysis relies fundamentally on one or more analysts becoming intimately familiar with the data and serving as an interface between the data and the users and products (Fayyad et al., 1996). These methods no longer satisfy the needs today.

Without powerful tools, large data repositories become “data tombs”—data archives that are seldom visited. Moreover, misinterpreted data can lead to misguided decisions and unwanted consequences. There is a need for methodologies and tools at least partially automated to assist human in this task.

1.3.2 Service Industry Challenge

As in any business, the service industry, also known as tertiary sector of industry, is an arena where the buyer pays the seller in exchange of products. The difference is that instead of extracted natural resources (as in the Primary Industry) or manufactured goods (as in the Secondary Industry), the suppliers earn revenue through intangible products and services (BusinessDictionary, n.d.; Wikipedia, 2013). It includes a wide range of sectors from quasi-manufacturing systems with low customer contact (for example, financial institutes, wholesale, postal service) to pure service systems with high customer contact (for example, health centers, hotels, schools) (Chase, 2010). According to Chase, the extent of required customer contact in the creation of the service product distinguishes one service system from another. Consequently, higher customer contact systems

are more difficult to control and rationalize due to the involvement of the customer.

Customer needs and customer expectations are central to service businesses being able to create the satisfaction and loyalty they require for sustainable competitive advantage (Schneider & Bowen, 2010). The higher customer contact a service sector has, the more important it is to understand the customers and their role in the system in order to gain success and profits. The understanding needs to be presented in a way that can be integrated into the design and operations of the industry, which, in some cases, means to reengineer the systems.

Most of the studies of the service profit chain relationships rely on large amounts of data. This may require that researchers relinquish control over the collection of at least a portion of the data needed, relying on already-existing data in organization under study (Heskett & Sasser, 2010). Due to the difficulty in maintaining consistency and obtaining data access, factor analysis regarding retrospective or prospective behaviors are used more frequently than longitude case-effect study (Heskett & Sasser, 2010).

Conventionally, data collection for behavior analysis targets individual subjects using methods including observation or self-report, such as survey and questionnaire (Fishe, Groff, & Roane, 2011). But in practice, these methods may suffer lack of reliability due to subjectivity on both sides - the researcher and the

responder - let alone the difficulty and costs in conducting data collection and obtaining valid responses. On the other hand, existing data are not utilized because they don't meet the criteria of research methods, especially when they are aggregate data without personal identification.

1.4 Purpose of the Study

Given the challenges mentioned above, there is a need for an effective method to help suppliers in the service industry understand what the consumers need and why. Due to the structure of the industry, the answers must come from a comprehensive and systematic study of all the factors and relationships in the service dynamics. The study needs not to initiate another effort to collect self-report data from consumers, but rather makes good use of the existing data and interpret them in an innovative method.

The similar approach has been tested in causal relationship analysis for road traffic volume and mental health (L. Zhang, Gao, Bidassie, & Duffy, 2014). In this conference paper, the researchers initiated an effort to apply Bayesian network in two case studies: In the first case, individual instances of daily vehicle miles traveled (DVMT) in each county in the state of Indiana from 2006 to 2010 were used to find out that road type has the most significant impact on DVMT. A Bayesian network was built based on the learning algorithms and the training data set. In the second case, a network model was constructed using existing causal relationships from a prior study on veterans' mental health, and the

aggregate data set in the study was used to test the model using Bayesian Belief network algorithms. The model showed similar inference results as the original study.

Inspired by these two case studies, the author attempted to take the application of Bayesian networks to a further depth and larger scale, and most importantly, to use aggregate data in an application that's similar to the environment of the first case study (DVMT analysis). That is, to build a network model without existing knowledge using aggregate data.

The purpose of this study is to explore the application of Bayesian networks in analyzing relationships among multiple factors in the consumer service industry, and to verify the analysis approach using aggregate data instead of individual data.

Through a case study of the travel and tourism sector in Hawaii, this study will develop a systematic approach to model the relationships among multiple factors, and examine the results. Recommendations for the industry will be provided based on interpretation of the results.

1.5 Assumptions and Hypothesis

This study intends to approach a service system based on the assumption that no prior knowledge is available about the relationships and interactions among multiple factors in the system. It is through the data analysis that such information will be obtained. In a specific real situation, professional opinions and

expert experience may exist to help guide the analysis or interpret the results.

But in this study, the main purpose is to explore an approach to draw information from existing data as a generic method.

At the beginning stage of the study, two hypotheses were formed for the research:

1. Aggregate data can be used as input to Bayesian networks to analyze complex system and provide valuable insights on the relationships among multiple factors.
2. In the travel and tourism section, visitors from different regions have different behaviors which will affect the outcomes evaluated by measurable metrics, such as arrivals, length of stay, expenditure.

CHAPTER 2. LITERATURE AND CONTRIBUTIONS

This chapter introduces the literature of the relevant research and application fields and identifies and existing gaps. For each gap, the contributions of this study will be discussed.

2.1 Theoretical Contributions

This study presents significance in two ways in the theoretical research areas.

2.1.1 Bayesian Network Modeling

2.1.1.1 Literature and Gap

Bayesian network models present the probabilistic inference of uncertainties between variables by deriving posterior probabilities based on prior probabilities. In statistics, the probability density indicates the distribution of individual events in a sample. If it is not possible to construct a probability distribution function due to data availability problems, expert knowledge or experience may help. For example, to forecast the performance of a stock, data of the historic prices are needed to project the trend, and/or information and knowledge of the stock market and global events should be considered.

When human subjects are involved, individual events data often include the attributes of each individual person. This adds to the difficulties in data collection because such data are often aggregated due to data confidentiality, the

protection of privacy or the limited size of database (Park, 2011). In marketing and economics, many researchers have relied on aggregate data to understand consumer choices and preferences because they are cheaper and easier to get. While limited by data availability, the analysis of consumer demands is conducted using aggregate consumption and expenditure data which are typically all that is available to draw conclusions based on the theory of individual consumer behavior (Cranfield, 1999). But researchers have reported that the knowledge obtained from individual survey could be rejected through aggregate data analysis, indicating that the individual theories or assumptions don't always fit when considering consumers as a group (Cranfield, 1999; Sabelhaus, 1990). Musalem et al (Musalem, Bradlow, & Raju, 2009) argued that the traditional use of aggregate data did not incorporate heterogeneity, and proposed a method of using Bayesian methods normally 'reserved' for data that arrive in the form of individual-level choices, for estimating demand models from aggregate market share data. This method was further developed in other customer choices research (Park, 2011; Rutz & Trusov, 2011), but the models were all based on a simple choice scenario: only one choice (purchase) is made at a time and no other factors in the service system was included (customer characteristics, environmental factors, supplier inputs, etc).

2.1.1.2 Contribution: Using Aggregate Data

The method introduced in this thesis is innovative in the sense that it uses aggregate data as input for a Bayesian network model, instead of a simple choice problem. To the author's best knowledge, it is the first in this area.

2.1.2 Consumer Behavior Analysis

2.1.2.1 Literature and Gap

Application of the scientific method to the investigation of human behavior, and psychology should be as free as possible from the various species of bias in order to yield reasonably reliable and valid results (Felthous, 2014). However, as mentioned in Section 1.2.2, behavioral analysis often rely on observational method (Bakeman & Quera, 2011; Fiske et al., 2011; Moutinho, 2000) and self –report by research subjects (Moorman & Podsakoff, 1992). However, the validity of these data is often at question. The significant reliance on self-reports has been identified as a major short-coming in organizational behavioral analysis with one of the major reason of the tendency for individuals to respond in socially desirable ways (Donaldson & Grant-Vallone, 2002; Moorman & Podsakoff, 1992). As for observational method, it is largely affected by the accuracy, validity and reliability of the measurement (Kahng, Ingvarsson, Quigg, Kimberly E. Seckinger, & Teichman, 2011). Researchers could also be a source of bias. It's been found that researcher too often find what they seek by statistically exaggerating findings (Bower, 2013).

Digital technology has brought advances in measurement and recording, but due to the nature of the data source and research method, the bias of responders and researchers are still difficult to control.

2.1.2.2 Contribution: Using Unbiased Data

The study described in this thesis will use a different type of data set. Using aggregate data not only reduces the costs and efforts of conducting survey, but also eliminates the potential self-report bias. It is truly “the voice of data”. The modeling approach follows probabilistic and statistical theories instead of personal judgment, therefore, the results will not be affected by researcher’s personal opinion. When interpreting and discussing the results, the research will refer to literature and empirical findings, but no more than non-behavioral analysis.

2.2 Practical Contributions

The results of this study provide practical recommendations to managers and employees that may help enhance the outcomes and achieve higher mutual values in the service industry.

2.2.1 Literature and Gap

2.2.1.1 I/O Model of Consumer Service Process

There are a great variety of sectors in the service industry, but they all share a similar operation and profiting structure. As Heskett et al (Heskett & Sasser, 2010) stated, the service profit chain posits, simply, that profit (in a for-profit organization) and growth (or other measures of success in for-profit or not-for-

profit organizations) results from customer loyalty generated by customer satisfaction, which is a function of value delivered to customers. Value for customers in turn results from employee loyalty and productivity, a function of employee satisfaction, which is directly related to the internal quality (or value) created for employees. Compared to the Primary and Secondary industries, this tertiary industry is more human-centric. Therefore, the definition, measurement and evaluation of values should not be considered without taking human factors into consideration.

This can be illustrated by the Input/Output (I/O) models used in the operations management process of the service and non-service industries. The figures below are from Sampson's summary (Sampson, 2010). Figure 2.1 represents the traditional paradigm about service, referring to it as a product delivered from the supplier to customers. The Unified Service Theory (UST), however, holds that service is a process wherein customers supply on or more input components for the production process of service. The participation of individual customers in the service process is the distinguishing feature of service industry.

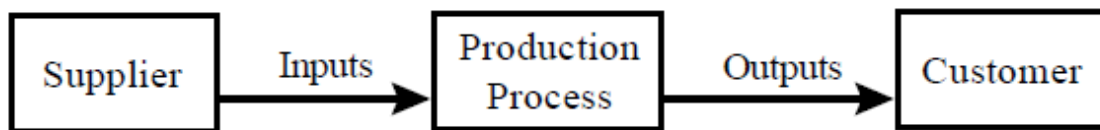


Figure 2.1 Traditional I/O Model

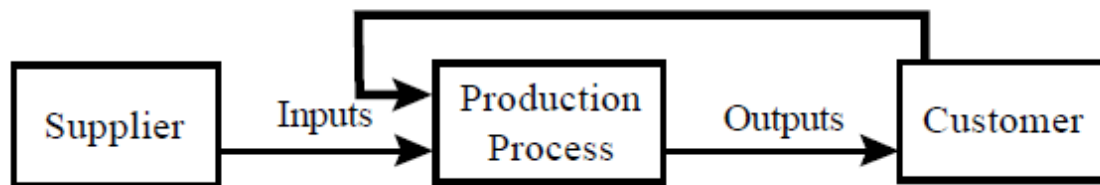


Figure 2.2 Service I/O Model

The abstract service I/O model can be expanded with details into a framework as shown in Figure 2.3. Consumers make decision and make consuming behavior with influence from multiple factors. On the service suppliers' side, the provision of service is also the result of multiple factors. Together, with the contribution of external factors, they form the service system with output values on both sides, monetary and non-monetary: customer satisfaction and loyalty, brand reputation, profits.

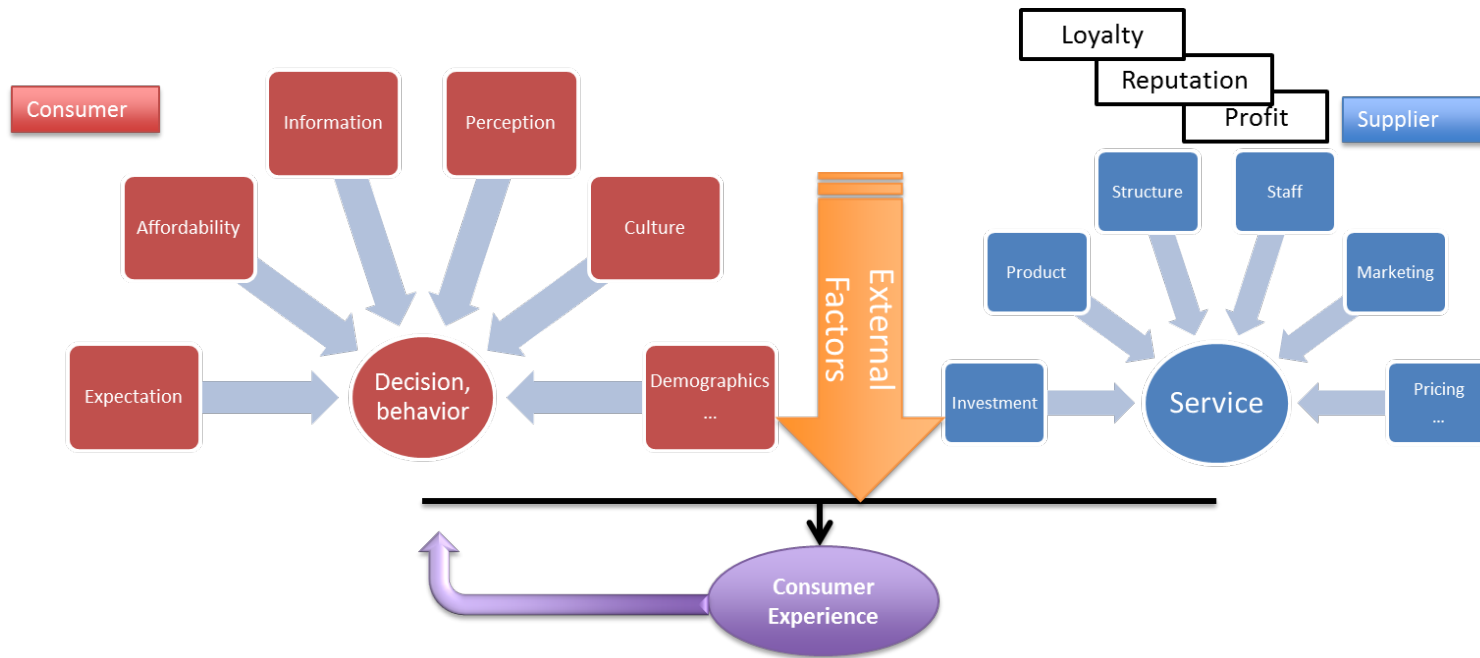


Figure 2.3 Consumer Service Industry Framework

2.2.1.2 Features of the Travel and Tourism Industry

Look closer at the case study area: the travel and tourism sector. It is one of the leading industries worldwide which involves many elements: history, culture, environment, transportation, infrastructure, economy, service, management, safety, policy making, etc. Tourism involves the greatest flows of goods, services, and people on the surface of the earth, and it is, therefore, the most visible expression of globalization, described by the movements of services and flows of information and capital (Reisinger, 2008). In the twentieth century mechanized mass transportation opened up exciting new experiences for people of all classes (Votolato, 2007). With the advances in transportation and digital technology today, consumers are exposed to many choices of available in the global travel and tourism market to suit their budget and needs, and thus have many decisions to make. Therefore, in the I/O model, factors that could influence customer decision making should all be considered as inputs into the service production process, and customer experience must be viewed as part of the outcome values. To gain an advantage against competitors all over the world, service suppliers need to correctly identify these input and output factors, understand their relationships, in order to control the controllable ones, and prepare for the uncontrollable ones.

Reisinger describe the tourist in the globalized travel and tourism market as a “new type of tourist” (Reisinger, 2008) who demands new products, variety, flexibility, and personalization. Their demands often come from their cultures.

For example, Figure 2.4 is based on the monthly tourist arrival data from Hawaii Tourism Authority (Tourism Research Division of Hawaii Tourism Authority, 2014). One would easily notice the rising trend from January to February, and the spike in February. Late January or early February is the time of Chinese New Year with an extended national holiday. During the winter, people like to take vacation in warm places. This results in an increase of tourist number to Hawaii. While it is difficult for a Hawaii service supplier who doesn't know about Chinese festivals to forecast this trend, it can be reflected by airline seat occupancy.

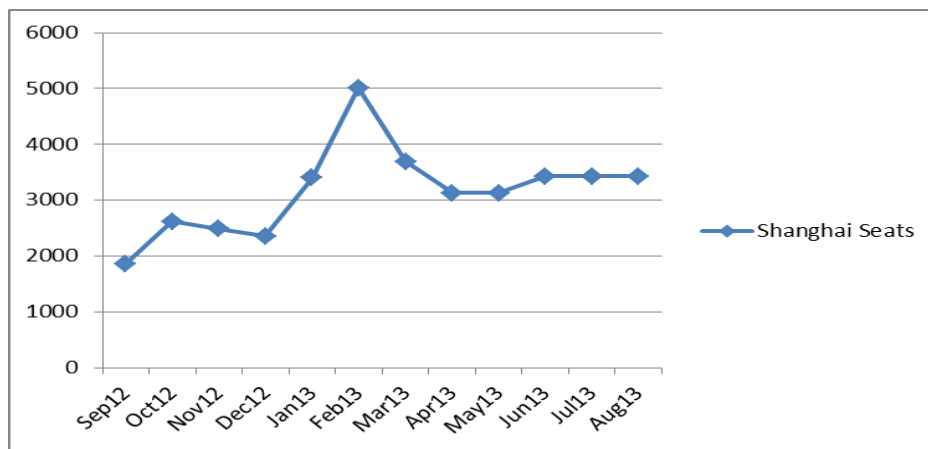


Figure 2.4 Airline seat occupancy of flights to Hawaii from Shanghai

Knowing this, not only the airline company is better prepared for the passenger volume, local service providers in Hawaii, such as hotels, restaurants, car rental companies, can also plan in advance to ensure that visitors' needs are met. On the opposite side, poor preparation due to lack of information could lead to problems like hotel being overbooked and understaffed, and result in customer values being undermined. We know from the service I/O model that this will harm

the mutual value production, and in turn reduce the outcomes on the supplier side.

Organizations need to know more about their final customers, but the reality is that they are typically widely separated from the consumers (Moutinho, 2000).

Destination service suppliers are usually local businesses. While the owner might be either local or global chain operations, the staff who work directly with the customers day-to-day are most likely hired locally. For them, to understand different cultures of other parts of the world is a big challenge. At the manager level, marketing strategies and operations need clear, result-oriented and reliable advice.

2.2.2 Contribution: Holistic Model, Applicable Recommendations

The example of Figure 2.4 uses a univariate analysis that considers the relationship between only two variables. It is an overly simplified representation. In reality, higher volume of Chinese visitors during January and February will cause overbooking of flights and might increase the airfare. This could make some people change their travel plan or even switch to another destination of the similar type, like the Maldives or Guam. Therefore, the relationship between airline occupancy and airfare should be included in the analysis.

The model developed in this research's case study is a multivariate analysis which takes into consideration the values on both customer and supplier sides. The service system will be treated as a truly systematic model, with multiple

layers, omni-directional relationships, and intermediate factors which are both the result and the cause of other factors. In this way, the model is a comprehensive representation of the real world situation.

Because it uses measurable and meaningful factors, the results will show direct influences on the values of significance. Based on historic data and the learning ability of Bayesian network, the model enables probabilistic projection for the future.

The model is scalable depending on the data availability and users' priority. The graphical presentation of the network helps destination service suppliers to easily identify the most important relationships, and then also allows them to take a closer look at the problems of most concern.

In summary, this study helps service suppliers to make informed decision, avoid costly mistakes, make marketing strategies, plan for resource allocation and investments to improve their profits while enhancing customer satisfaction and loyalty. From human resource point of view, it also provides an opportunity to guide and educate the employees which are of a great value in the information-rich service industry.

CHAPTER 3. METHODOLOGY

3.1 Multivariate Analysis

Marketing research is widely used in tourism management organizations and service suppliers. Some are beginning to incorporate marketing research into their marketing information systems (MIS) designed to provide managers with the relevant information needed to solve recurring problems and make decisions.

Moutinho (Moutinho, 2000) emphasized that the importance of MIS by distinguishing it from traditional marketing research:

1. It is oriented not only to solve problems, but also to prevent problems through control.
2. It operates as a true system rather than intermittent projects.
3. It uses projection techniques for acquiring future oriented data.

In the actual implementation, powerful techniques are required to achieve best results of the marketing research and MIS tools. Given the natural complexities of the service systems, any researcher who examines only two variable relationships and avoids multivariate analysis is ignoring powerful tools that can provide potentially very useful information (Moutinho, 2000). With the assistance of computerized data analysis technology, multivariate analysis has become an essential approach. However, to be considered as truly multivariate analysis, all

of the variables must be random variables that are interrelated in such ways that their different effects cannot meaningfully be interpreted separately (Moutinho, 2000).

Multivariate analysis includes dependence and independence methods (Moutinho, 2000; Rencher & Christensen, 2012). This study focuses on dependence methods as the objective is to explain and predict the relationships of a set of variables. Multiple regression analysis and canonical correlation analysis are among the most important dependence methods.

When the type of data and problem is simple and similar, analysis on multiple variables can be an extension of the basic data types and analysis (Rencher & Christensen, 2012). However, in many cases simple univariate or bivariate analysis techniques that only model the relationship and trend between 2 variables are insufficient. These methods fail to approach the problems in a systematic way because they see only single pairs of factors instead of the whole network. In real practice, the amount of sample groups and factors involved often makes it overwhelming for the attempt to carry out an analysis using insufficient tools. For example, in a system with 10 factors, it would take 45 rounds of univariate analysis to examine the relationships between each 2 variables. Expert knowledge may help narrow down the scope, but it also induces the risk of missing out important unknown information, which could lead to unaffordable mistakes. Moreover, it is very likely to have more than 2 variables interacting with

each other as a subgroup. Simplified assumption and reliance on individual's knowledge is risky, especially for a new problem.

In summary, a technique that is powerful enough to approach multiple factors efficiently, and model complex systems in a holistic way is required.

3.2 Machine Learning

As discussed earlier, the purpose of this study is to build a network model to represent the service I/O system, specifically, the relationships in the multi-variable system - not only between inputs and outputs, but between any two factors of interest. The course of discovering patterns in existing data to solve problems is known as data mining (Fayyad et al., 1996; Han et al., 2012a; Maimon & Rokach, 2010; Witten & Frank, 2005b). The computer technology has enabled automatic or semi-automatic data mining in large quantities of data based on memory capacities, instruction operations and algorithms (Sebe, Cohen, Garg, & Huang, 2005; Witten & Frank, 2005b). The goal of machine learning is to use computers to extract knowledge from experimental data for complex decision-making (Huang, Kecman, & Kopriva, 2006a).

3.2.1 Typology of Machine Learning

In terms of the typology of machine learning, researchers have used similar terminologies, represented by the 4 basic styles of learning: Witten and Frank (Witten & Frank, 2005b) suggested classification, association, clustering, and numeric prediction (a variant of classification learning). As for the general

learning algorithms to solve different types of problems, there are two well-recognized major types: unsupervised learning and supervised learning.

Unsupervised learning algorithms work with unlabeled data with the objective to discover structure in the data, while supervised learning models are trained with labeled data, i.e., a desired output, to speculate the output for an input that has not been observed. Figure 3.1 shows the 2 phases of supervised learning.

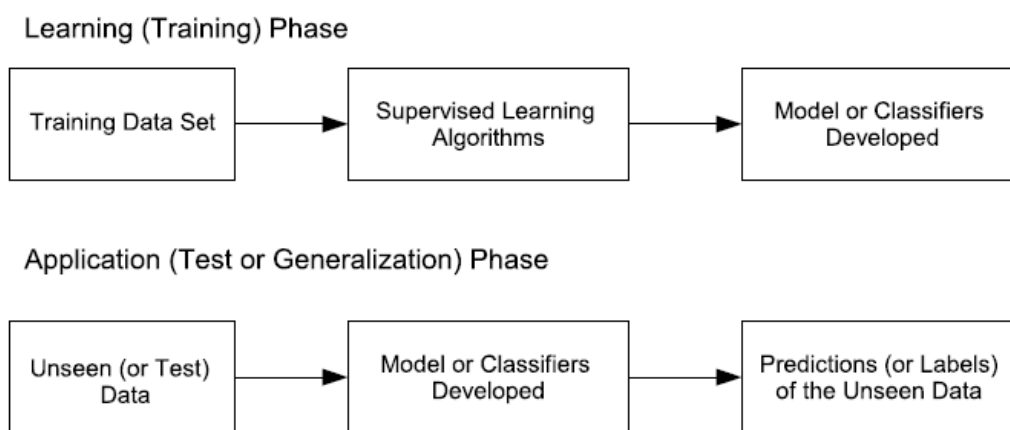


Figure 3.1 Supervised Learning Algorithms in Two Phases

In practice, unsupervised learning algorithms are usually used for clustering and association detection, while supervised learning algorithms are used for classification, regression, and prediction (Huang et al., 2006a; Karayiannis & Mi, 1997; Kasabov, 2001; Zhu & Goldberg, 2009). A combination of supervised and unsupervised learning techniques is known as semi-supervised learning (Huang et al., 2006a; Karayiannis & Mi, 1997; Kasabov, 2001; Sebe et al., 2005; Witten & Frank, 2005b; Zhu & Goldberg, 2009), which is popular in applications due to its ability to use readily available unlabeled data to improve supervised learning tasks when the labeled data is scarce or expensive, and its potential as a

quantitative tool to understand human category learning, where most of the input is self-evidently unlabeled (Huang et al., 2006a).

Considering the objectives of this study and the characteristics of the application area, the ideal method should be a hybrid of different types of learning algorithms at different stages: using unsupervised learning to obtain qualitative knowledge; then clustering and numeric prediction for quantitative knowledge.

3.2.2 Choice of Techniques and Methods

There are many schemes and techniques of machine learning in real world implementation, including linear modeling, decision trees, support vector machines, artificial neural networks, Bayesian networks, etc (Alpaydin, 2004; Witten & Frank, 2005a). Because this study attempts to approach a system without prior knowledge and to find out the structure based on existing data, the information needed to construct the consequences and hierarchies of a decision tree is not available, There is no evidence that the unknown relationships follow a linear regression function. Support vector machine method is powerful in classification and categorization (Huang, Kecman, & Kopriva, 2006b; Witten & Frank, 2005a), but not in association discovery and inference. After excluding these methods, the next section will compare Bayesian networks and artificial neural networks in detail to explain why Bayesian network was chosen as the research method in this study.

As shown in Figure 3.1, prediction is the final stage of supervised machine learning. The researcher also attempts to achieve the predictive inference of concerned factors through the system modeling. In other words, in addition to understand how one factor is related to another (qualitative), the research is also designed to find out how much influence the relationship has on the factors (quantitative). When introducing predictive approaches, Geisser (Geisser, 1993) divided them into 2 big categories: non-Bayesian and Bayesian, and recommended Bayesian predictive modeling as not only a substitute for parametric analysis, but also presents predictive analysis that have no real parametric analogues, which fits the situation of this research. This also supports the choice of Bayesian networks.

3.3 Theoretical Basis: Bayesian Networks

Bayesian networks is a type of probabilistic graphic model based on Bayes' theorem (also known as Bayes' law or Bayes' rule). This section provides an introduction of this modeling technique, and explains why it is appropriate for this study.

3.3.1 Bayes' Theorem

Named after the British mathematician Thomas Bayes who first developed this theorem in the 18th century, posthumously updated and published by his colleague Richard Price, and put into the modern formulation by French mathematician Pierre-Simon Laplace in 1784 (Rawlins, 2011), Bayes' theorem gives the posterior probability function for an event A which is conditioned by a

joint input event B under the assumption that we can express the joint likelihood density $P(A|B)$ as a product of the probability of A and the conditional probability of B given A, $P(B|A)$ as:

$$P(A|B) = \frac{P(B|A)P(A)}{P(B)}.$$

Equation 3.1 Bayes Theorem

In Bayes' Theorem, each probability above has a conventional name. $P(A)$ is called the prior probability, also known as “unconditional” or “marginal” probability. The term “prior” doesn't mean it happens earlier than B in the time sequence, but means that it doesn't take into account any information about B (Conrady & Jouffe, 2013b). $P(B)$ is the prior, or marginal probability of B. $P(A|B)$ is the conditional probability of A given B. It is also called the posterior probability because it is derived from or depends upon the value of B. $P(B|A)$ is the conditional probability of B given A.

3.3.2 Introduction of Bayesian networks

Bayesian classifiers are statistical classifiers which can predict the probabilities of belonging to a particular class. Studies comparing different algorithms have found a simple Bayesian classifier known as the naive Bayesian classifier to be comparable in performance with decision tree and selected neural network classifiers (Han, Kamber, & Pei, 2012b). Naive Bayesian classifiers assume that the effect of an attribute value on a given class is independent of the values of the other attributes. This assumption is called class-conditional independence. It

is made to simplify the computations involved and, in this sense, is considered “naive” (Han et al., 2012b).

A Bayesian network, also known as Bayesian belief network or Bayesian model, allow the representation of dependencies among attributes (or variables). In practice, this is more useful than the simplified assumption with naive Bayesian classifier.

The origins of Bayesian networks can be traced back as far as the early decades of the 20th century, when Sewell Wright developed path analysis to aid the study of genetic inheritance in crops (Sebastiani, Abad, & Ramoni, 2010). In the late 1970s, their development was motivated by the need to model the top-down (semantic) and bottom-up (perceptual) combination of evidence in reading (Conrady & Jouffe, 2013b). In the early 80s, Bayesian networks were introduced as a knowledge representation formalism to encode and use the information acquired from human experts in automated reasoning systems to perform diagnostic, predictive, and explanatory tasks (Sebastiani et al., 2010). Feature by their intuitive graphical representation, support for bi-directional inferences, and the theoretical basis of probabilistic foundation, Bayesian networks rapidly became a well-received choice when it comes to uncertain reasoning in artificial intelligence (AI) and the data mining and knowledge discovery community. This highly symbolic formalism, originally developed to be used and understood by humans, well-grounded on the sound foundations of statistics and probability

theory, is able to capture complex interaction mechanisms and to perform prediction and classification (Sebastiani et al., 2010).

As a graphic model, a Bayesian network is built up with two components: a directed acyclic graph, and a probability distribution which is often provided as a table. In the graph, the model consists of two important building blocks: nodes and arcs. All the variables are represented by nodes, whether they have categorical, continual or discrete values. Arcs indicate the directed probabilistic dependencies between two variables. If an arc is drawn from node A to B, then A is a “parent” or immediate predecessor of B (Han et al., 2012c). Arcs can be bi-directional.

3.3.2.1 Example of A Simple Bayesian Model

This section uses a simple example to illustrate the characteristics of Bayesian network. Please note that data in this example is only used for explanation purpose and doesn't represent any actual study.

Figure 3.2 is an adaption of the known fact in clinical research that cigarette smoking is the number one risk factor for lung cancer (Centers for Disease Control and Prevention, 2013). The arrow from “Smoker“ to “Lung Cancer” indicates the causal relationship that the chance of having lung cancer is influenced by whether or not the person is a smoker (among other factors which are not shown in the figure as a simplified example). While there is a causal relationship between the two variables, being a smoker does not definitely lead to

the conclusion that the person must have lung cancer. In a Bayesian network, each variable has a conditional probability distribution table showing the conditional probabilities of the variable, given the value of its parent (or combinations of its parents). The marginal and joint probabilistic distribution table of the parent node “Smoker” and descendant node “Lung Cancer” is shown in Table 3.1.



*Arrow indicates causal relationship: being a smoker could cause lung cancer

Figure 3.2 Example of a Two-Node Bayesian Network

Table 3.1 Marginal and Joint Probabilistic distribution table of Lung Cancer and Smoker

	Lung Cancer (LC)	No Lung Cancer (NLC)	Marginal Probability (Smoker)
Smoker (S)	0.15	0.25	0.4
Non-Smoker (NS)	0.05	0.55	0.6
Marginal Probability (Lung Cancer)	0.2	0.8	1

Given the information above, the conditional probability can be deducted. For example:

$$P(LC|S) = P(LC \cap S) / P(S) = 0.375$$

Therefore, the complete conditional probability distribution is:

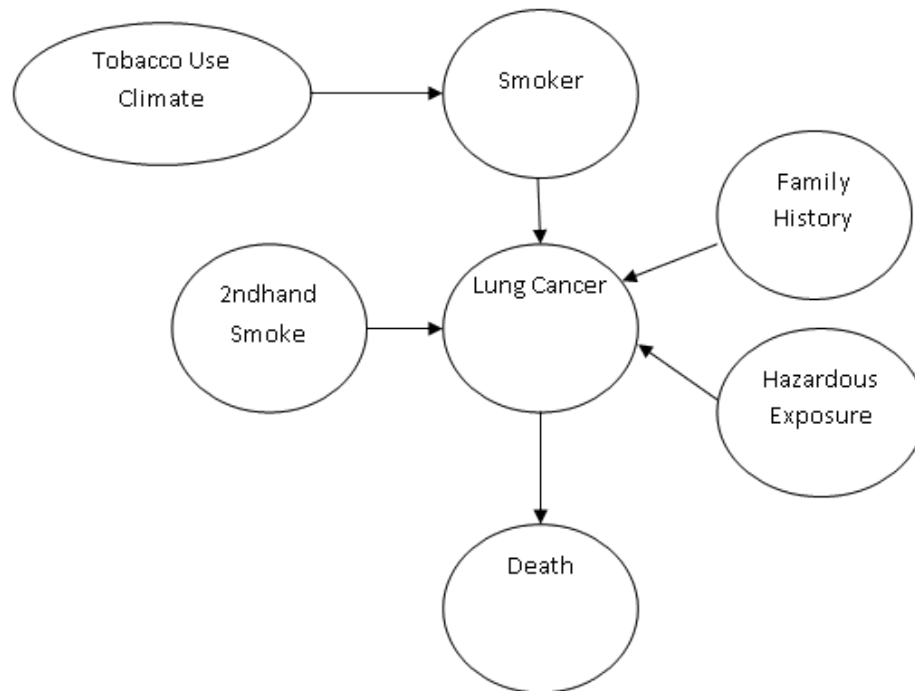
$$P(LC|S) = 0.375 \quad P(NLC|S) = 0.625$$

$$P(LC|NS) = 0.083 \quad P(NLC|NS) = 0.917$$

Similarly, given the data of conditional and marginal probability distribution, joint probabilities can be deducted reversely. In real cases, data could be available in either condition.

With more information, this simple example can be extended to a more complete model. As proven in clinical research (Centers for Disease Control and Prevention, 2013), smoke from other people's cigarettes, pipes, or cigars (secondhand smoke) also causes lung cancer, family history and exposure at home and work to hazardous gas or substances can increase the chance of having lung cancer, and tobacco use climate enhances the chance of a person being a smoker. If there are smokers in their households and to have spouses, friends and family members who smoke, they are more likely to be smokers (Center for Tobacco Research and Intervention, 2002). Obviously, a tobacco-friendly environment also increases the chance of suffering from secondhand smoke. Finally, lung cancer can lead to death. Based on these information, an extended model is shown in Figure 3.3. Similar to Figure 3.2, the arrows indicate causal relationships (e.g. being exposed to secondhand smoke could lead to lung cancer).

.



*Arrow indicates causal relationships

Figure 3.3 An extended example of Bayesian network

In this case, the conditional probabilistic distribution of lung cancer is based on the combination of 4 parents. The conditional probability distribution of having lung cancer follows the multiplicative rule:

$$P(B_i | A) = \frac{P(A | B_i)P(B_i)}{\sum_{i=1}^n P(A | B_i)P(B_i)}$$

Equation 3.2 Bayes Theorem with Multiplicative Rule

The denominator is the total marginal probability of event A which, in this example, is having lung cancer or not. B_i stands for each of the attributes that affect A. They must be conditionally independent of each other to satisfy the total probability law.

3.3.2.2 Features of Bayesian Network

Although simple, this example demonstrated several important features of Bayesian networks that made it an appropriate method for this study.

1. Graphical representation. Bayesian networks use graphical network model to present the causal relationships in a multivariate system. It provides an intuitive perception at a glance, which is especially useful as the sizes and complexities of the data set increases. Computer-generated graph has long been recognized as a useful tool for communicating information efficiently and effectively (Lohrding, Johnson, & Whiteman, 1978; Woo, 2012). Data table is an organized form of complete and accurate original data but doesn't tell what they mean. In probabilistic and relationship analysis, commonly used conventional data charts like histograms, distribution function plots, trend line plots, scattered or cluster plots (Chandoo, 2010; Lohrding et al., 1978) can only display data samples or attributes in a 2 to 3-dimensional way. For large volume, multivariate analysis, network models are the most vivid reflection of the real problems. Through software functions, Bayesian network models can even visualize the direction and strength of the relationships, making it much easier to identify the most noteworthy issues and enhancing the efficiency of system analysis.
2. Omnidirectional relationships. In a Bayesian network, there is no unitary direction for the relationships. Each node can have one or more parents

and also be a parent itself. Compared to an artificial neural network (ANN), another widely used technique in modern data mining, a Bayesian network reflects complex systems like the case in this study more accurately.

ANNs are inspired by the nervous systems of animals, especially the brain. Roughly speaking, a neural network is a set of connected input/output units in which each connection has a weight associated with it (Han et al., 2012c). They are also graphical models with multiple layers of perceptrons from input to output. An example of a feed-forward ANN is shown in Figure 3.4 (Han et al., 2012c). Comparison with Figure 3.3 shows that Bayesian network has no certain layer where a group of factors receive input and generate output at the same level and in the same direction. Instead, the “arcs” in a Bayesian network is omnidirectional.

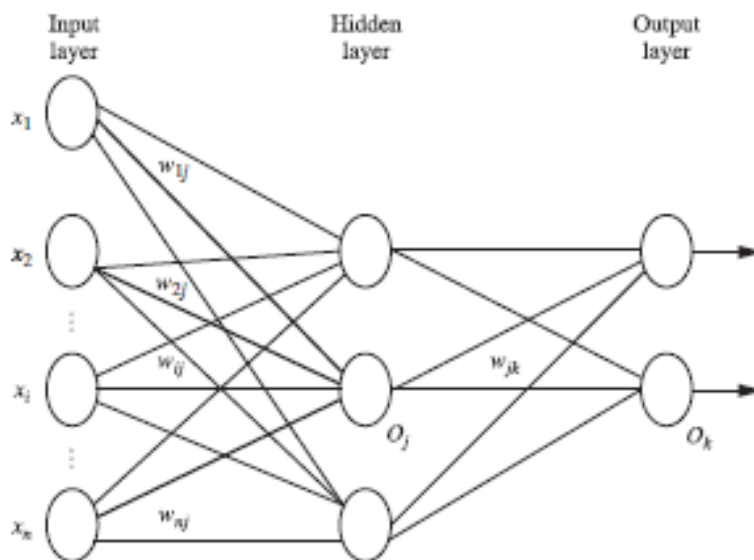


Figure 3.4 An example of a multilayer feed-forward neural network

3. Learning ability. According to the probability interpretation of Bayes' Theorem, the degree of belief in a certain proposition is related to the knowledge of prior evidence, which is to learn from the data. Even a small set of observations can be used to train the network in order to find out the optimal solution.

One of the most important features of ANN is also learning ability. Neural networks involve long training times and are therefore more suitable for applications where this is feasible (Han et al., 2012c). They require a number of parameters that are typically best determined empirically. Techniques like neural networks are designed solely to achieve accuracy. However, as their classifiers are represented using large assemblages of real valued parameters, they are also difficult to understand and are referred to as black-box models (Maimon & Rokach, 2010).

In comparison, Bayesian network provides an elegant formalism for representing and reasoning about uncertainty. It specifies a joint probability distribution over a finite set of random variables and consists of both qualitative and quantitative components (Kersting & De Raedt, 2001).

3.3.2.3 Bayesian Network and Artificial Neural Network

This section provides an extension and summary of the research method selection. In Section 3.3.2.2, ANN was introduced as a similar graphical network modeling technique to be compared with Bayesian network. They both can be

used to analyze and extract information from large and complex data sets with a number of variables to extract explicit information which can be used for diagnosis, forecasting, optimization and other issues in a wide range of industries. Through a thorough comparison and literature study (Conrady & Jouffe, 2013b; Han et al., 2012c; Stassopoulou & Petrou, 1998), the advantages of Bayesian Networks are summarized below:

1. They allow bidirectional flow of information between causes and effects. In a Bayesian network, an arc from cause to effect indicates deduction, prediction or simulation, while an arc in the opposite direction enables diagnosis and reasoning. ANN flows are from input to output only.
2. They allow input data to be inserted at any node. In ANN, there is only one input layer where data can be entered.
3. Since the model deals with dependencies among all variables, they can cope with incomplete and uncertain data. ANN relies more on the accuracy of input data.
4. They can cope with uncertain rules of reasoning, strengthening the power of diagnosis and prediction. ANN is based on empirically predetermined parameter structure.
5. All the nodes and arcs are displayed and transparent to the analyst. In ANN, there are hidden layers and hidden units. Although the accuracy of the results is not affected, it's less flexible and more difficult to transfer, modify and understand.

The similarities and differences between the two methodologies are summarized in Table 3.2.

Table 3.2 Similarities and Differences between Bayesian Network and Artificial Neural Network

	Bayesian Network	Artificial Neural Network
Similarities	Reflect the relationship and dependencies among multiple variables	
	Data mining, pattern recognition	
	Learning ability	
	Graphic presentation	
Differences	Based on Bayes' Theorem	Inspired by brain nervous system
	Acyclic graphs	Can be acyclic or cyclic
	Allows bidirectional causal relationship	From input to output only
	Input data can be inserted at any node	Layered structure, input data at the initial input data only
	Can cope with incomplete/uncertain data and uncertain rules of reasoning	Relies more on training data set from observations

The concerned field in this study is the consumer service system, specifically, the tourism market. Previous chapters have discussed that in a network view of the service system, there are many intermediate factors which are both input and output, and relationships could exist any two factors in either direction, empirical assumptions are often limited or risky to rely on, and data records may be inconsistent or incomplete. Targeted application defines the most desirable features of the research method. Based on these considerations, Bayesian network has been identified as the most appropriate method for this study.

For any market, especially in the service sector, consumer satisfaction is a key measurement of the quality of product and services. Past studies have shown that destination and tourist satisfaction have a significant impact on destination loyalty (Rajesh, 2013). Impaired destination loyalty does not only reduce the chance for revisiting, but also leads to negative word-of-mouth advertising which, in the age of social network, microblog and social network, will be magnified and influence more potential visitors. Therefore, understand and forecast tourists' behavior is very important.

These new tourists have multiple demands, often borrowed from other cultures; they are more dependent on information technology and self-service; they have become more individualistic and require more customized and highly developed products (Castillo-Manzano, López-Valpuesta, & Gonzalez-Laxe, 2013). Such changes in consumer behavior have also brought changes to destination marketing and called for the development of more targeted and customized products. Complexities of globalization call for understanding and accommodating different worldviews, variations in employers' business practices, and differences in national cultures of employees and consumers (Reisinger, 2008). Global service suppliers must develop high levels of intercultural communication and competencies and make appropriate adjustments to their business practices to suit particular customer needs.

3.3.3 Information Theory and Statistics Theory

An important theoretical support part of the research approach in this study is based on Information Theory and statistics, especially the mutual relationship (MI) and Pearson Correlation Coefficient (PCC).

3.3.3.1 Mutual Information

Information is an umbrella term, too broad for a single definition. In information theory, information is simply the outcome of a selection among a finite number of possibilities measured by entropy (Cover & Thomas, 2006; Feixas, Bardera, Rigau, & Xu, 2014). Mutual Information, a special form of relative entropy, is a representation of the information shared between a pair of nodes. It is used to measure the dependence between two random events - how much can be known about one node given that the knowledge of the other. The formal mathematical definition of MI is shown in Equation 3.3 (Conrady & Jouffe, 2013c; Cover & Thomas, 2006; Feixas et al., 2014).

Formal Definition of Mutual Information

$$I(X;Y) = \sum_{y \in Y} \sum_{x \in X} p(x,y) \log \left(\frac{p(x,y)}{p(x)p(y)} \right)$$

Equation 3.3 Formal Definition of Mutual Information

3.3.3.2 Pearson Correlation Coefficient

Prof. Karl Pearson first developed a coefficient of correlation in 1895 in an inheritance study in the case of two parents (Pearson, 1895). In statistics,

Pearson Correlation Coefficient is a measure of the linear correlation between two variables, usually denoted by r when applied in a population (Wikipedia, 2014b). In a given population, the coefficient is calculated, where cov is the covariance, σ_X is the standard deviation of X , μ_X is the mean of X , and E is the expectation. The mathematical definition is shown in Equation 3.4 (Wikipedia, 2014b).

$$\rho_{X,Y} = \frac{\text{cov}(X, Y)}{\sigma_X \sigma_Y} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sigma_X \sigma_Y}$$

Equation 3.4 Formula of Pearson's Correlation Coefficient

Pearson's distance is a distance metric for two variables X and Y , defined from their correlation coefficient as: $d_{X,Y} = 1 - \rho_{X,Y}$.

CHAPTER 4. STUDY DESIGN

4.1 Design of Analysis

This study conducted an analysis of an existing data set to develop and verify the systematic analysis approach. This section will introduce the background of the study, the data set, and the preparation of data for it to be analyzed.

4.1.1 Hawaii Tourism Industry

The State of Hawaii, with its six islands, is one of the most desirable tourist destinations in the world. Being the southernmost state of the United States, it's not geographically located in the North America Continent, but in the North Pacific Ocean. This unique combination of territorial property and geographic position has made Hawaii accessible to a large population of tourists from the North America, the Asia-Pacific region, and Europe. Its charming scenery, pleasant tropical climate all year round, rich natural resources, abundant beaches, and the historic and cultural heritage have attracted visitors with a variety of purposes including vacation or family trip, wedding or honeymoon, seaside activities, and biology and geological research. Furthermore, Hawaii has also become a popular choice of national and international events including meetings and conventions, film festivals, golf championships and more.

Tourism plays an important role in the state's economy. From 1974 to 2013, visitor expenditure has constantly been on top of the export industries in terms of expenditure (Department of Economic, Development & Tourism, 2014), higher than the total of the expenditure of the 3 industries that follow it. In 2013, visitor expenditure reached 14,520.5 million dollars (Department of Business, Economic Development & Tourism, 2013). In the 2012 Annual Report of the State of Hawaii, Hawaii saw steady economic growth led by key areas such as tourism and construction, and sectors like Food Services and Drinking Places, Accommodation, Trade and Transportation, Warehousing and Utilities are the leading contribution to job gain compared to same period of 2011 (Department of Business, Economic Development & Tourism, 2012).

4.1.1.1 Slow-Down and Decline

However, after a strong and sustained growth for more than 30 years, Hawaii's tourism industry struggled during the 1990s (Hibbard & Salbosa, 2006; Mak, 2008; State of Hawaii & Hawaii Tourism Authority, 2004). As shown in Figure 4.1 and Figure 4.2, although the overall visitor numbers still kept an uprising trend, year-to-year declines appeared since 1990, and the average annual increase rate started to slow down. A more obvious down trend is observed in visitors expenditures starting from 1995. These 2 figures originally appeared in Mak's book (Mak, 2008).

There are a range of reasons causing this trend (Hibbard & Salbosa, 2006; Mak, 2008; State of Hawaii & Hawaii Tourism Authority, 2004). Externally, the industry appeared to have become more susceptible to negative domestic and global events such as the prolonged economic recession in California, the first Gulf War, the economic bubble collapse in Japan, the Asian financial crisis and hazardous climate attacks like Hurricane Iniki (Mak, 2008). Entering the 21st century, following the dramatic shock of 9/11, a rise in global terrorist attacks and military outbreaks in Afghanistan and Iraq, the SARS epidemic, and the global financial crisis in the late 2010s all reduced people's desire and abilities to travel. While the influences of uncontrollable factors were acknowledged, some researcher and local observers also held the view that there is a gap in Hawaii's tourism management strategy. While the globalization of tourism market brought more destination choices to potential visitors, Hawaii was transitioning into a "mature market" with increasingly more repeat visitors. At the same time, hotels, resorts and facilities in the major tourist destinations needed renovation and redesign (Hibbard & Salbosa, 2006). Changing consumer preferences, shakeups in the airline industry, and technological advances have also recently contributed to revolutionary changes in the industry (State of Hawaii & Hawaii Tourism Authority, 2004). At the same time, people began to realize the impact of the tourism industry to the island's natural and cultural resources.

In 2004, the Hawaii Tourism Authority (HTA) adopted the Hawaii Tourism Strategic Plan, 2005-2015 (State of Hawaii & Hawaii Tourism Authority, 2004). It

is a more comprehensive and inclusive plan that addressed the needs and identified the responsibilities of all Hawaii's visitor industry stakeholders. The strategic Plan set a collective vision to move towards a sustainable and responsible tourism industry for the State, described as:

By 2015, tourism in Hawaii will:

- honor Hawaii's people and heritage;
- value and perpetuate Hawaii's natural and cultural resources;
- engender mutual respect among all stakeholders;
- support a vital and sustainable economy; and
- provide a unique, memorable and enriching visitor experience.

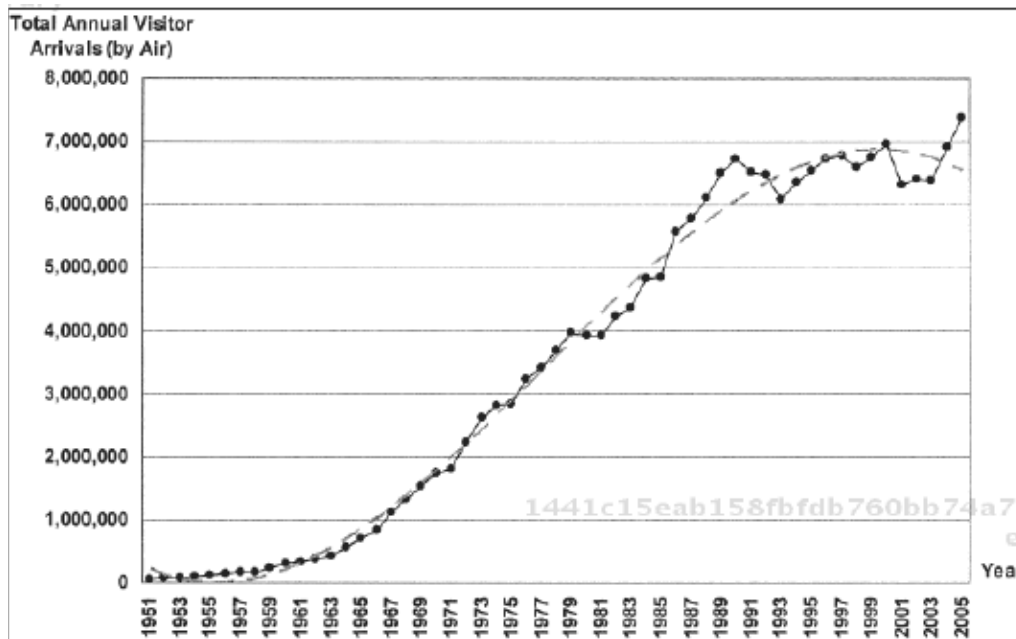


Figure 4.1 Hawaii Tourist Arrivals by Air 1951-2005

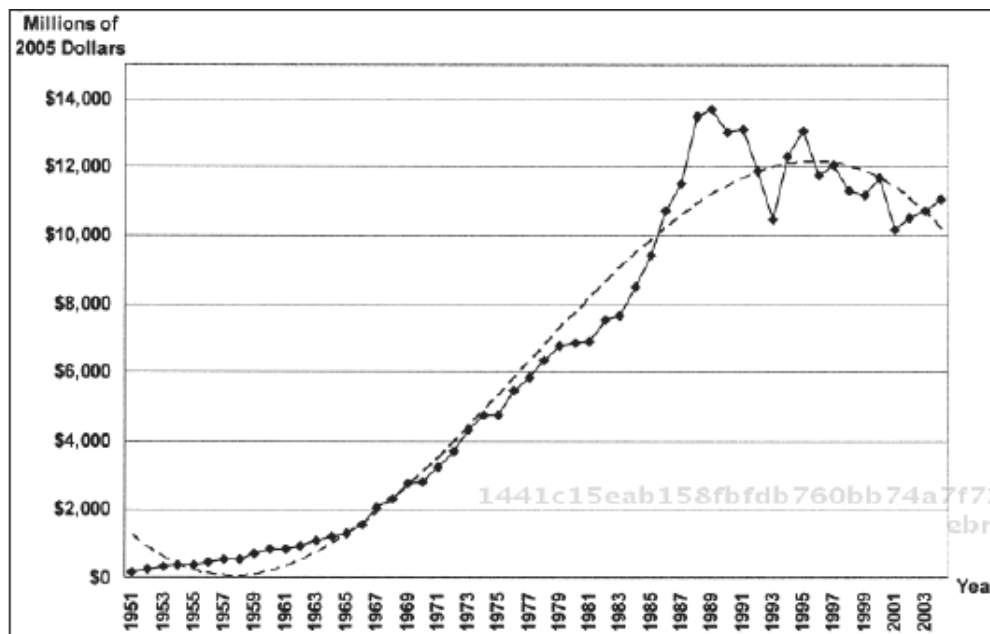


Figure 4.2 Hawaii Visitor Expenditures 1951-2005

4.1.2 Data

This study is indebted to HTA for providing the source data and publishing on its website (<http://www.hawaii tourism authority.org/research>) with public access, and giving permission for the use of these data in research. HTA's Tourism Research Division (TRD) develops statistical and analytical information and conducts special research on Hawaii's visitor industry that helps aid state marketing and product development efforts, industry planning and tourism policy-making (Tourism Research Division, 2014).

The Visitor Highlights section provides monthly visitor statistics reports highlighting the primary visitor characteristics, expenditure and other information for visitors arriving Hawaii from the four major marketing areas (MMA): U.S. West, U.S. East, Japan and Canada. The reports follow similar formats and summarize

the market performance in visitor arrivals, length of stay by days (visitor days), and expenditures (for example, Table 1 in the 2012 Annual Visitor Research Report (Tourism Research Division of Hawaii Tourism Authority, 2014)), As these parameters are closely related to the profits of the overall Hawaii tourism market, they are set as the measurable outcome values of the destination management organization (DMO).

Other parameters included in the reports are visitor characteristics (e.g. MMA, travel purpose, repeat or first-time visitor) and their consumer behaviors (e.g. accommodation choice, travel with group or not, take a package trip or not). As the annual report is summed up from each month's statistics, this study took the common parameters included in each monthly visitor highlight release. These 14 factors are:

- 3 Outcomes factors: visitor arrivals, average length of stay per person, expenditure per day per person. These are identified by HTA in annual report as performance measurers.
- 11 predictors: MMA, month, number of visitors staying at hotel / at Bed & Breakfast (B&B) / with friends or relatives, purpose of travel being for pleasure / for meeting or conference / for visiting family or friends, percentage of repeat visitors, number of visitors who traveled with a group, number of visitors who traveled on a packaged trip. In modeling (prediction modeling), predictors are referred to variable used as input or

causes. In this text, they are the 11 factors from the consumers' side or nature (Month), in contrast to the 3 outcome factors..

For each month, the Visitor Highlight data was obtained in an Excel spreadsheet. Data of each visitor were aggregated by MMA. An example of the raw data for the 3 outcome variables is shown in Table 4.1.

Table 4.1 Example of Monthly Visitor Highlight Raw Data

Year	Month	MMA	Arrivals	Average Length of Stay (days)	Per Person Per Day Spending (\$)
2013	December	Air_US_W	263919	10.31	155.2
2013	December	Air_US_E	142212	11.53	197.8
2013	December	Air_JP	138190	5.83	283.4
2013	December	Air_CA	67535	13.78	150.5

This study used monthly data from January, 2011 to December, 2013.

4.1.2.1 Data Preprocessing for Analysis

In addition to the original source data, some data were preprocessed for better analysis results. In the Visitor Highlights reports, most of the data were in absolute values, To eliminate the multiplicative and confounding effect, they were turned into percentage values. For example, U.S. West constantly has the highest visitor arrivals. When considering visitors from which MMA would be more likely to stay at hotel during their visit in Hawaii, U.S. West might have a higher absolute value of visitors staying at hotel, but this could be due to the larger sample. Therefore, these absolute values were divided by the total number of visitors in the same month to form a fair dataset for likelihood analysis. An example of this preprocessing is shown in Table 4.2.

Table 4.2 Example of Converting Absolute Values to Percentage

Year	Month	MMA	Arrivals	Number of Visitors	
				Staying at Hotel	Hotel%
2013	December	Air_US_W	263919	123368	46.74
2013	December	Air_US_E	142212	81250	57.13
2013	December	Air_JP	138190	121638	88.02
2013	December	Air_CA	67535	26680	39.51

For a clearer view in the model, the names of the variables are shortened into abbreviations. Table 4.3 is a complete list of the variable names and definitions.

Table 4.3 Complete List of Variables and Definition

Type	Name	Definition	Unit
Outcome	Arrivals	Number of visitors arriving in Hawaii	person
	Avg_Stay	Average length of stay by days	day
	Exp_pp/D	Per person per day spending by USD	USD
Predictor	MMA	Major Market Area, the original country/region visitors came from by air	
	Month	The month when the data were collected	
	Stay_Hotel%	Percentage of visitors who plan to stay at hotel during their stay in Hawaii (including hotel only and hotel + other accommodations)	
	Stay_B&B%	Percentage of visitors who plan to stay at Bed & Breakfast during their stay in Hawaii	
	Stay_F&R%	Percentage of visitors who plan to stay with Friends/Relatives during their stay in Hawaii	
	POT_Pls%	Percentage of visitors whose Purpose of Travel was Pleasure, including Pleasure/Vacation, Wedding and Honeymoon.	
	POT_Mtg%	Percentage of visitors whose Purpose of Travel was Corporate Meeting, Convention or Incentive	
	POT_Vst%	Percentage of visitors whose Purpose of Travel was to Visit Friends or Relatives.	
	Rep%	Percentage of Repeaters whose recorded visits were not their first trips to Hawaii.	
	Style_Grp%	Percentage of visitors who traveled with a group.	
	Style_Pkg%	Percentage of visitors who traveled on a purchased package trip.	

4.2 Modeling Tool: BayesiaLab

Since its recent widespread in scientific research and industries since 1970s, Bayesian networks have seen frequent uses in real world applications, such as diagnosis, forecasting, automated vision, sensor fusion, and manufacturing control (Heckerman, Mamdani, & Wellman, 1995). Most recent innovative

applications of Bayesian networks include bioinformatics (Husmeier, Dybowski, & Roberts, 2005), computational intelligence (Holmes & Jain, 2008), brain injury detection (Herskovits & Gerring, 2003), ecology and natural resource management (McCann, Marcot, & Ellis, 2006), dependability, risk analysis and maintenance areas (Weber, Medina-Oliva, Simon, & Lung, 2012), among others. The development of personal computer provided small, powerful devices on which modeling tools can run, and the advance of graphical user interface (GUI) stimulated the emergence of various software applications. Most of them support graphical modeling, pattern mining, learning and simulation. Some of the most well-known applications include AgenaRisk (<http://www.agenarisk.com>), BayesiaLab (<http://www.bayesia.com/en/products/bayesialab.php>), Bayes Server (<http://www.bayesserver.com>), Netica (<http://www.norsys.com/netica.html>), PrecisionTree (<http://www.palisade.com/precisiontree/>), and many more. Indeed, all these are very helpful tools with Bayes' rule embedded. While deciding which tool to use, several factors were taken into consideration: the capability to deal with multi-factor system and conduct predictive inference, the transparency of the algorithms, the cost and the availability of a free version for evaluation, the availability of tutorial material and examples, the easiness of data import from external files, the form of result presentation, and the user-friendliness. The author specifically studies AgenaRisk and BayesiaLab. For AgenaRisk, Fenton's book (Fenton & Neil, 2012) provides a good knowledge in application examples, but not so much in its algorithms. It has a lite version for free download, but with reduced functions in analysis and data visualization.

Most importantly, even Fenton himself stated that when a node has more than 3 parents, the calculation in AgenaRisk becomes very inefficient (Fenton, 2013). With BayesiaLab the calculation inefficiency for nodes with more than 3 parents was not observed, and the software supplier provided an online library with extensive information about the algorithms and interpretation of the software. It provides a free trial version as well, with a limitation of nodes quantity in a model. But other than that, the trial version supports full analysis features. The supplier also offers an elastic pricing purchase option, which allowed running the complete version at a much more affordable cost. During this study, BayesiaLab was chosen after an in-depth study through hands-on experience. It proved to be a dependable and comprehensive tool.

4.2.1 Introduction of BayesiaLab and Features

BayesiaLab is the modeling software developed and supported by Bayesia (<http://www.bayesia.com/en/index.php>), a designer of decision aid software packages, world leader in Bayesian networks for data mining (Bayesia, n.d.-a). It is a Bayesian network publishing and automatic learning program which represents expert knowledge and allows one to find it among a mass of data. BayesiaLab provides a complete laboratory for handling Bayesian networks to develop, communicate with and use readable illustrated decisional models that are strictly faithful to reality (Bayesia, n.d.-b). It has outstanding features and advantages that are desirable for this study, summarized below (Bayesia, n.d.-c):

1. Network modeling

- Highly intuitive graphic development of networks
- Easy data import/export in main formats in the market

2. Learning/data mining

- Powerful filter to identify unused values, discretize continuous variables, and incorporate discrete modalities
- Very wide range of learning algorithms

3. User interface

- Visually analyzing that presents models in a highly readable way
- Doesn't not require a statistics expert to use it

4. The power of Bayesian networks

- Take advantage of the Bayesian power of inference for scenario simulation and subject classification
-

4.2.1.1 Algorithm

The Information described in this section mainly comes from BayesiaLab's online library regarding its Score-Based Learning Algorithm (Bayesia, 2014).

BayesiaLab uses a proprietary score-based learning algorithms in modeling and visualization. It utilizes Minimum Description Length (MDL score) to measure the quality of candidate networks with respect to the available data. Derived from Information Theory, this score allows to automatically take into account the data likelihood with respect to the network and the structural complexity of the network.

MDL score is a two-component score traditionally used in the Artificial Intelligence community for estimating the number of bits required to represent a model and the data given this model. For structural learning of Bayesian networks, the model is the Bayesian network (graph plus probability tables), whereas the number of bits for representing the data given the Bayesian network is inversely proportional to the probability of the observations returned by the model. These are represented by Equation 4.1 to Equation 4.6.

$$\text{MDL}(D, B) = \alpha \text{DL}(B) + \text{DL}(D | B)$$

Equation 4.1 Expression of MDL

$$\text{DL}(B) = \text{DL}(G) + \text{DL}(P | G)$$

Equation 4.2 Expression of DL(B)

$$\text{DL}(D | B) = \sum_{j=1}^N \text{DL}(e_j | B) = \sum_{j=1}^N \log_2 \left(\frac{1}{P_B(e_j)} \right) = - \sum_{j=1}^N \log_2 \left(\prod_{i=1}^n P_B(x_{ij} | \pi_{ij}) \right)$$

Equation 4.3 Expression of DL(D|B)

$$\text{DL}(G) = \sum_i^n (\log_2(n) + \log_2(\|\pi_i\|))$$

Equation 4.4 Expression of DL(G)

$$\text{DL}(P | G) = \sum_i^n \left(\prod_j^{\|\pi_i\|} \text{val}(\pi_i^j) \times (\text{val}(X_i) - 1) \times \text{DL}(p) \right)$$

Equation 4.5 Expression of DL(P|G)

$$DL(p) = \frac{\log_2(N)}{2}$$

Equation 4.6 Classical Heuristic Expression of DL(p)

In these expressions:

- **MDL(D,B)**: the number of bits to represent the model,
- **DL(B)**: the number of bits to represent the Bayesian network **B** (graph and probabilities),
- **DL(D|B)**: the number of bits to represent the dataset **D** given the Bayesian network **B**,
- **α** : the BayesiaLab Structural Coefficient (the default value is 1), a parameter that allows changing the weight of the MDL structural part,
- **G** refers to the Graphical structure, and **P** to the set of Probability tables,
- **n** is the number of random variables (nodes) **X₁, ..., X_n**, **N** is the size of the dataset
- **π_i** is the set of the random variables that are parents of **i** in the graph **G**,
- **|| π_i ||** is the number of parents of random variable,
- **val(X)** represents the number of states of random variable **X**,
- **p** is the probability recorded in the cell,
- **e_j** is the n-dimensional observation described in row **j**, and
- **P_B(e_j)** is the joint probability of this observation returned by the Bayesian network **B**.

4.2.1.2 Limitations with Trial Version

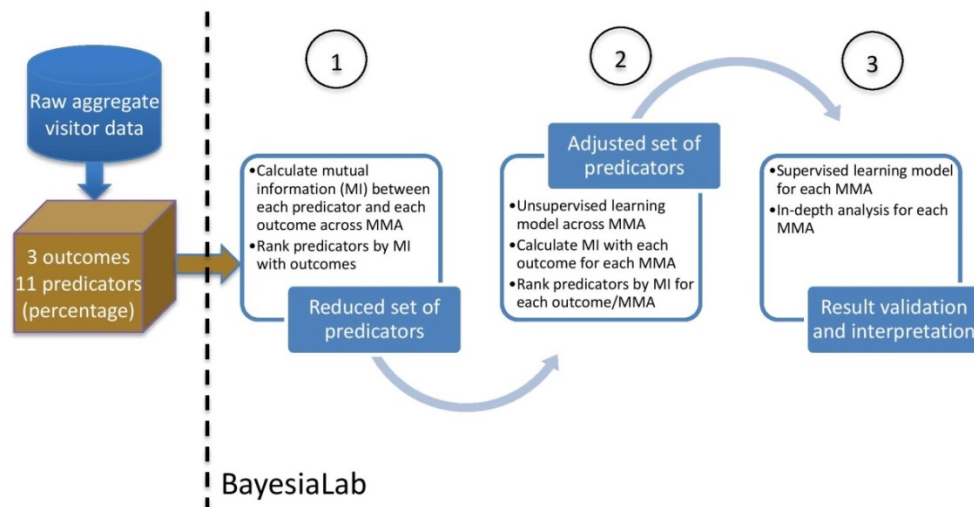
BayesiaLab is a commercial software charging license purchase fees. A single user 1-year standard edition costs €3,000. This was beyond affordability of the researcher. A trial version was used for this analysis. However, the trial version has two major limitations:

1. The model can't be saved.
2. It only allows maximum 10 nodes (variables) in a model.

The first limitation didn't cause much trouble. As all the import data files were saved separately, it didn't take long to recreate the model. But the second limitation forced the researcher to reduce the number of variables from the original 14-factor data set. The reasoning and verification of the variable selection process will be discussed in CHAPTER 5.

CHAPTER 5. RESEARCH APPROACH AND RESULTS

The process of analysis is also a process of exploring the developing an innovative approach. This procedure can be summarized in 3 steps, illustrated in Figure 5.1. This section will introduce each step in detail, and present the results with each step. The software features utilized will be explained along the way as well.



5.1 Step One: Initial Predictor Ranking and Screening

When studying a system with multiple factors and relationships, the analyst usually needs to clear out the factors that are insignificant or irrelative to the problem of interest.

5.1.1 Ranking Criteria and Method

The complete version of BayesiaLab can conduct unsupervised learning to help prioritize the factors that are most information-rich. But with the limited version, it's impossible to analyze all 14 factors at one time. However, knowing that for the DMO, the outcomes (higher visitor arrivals, longer period of stay, and more spending) are the values of most interest, it is reasonable to prioritize predictors that are most related to the outcome nodes. According to its definition (Section 3.3.3.1), mutual information only depends on the 2 nodes, regardless of the number of total nodes in the model. Therefore, MI with each of the 3 outcomes nodes was used as the index for initial predictor ranking and screening.

There are two types of data in this data set: MMA and Month are discrete data, and the rest are continuous. For discrete data, discretization is needed to calculate MI. This can be calculated manually or using computerized tools. BayesiaLab also has this feature. During data import (from database or text file in .csv or .txt format), there is an option to choose the discretizing type and intervals. K-Means was chosen in this analysis with 4 intervals.

5.1.1.1 K-Means Clustering

Originally used by James McQueen in 1967 (MacQueen, 1967), K-Means clustering is a widely used method in data mining to partition n observations into k clusters. It is a simple algorithm aiming to partition the n observations so as to minimize the within-cluster sum of squares (WCSS) (Department of Electronics

Information and Bioengineering, n.d.; Hartigan & Wong, 1979; Wikipedia, 2014a). It does so through an iterative optimization procedure to calculate the cluster prototype matrix of the partition until there is no change for each of the k cluster (Xu & Wunsch, 2008). The initial partition may be based on prior knowledge or set randomly, and the clustering in the next iterations follow the nearest-neighbor rule (MacQueen, 1967; Xu & Wunsch, 2008). In mathematical description, for a set of n observations, K- means algorithm aims to find the value as indicated in Equation 5.1, where S is the partitioned sets of k sub-samples: $S = \{S_1, S_2, \dots, S_k\}$, and μ is the mean of data points in S_i .

$$\operatorname{argmin} \sum_{i=1}^k \sum_{x \in S_i} (x - \mu_i)^2$$

Equation 5.1 Objective of K-Means Algorithm

Because of its ability to minimize the distance between data points and the nearest centroid, the classic K-Means clustering algorithm proved to be useful in unsupervised learning module (Coates & Ng, 2012), which will be the next step. It was chosen for discretization in this study especially due to the multiple types of data involved: The continuous variables have different practical implications, units, ranges and distributions. It is difficult to use a unique parametric discretization function for all variables. Under this condition, clustering the data to minimize the within-cluster distance fits the objective of classification prediction, and suits all different variables. While K-Means clustering is often used in applications with multi-dimensional, large scale data (Kanungo et al., 2002; Xu &

Wunsch, 2008), BayesiaLab uses it in one-dimensional data following the same iterative optimization procedures (Bayesia, 2013b).

To determine the initial setting of k , the number of intervals in K-Means clustering, literature showed that there is no universal efficient method, but rather rely on heuristics and empirical approaches (Bradley & Fayyad, 1998; Jain, 2010; Ray & Turi, 1999; Xu & Wunsch, 2008). Dynamic techniques are also available to determine K , including the ISODATA (Iterative Self-Organizing Data Analysis Technique) method developed by Ball and Hall (Ball & Hall, 1967). But the implementation without computer automation assistance would require significant calculation efforts. This study took an experimental method to try out different selections of K and compare the results. Starting from the initial setting of $K = 4$, and comparing with the modeling results with $K = 3$ and $K = 5$. It was observed that the classification is not distinguishing enough when $K = 3$, while increasing K to 5 doesn't provide new knowledge. Figure 5.2 shows an example of the posterior inference classification of the three outcomes when MMA is U.S. East. The interpretation of the inference will be explained in detail in Section 5.2.2. Here, by comparing the 3 groups of posterior probabilities, it can be observed that when $K = 3$, the distribution of outcomes were very extreme as the possible values mostly fell into one interval. It was not accurate and maybe misleading. Compared to $K = 5$, the clustering when $K = 4$ provided narrower intervals, but it didn't add much to the knowledge obtained. As unsupervised learning serves mainly as qualitative analysis and is used to guide the direction of supervised

learning, $K = 4$ was considered the proper setting of the number of clustering intervals.

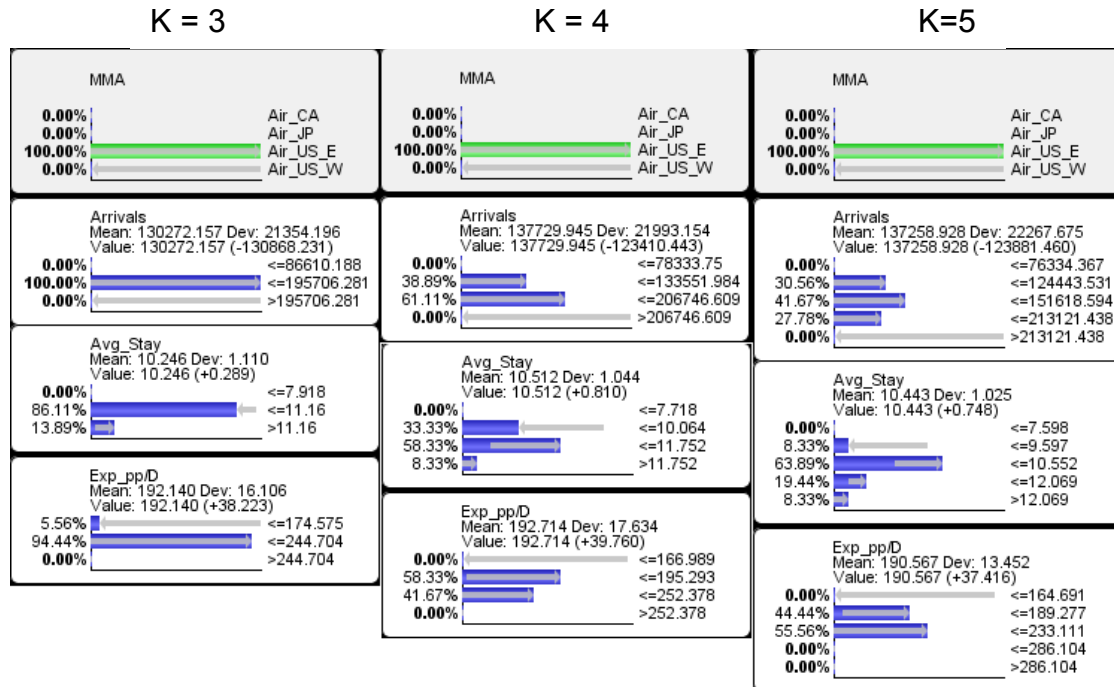


Figure 5.2 Comparison of Different Discretization Binning Selections

5.1.2 Step One Results

For each of the 3 outcomes, the MI value with each of the predictor was calculated. Because there is no evidence to suggest any of the outcome is more valuable than the others, no weight is assigned. The MI calculation results are summarized in Table 5.1. For each outcome, the 11 predictors are ranked by MI from the highest to the lowest. This ranking unveils 2 important commonalities for all the outcomes:

1. MMA has the highest MI values, and
2. POT_Mtg% and Month rank lowest.

Table 5.1 Ranked Predictors by MI with Each Outcome

Arrivals		Avg_Stay		Exp_pp/D	
Factors	MI ₁	Factors	MI ₂	Factors	MI ₃
MMA	1.5443	MMA	1.2746	MMA	1.4065
Stay_F&R%	1.125433	Stay_F&R%	1.226433	Stay_Hotel%	1.322833
Stay_Hotel%	1.056567	POT_Vst%	1.1325	Style_Grp%	1.183667
POT_Vst%	1.0104	Stay_B&B%	1.070633	Stay_B&B%	0.996
Rep%	0.936267	Stay_Hotel%	1.049133	Stay_F&R%	0.957933
Style_Grp%	0.845467	Style_Pkg%	0.885967	Style_Pkg%	0.880067
Stay_B&B%	0.845033	Style_Grp%	0.883667	POT_Vst%	0.854267
POT_Pls%	0.786933	Rep%	0.629433	POT_Pls%	0.4774
Style_Pkg%	0.645567	POT_Pls%	0.571467	Rep%	0.469233
POT_Mtg%	0.2757	Month	0.3361	POT_Mtg%	0.416667
Month	0.1878	POT_Mtg%	0.2062	Month	0.261867

For each predictor, the final MI is the average of the MI value with each of the 3 outcome variables. Based on the results above, Table 5.2 shows the reduced predictor set after screening at the end of Step 1. This 7-factor data set, together with the 3 outcomes, form the 10-node input data set for Step Two.

Table 5.2 Reduced Predictor Set after Step 1

Selected Predicators	
Factors	MI
MMA	1.408467
Stay_Hotel%	1.142844
Stay_F&R%	1.103267
POT_Vst%	0.999056
Style_Grp%	0.970933
Stay_B&B%	0.970556
Style_Pkg%	0.803867

5.2 Step Two: Unsupervised Learning

From Step One, it's known that MMA shares the most mutual information with all the 3 outcomes. In other words, given the knowledge of MMA, the uncertainty of these 3 nodes is reduced most. In Step Two, BayesiaLab's unsupervised learning feature is used to construct an initial Bayesian network. Further understanding of the relationships among all the factors is obtained through the unsupervised model.

5.2.1 Unsupervised Learning and Supervised Learning

In general, there are two ways to construct a Bayesian network. The first one is to build up a network according to the already known conditional dependence, similar as Figure 3.2 and Figure 3.3. But this method requires confidence in the initial structure which is not possessed in this analysis. Therefore, the other approach is used to define an evaluation function (or score) which accounts for the quality of candidate networks with respect to the available data and to use some kind of search algorithm in order to find an optimized network given the

conditions (Munteanu & Bendou, 2001). In other words, the network is built up based on the learned knowledge from given data sets.

As mentioned in the previous section 3.2, machine learning can be categorized as unsupervised and supervised learning algorithms. BayesiaLab supports these two learning modes too. In the domain of machine learning, unsupervised learning (or “learning without a teacher”) is to discover unknown structures of a data set, or in statistics term, the properties of the joint probability density $P(X)$ for a set of N observations (X_1, \dots, X_n) , without prior knowledge of the association between the observations and the output (Hastie, Tibshirani, Friedman, & Franklin, 2009b; Huang, Kecman, & Kopriva, 2006c). As opposed to supervised learning, a result of unsupervised learning is a new representation or explanation of the observed data (Huang et al., 2006c). In supervised learning, the goal is to use the inputs to predict the values of the outputs (Hastie, Tibshirani, Friedman, & Franklin, 2009a).

In recent studies, unsupervised learning and supervised learning have been used together as hybrid methods to solve problems (Huang et al., 2006c; Karayiannis & Mi, 1997; Zhao & Liu, 2007). When facing a new problem or a new domain, unsupervised learning is often used to obtain an initial understanding and to guide a more informed supervised learning that follows.

5.2.1.1 Unsupervised Learning Algorithm

In BayesiaLab, from a user's point of view, the difference between the two learning methods is that supervised learning must have a predefined target node. Unsupervised learning can be performed directly after data import. The software provides several algorithms to discover the probabilistic associations in the data, including Maximum Spanning Tree, Taboo, EQ, SopLEQ and Taboo Order. This study uses EQ framework for unsupervised learning. Compared to Maximum Spanning Tree, it results in a more optimal network (Bayesia, 2013a). Unlike the Taboo algorithm which is particularly useful for a network built by human experts or for updating a network learned on a different data set, it looks for the equivalence classes of Bayesian networks and applies to general data sets (Bayesia, 2013a). Compared to greedy search algorithms, the EQ algorithm is very efficient in avoiding local minima and reducing the search space size (Bayesia, 2013a; Munteanu & Bendou, 2001)

5.2.2 Step Two Analysis

After importing the data set from Step One and running unsupervised learning in EQ algorithm, the results Bayesian network is shown in Figure 5.3

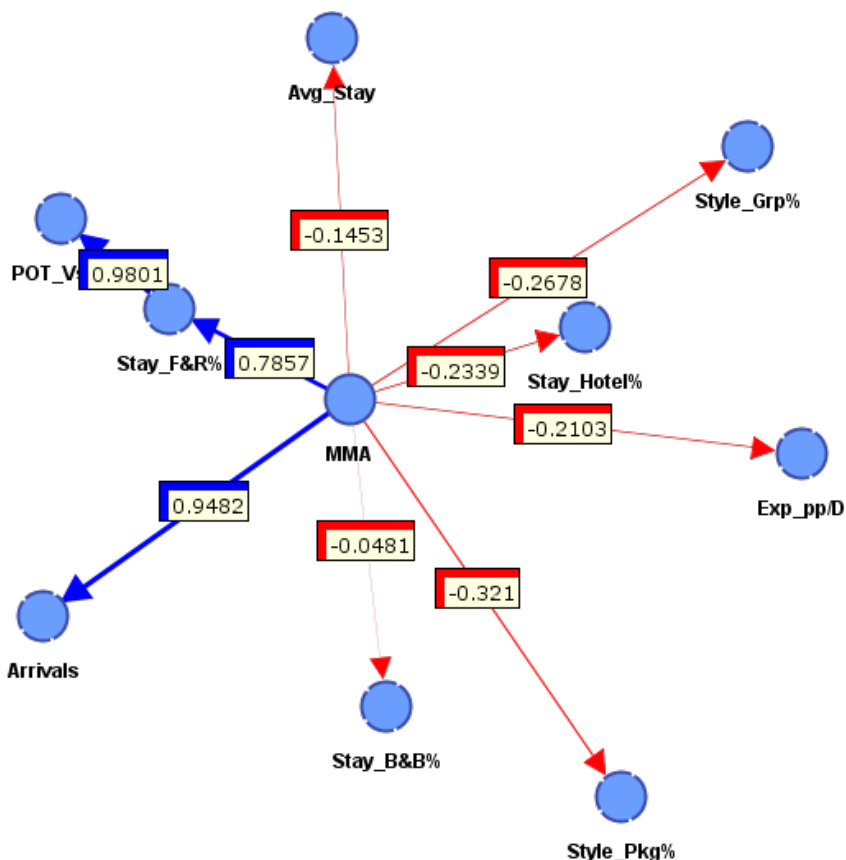


Figure 5.3 Unsupervised Learning Model

This graphical model uses the “distance mapping” feature of BayesiaLab. The length of the arcs is inversely proportional to the mutual information between the 2 connected nodes. Longer arc corresponds to smaller MI. In other words, if two nodes are close to each other in the 2-dimensional network, the mutual information between them is strong.

The color and numeric values of the arcs represent the Pearson correlation coefficient. Each number is the PCC value between the two nodes connected by the arc. Red color indicates negative correlation, blue positive. For example, at the upper-left corner of the figure, POT_Vst% is connected with Stay_F&R% with

a positive PCC value of 0.9801. It means the percentage of visitors whose purpose of visit was to visit friends and families is highly positively correlated with the percentage of visitors who chose to stay with friends and relatives. This is plausible based on common sense.

This unsupervised model shows what the data set tells without prior knowledge: An outstanding observation is that MMA is related to most of the other variables. In terms of probabilistic relationship, it means given the knowledge of MMA (knowing the region the visitors come from), the uncertainty of most of almost all the other factors are reduced (it is easier to infer how many visitors would arrive, how much they would spend, how long they would stay, what their choices for accommodation would be, etc).

This can also be illustrated by Figure 5.4, visual mapping of the model. In this figure, the arcs still indicate MI as with Figure 5.3, while the size of each node is proportional to its node force. In BayesiaLab, the total node force (NF) of a node is defined as the sum of the incoming forces and outgoing forces. The incoming force is the value of MI of an arc that goes into the node. The outgoing force is the value of MI of an arc that goes away from the node. The definition of the node force for node i is represented by Equation 5.2.

$$NF_i = \sum_{j=1}^n MI_j + \sum_{k=1}^m MI_k$$

Equation 5.2 Definition of Node Force

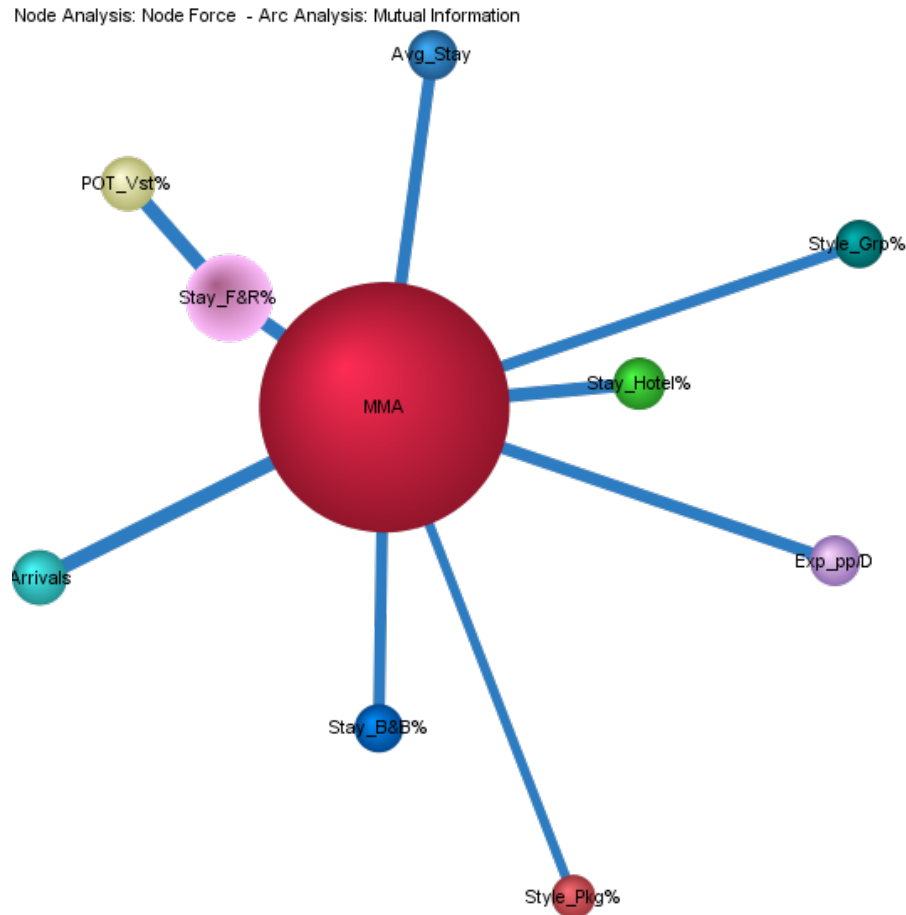


Figure 5.4 Node Force Mapping

5.2.2.1 Posterior Probability Inference

BayesiaLab's validation mode can simulate the posterior probability distribution of a node by setting the marginal probability distribution of another node connected to it. With the omnidirectional feature of Bayesian network, it is possible to look into the relationship between any two connected nodes. Figure 5.5 to Figure 5.7 show the posterior probability distribution of MMA, given the maximum setting of each outcome. Figure 5.5 shows that when the visitor arrivals reaches the highest level (more than 206,747 persons), the majority

visitors are most likely to come from U.S. West. Figure 5.6 shows that when the visitors stay for the longest period in Hawaii (longer than 11.75 days), they are most likely to come from Canada. From Figure 5.7, it's understood that visitors from Japan are most likely to spend most per person per day during their stay in Hawaii (more than \$252.378).

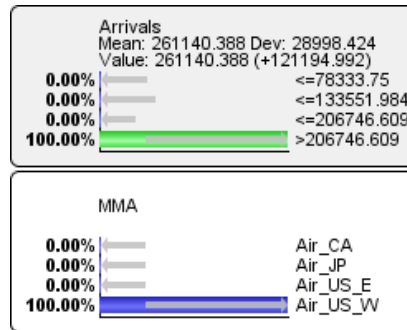


Figure 5.5 Posterior Probability distribution of MMA given Highest Arrivals

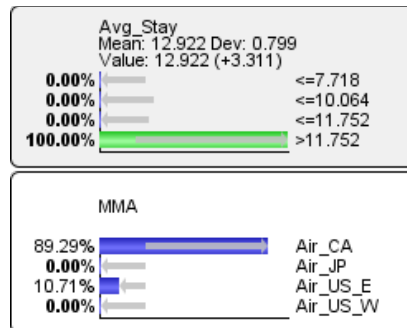


Figure 5.6 Posterior Probability distribution of MMA given Highest Avg_Stay

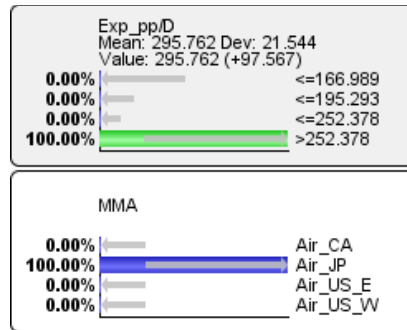


Figure 5.7 Posterior Probability distribution of MMA given Highest Exp_pp/D

Take a further step to investigate the relationships between MMA and the other predictors. Similarly, omnidirectional posterior probability distribution is used, but this time MMA is set as prior marginal probability to observe the changes in other visitor characteristics. Figure 5.8 shows a comparison of the visitor characteristics posterior probability distributions when MMA is set to Japan and U.S. West, respectively. Some very interesting findings include: visitors from Japan are most likely to stay at hotel and least likely to stay at relatives' or friends' home. They tend to travel with a group and purchase a package trip. Visitors from the U.S. West are almost the opposite.

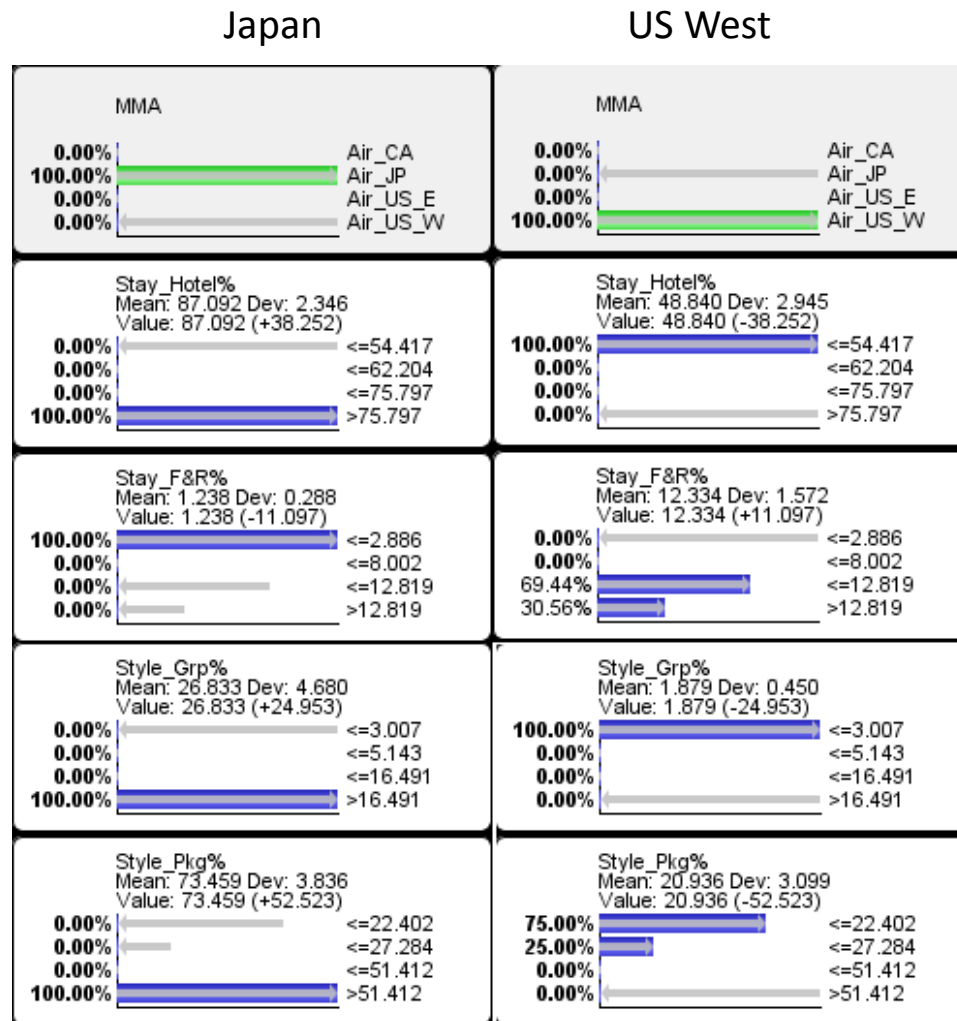


Figure 5.8 Visitor Characteristics Posterior Distribution: Japan v.s. US West

5.2.3 Predictor Ranking for Each MMA

Knowing that MMA is an effect modifier in the model, the analysis should go a further step to investigate the interactions within each MMA subgroup. Therefore, the data set is divided by MMA, and a second round of ranking is performed to clear out the predictors for each MMA. The ranking uses the same method as in Step One. In addition, the two factors excluded from Step One were included back to have a more comprehensive view. Table 5.3 shows the top 5 ranking

predictors for each MMA/outcome. Comparing the 4 MMAs, this ranking shows some noteworthy findings:

Commonalities across MMA:

1. For all the MMAs, **Month** has the strongest relationship with all 3 outcomes.
2. **Rep%** is strongly related to Avg_Stay and Exp_pp/D for all MMAs
3. **POT_Pls%** is strongly related to Arrivals for all MMAs

Uniqueness for each MMA (factor ranked top 5 for all 3 outcomes):

1. US_West: **Stay_B&B%**
2. US_East: **Rep%**
3. Japan: **Stay_Hotel%, Rep%**
4. Canada: **Stay_Hotel%, Rep%**

It is noteworthy that in Step One, Month was ranked at the bottom for all three outcomes. But in Step Two, it is the predominantly highest ranked factor. It is because the influence of Month was masked in the cross-region analysis. This finding also confirms the value of this analysis approach.

Table 5.3 Top 5 Factors by MI for Each Outcome and MMA

MMA	Arrivals		Avg_Stay		Exp_pp/D	
US_West	Month	1.3183	Month	1.5721	Month	0.5896
	POT_Pls%	0.7867	POT_Vst%	0.685	Style_Pkg%	0.2724
	Stay_B&B%	0.6046	Rep%	0.5507	Stay_B&B%	0.2309
	POT_Mtg%	0.5588	Stay_Hotel%	0.5128	Stay_F&R%	0.2303
	POT_Vst%	0.4211	Stay_B&B%	0.492	Rep%	0.1941
US_East	Month	1.6666	Month	1.5108	Month	0.8625
	Rep%	0.6193	Rep%	0.7548	POT_Vst%	0.4337
	Stay_B&B%	0.5722	Style_Grp%	0.5837	Stay_F&R%	0.3897
	Stay_Hotel%	0.5429	POT_Mtg%	0.5757	Rep%	0.3462
	POT_Pls%	0.4795	Stay_B&B%	0.4792	Style_Pkg%	0.2951
Japan	Month	0.926	Month	1.2139	Month	1.1012
	Stay_Hotel%	0.3863	Stay_Hotel%	0.6758	Stay_Hotel%	0.4831
	Rep%	0.3355	Style_Pkg%	0.5857	Rep%	0.4178
	Stay_F&R%	0.3096	Rep%	0.5826	POT_Vst%	0.4092
	POT_Pls%	0.2923	POT_Pls%	0.5212	Style_Grp%	0.3856
Canada	Month	1.4591	Month	1.6348	Month	0.7451
	Rep%	0.7859	Rep%	0.7984	POT_Mtg%	0.5067
	POT_Pls%	0.6845	POT_Pls%	0.634	Rep%	0.2528
	Stay_Hotel%	0.6168	Stay_Hotel%	0.6075	Stay_B&B%	0.2527
	POT_Vst%	0.3664	Style_Pkg%	0.4938	Stay_Hotel%	0.2225

Month, Rep%, POT_Pls% are common across some or all of the 4 MMAs.

5.2.4 Step Two Results

The observations from the unsupervised network and posterior probability inferences validate **Hypothesis 2** in Section 1.5 that visitors from different regions have different behaviors which will affect the outcomes.

Table 5.4 shows the list of factors for each MMA after the second round of ranking in Step Two. Including the three outcomes, each set has 10 factors.

Table 5.4 Data Set for Each MMA After Step 2

US West	US East	Japan	Canada
Arrivals	Arrivals	Arrivals	Arrivals
Avg_Stay	Avg_Stay	Avg_Stay	Avg_Stay
Exp_pp/D	Exp_pp/D	Exp_pp/D	Exp_pp/D
Month	Month	Month	Month
Stay_B&B%	Stay_B&B%	Stay_Hotel%	Stay_B&B%
POT_Pls%	Stay_Hotel%	POT_Pls%	Stay_Hotel%
POT_Mtg%	POT_Mtg%	POT_Vst%	POT_Pls%
POT_Vst%	POT_Pls%	Rep%	POT_Mtg%
Rep%	Rep%	Style_Pkg%	Rep%
Style_Pkg%	Style_Grp%	Style_Grp%	Style_Pkg%

5.3 Step Three: Supervised Learning

After the previous two steps, the analyst has obtained an overview of the system, the important associations and factors. But in order to make the information has applicable values, more in-depth analysis is needed to understand how the factors interact with each other. The unsupervised network sets foundation for more focused supervised learning. In Step Three, variables in Table 5.4 will be used to construct supervised learning Bayesian networks for each MMA. This variable set includes all the 3 outcomes. In the previous 2 steps, all the analysis were performed for separate outcomes. But in reality, there is no evidence that these outcomes are not independent from each other. Now it's time to examine how the they interact.

The very first step in supervised learning is to select a target node and its target state. In each network, there can only be one target. So for each MMA, with the same data set, the three outcomes will be set as target one after another. Also

similarly as with the posterior inference setting in Step Two, the target state is set to maximize the outcomes.

5.3.1 Supervised Learning Algorithm

BayesiaLab provides several types of supervised learning algorithms, among which are the well-known Markov Blanket Learning and Naive Bayes (including Augmented Naive Bayes). The difference between Markov Blanket and Naive Bayes algorithms is the method to search for nodes in the candidate network. Markov Blanket algorithm looks for nodes that belong to the Markov Blanket (father, son, spouse) (Pearl, 1988) centered with the target node. It is a minimal set of variables conditioned on which all other variables are probabilistically independent of the target (Tsamardinos, Aliferis, Statnikov, & Statnikov, 2003). Based on Naive Bayes classifier (Han et al., 2012c; Rish, 2001), a Naive Bayes network has a predefined architecture where the target node is the parent of all the other nodes (Bayesia, 2012; H. Zhang, 2004). This study wants to examine the relationships between the target node and all the other factors, so Augmented Naive Bayes algorithm is used. Compared to the classic Naive Bayes algorithm, the Augment algorithm extends additional unsupervised search that is performed on the basis of the given naive structure (Bayesia, 2012; H. Zhang, 2004).

5.3.2 Step Three Analysis and Results

Following the similar procedures as in Step Two, the supervised learning Bayesian network for each MMA is shown from Figure 5.9 to Figure 5.12, using

target node as Arrivals for a demonstration of the layout, and Distance Mapping based on mutual information. The posterior probability distribution of each network is shown from Table 5.5 to Table 5.8. Each outcome is set to the optimum level as marginal probability to observe its influence on the other factors, including the other two outcomes. The effect on each factor, excluding MMA which is discrete data, is measured by the extent of change of its mean value, calculated in percentage (CP). The change is either positive, indicating an increase of the mean, or negative, meaning a decrease. The most significant changes are highlighted by “****”, representing $CP \geq 10\%$. For MMA, since the source data came from monthly data from January, 2011 to December, 2013, the marginal probability for each month is 8.22% (1/12). The inference shows the posterior probability of each month, as included in the tables below.

5.3.2.1 U.S. West

Table 5.5 shows the results of supervised learning model for U.S. West. Summers months from June to August seem to attract the most visitors, December and January are dominantly the period when visitors tend to stay for long, and similarly, December to March are when visitors tend to spend more, together with September. Highest arrivals tend to be associated with lower percentages of visitors staying at B&B and visiting for meeting, convention or incentive, or visiting friends and relatives, and higher percentages of people on a package trip and visiting for pleasure. Yet higher percentages of package trip travellers tend to indicate lower length of stay. Instead, higher percentages of

staying at B&B and visiting friends and relatives are positively correlated with longer length of stay. For daily expenditures per person, higher percentages of people taking package trip and visiting for meeting and convention seem to indicate a lower expenditure. But staying at B&B and visiting family and friends are positive indicators of higher daily personal expenditures.

Table 5.5 Posterior Probability distribution of U.S. West_Supervised

Target Node			Arrivals (person)				Avg_Stay (days)			Exp_pp/D (\$)		
Target State			≥ 283622				≥ 10.2			≥ 158.2		
Posterior Influence												
Month			Aug(30%), Jul(30%), Jun(20%), Dec(10), Mar(10%)				Dec (50%), Jan (50%)			Sep(20%), Dec(20%), Feb(20%), Mar(30%)		
Variable Name	Unit	Prior Mean	Posterior Mean	Change		Posterior Mean	Change		Posterior Mean	Change		
Stay_B&B%	%	0.886	0.732	-17.38%	****	1.019	15.01%	****	0.973	9.82%	***	
POT_Pls%	%	82.829	85.008	2.63%	**	81.912	-1.11%	**	83.113	0.34%	*	
POT_Mtg%	%	4.269	2.807	-34.25%	****	4.394	2.93%	**	4.039	-5.39%	***	
POT_Vst%	%	11.349	10.95	-3.52%	**	12.82	12.96%	****	11.958	5.37%	***	
Rep%	%	81.469	80.906	-0.69%	*	83.591	2.60%	**	82.088	0.76%	*	
Style_Pkg%	%	20.173	21.955	8.83%	***	17.124	-15.11%	****	17.901	-11.26%	****	
Interaction	Arrivals	person	261140.389				254956.792	-2.37%	**	265817.415	1.79%	**
with other 2	Avg_stay	day	9.593	9.560	-0.34%	*				9.555	-0.40%	*
Outcomes	Exp_pp/D	\$	151.250	151.566	0.21%	*	153.252	1.32%	**			
Number of * indicates the absolute value of change measured by percentage (CP):												
* 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP≥10%												

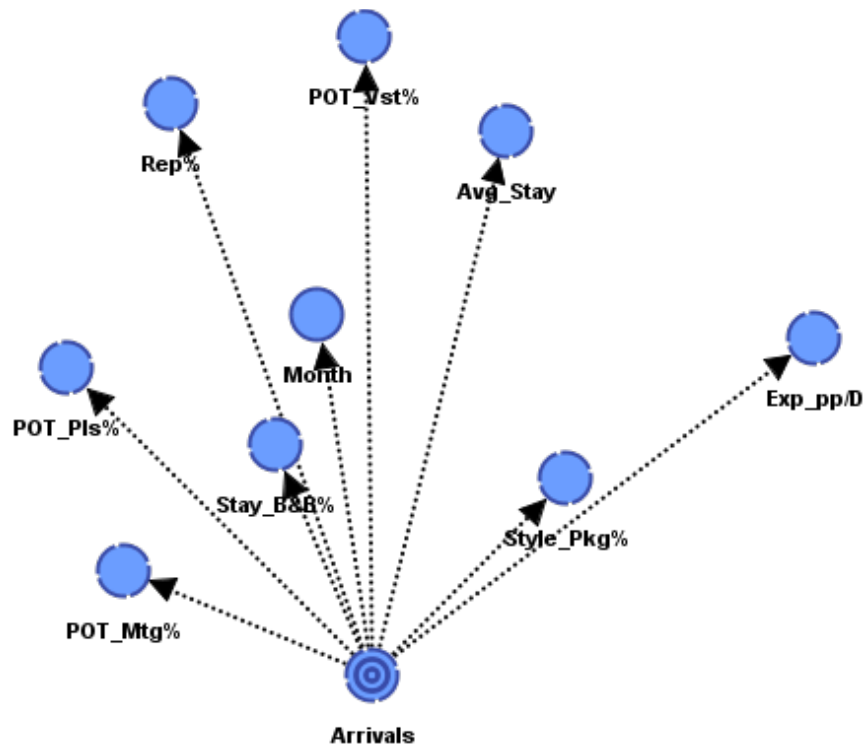


Figure 5.9 Supervised Model for U.S. West

5.3.2.2 U.S. East

Table 5.6 shows the results of supervised learning model for U.S. East. March, June and July seem to attract the most visitors, January is the single month contributing to the visitors staying for or above 12 days, and November and September are the months of higher expenditure. Similar as U.S. West, Hawaii is more likely to see larger volumes of visitors when less of them tend to stay at B&B or go to attend a meeting or convention, but when more of them visitor for pleasure. It also associated with lower percentage of visitors travelling with a group. But the same factor of smaller portion of group travellers tends to indicate longer lengths of stay, together with other factors including higher percentages of visitors staying at B&B, visiting for meetings, convention or incentive, and having

visited Hawaii before (Rep%), and lower percentage of visitors visiting for pleasure. When there is a higher portion of visitors going to attend a corporate meeting or convention, traveling with a group or stay at B&B, the daily personal expenditure tends to be higher. But meanwhile, the percentage of repeat visitors and the arrivals tend to go down.

Table 5.6 Posterior Probability distribution of U.S. East_Supervised

Target Node		Arrivals (person)				Avg_Stay (days)			Exp_pp/D (\$)		
Target State		≥ 156334				≥ 11.8			≥ 201.8		
Posterior Influence											
Month			Mar(33.33%), Jun(33.33%), Jul(33.33%)			Jan (100%)			Nov(33.33%), Sep(22.22%), Jan(11.11%), Apr(11.11%), Jun(11.11%), Aug(11.11%)		
Variable Name	Unit	Prior Mean	Posterior Mean	Change		Posterior Mean	Change		Posterior Mean	Change	
Stay_Hotel%	%	62.664	61.867	-1.27%	**	60.826	-2.93%	**	63.249	0.93%	*
Stay_B&B%	%	1.289	1.151	-10.71%	****	1.523	18.15%	****	1.336	3.65%	**
POT_Pls%	%	77.847	84.469	8.51%	***	69	-11.36%	****	76.344	-1.93%	**
POT_Mtg%	%	8.283	6.668	-19.50%	****	13.006	57.02%	****	8.924	7.74%	***
Rep%	%	57.983	56.87	-1.92%	**	64.156	10.65%	****	55.644	-4.03%	**
Style_Grp%	%	4.728	4.431	-6.28%	***	6.374	34.81%	****	4.899	3.62%	**
Interaction with other 2 Outcomes	Arrivals	person	140089.917			145090.077	3.57%	**	128197.020	-8.49%	***
	Avg_stay	day	10.454	10.123	-3.17%	**			10.370	-0.80%	*
	Exp_pp/D	\$	192.333	186.920	-2.81%	**	194.180	0.96%	*		

Number of * indicates the absolute value of change measured by percentage (CP):

* 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP≥10%

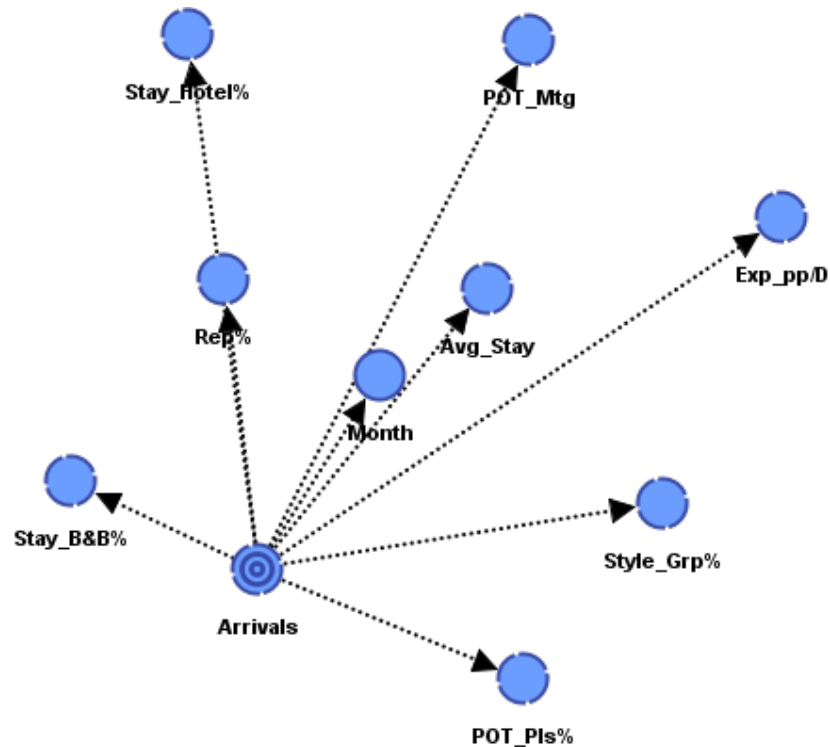


Figure 5.10 Supervised Model for U.S. East

5.3.2.3 Japan

Table 5.7 shows the results of supervised learning model for Japan. From August appears most likely to have large amount of visitors, followed by September, October and December. August is also the month most likely to see longer lengths of stay, followed by the neighboring months July and September. From October to January, together with July, the daily personal expenditures tend to be higher. A higher percentage of repeat visitors and a lower percentage of package trip travellers, as well as a lower daily personal expenditure tend to go with higher arrivals. The same factors are also associated with longer lengths of stay, except for the percentage of visitors whose purpose was to see family and friends – for higher arrivals this factor tends to be lower, while for length of stay it tends to be

higher. For daily expenditure per person, Japanese visitors have the highest values among all 4 MMAs. And a higher expenditure is associated with lower percentage of family and friends visitors and repeat visitors, and higher percentages of visitors going with a group and taking a package trip. In addition, higher expenditure tends to go with shorter average lengths of stay.

Table 5.7 Posterior Probability distribution of Japan_Supervised

Target Node			Arrivals (person)			Avg_Stay (day)			Exp_pp/D (\$)			
Target State			≥ 134867			≥ 6.2			≥ 313.5			
Prior Status			Posterior Influence									
Month			Aug(50%), Sep(16.67%), Oct(16.67%), Dec(16.67%)			Aug(60%), Jul (20%), Sep(20%)			Oct(20%), Nov(20%), Jan(20%), Dec(12.5%), Jun(12.5%)			
Variable Name	Unit	Prior Mean	Posterior Mean	Change		Posterior Mean	Change		Posterior Mean	Change		
Stay_Hotel%	%	87.092	84.697	-2.75%	**	82.753	-4.98%	**	88.681	1.82%	**	
POT_Pls%	%	83.866	86.479	3.12%	**	90.405	7.80%	***	82.823	-1.24%	**	
POT_Vst%	%	1.781	1.696	-4.77%	**	1.839	3.26%	**	1.739	-2.36%	**	
Rep%	%	58.492	65.332	11.69%	****	67.597	15.57%	****	56.515	-3.38%	**	
Style_Grp%	%	27.024	24.195	-10.47%	****	22.631	-16.26%	****	28.067	3.86%	**	
Style_Pkg%	%	73.459	69.982	-4.73%	**	66.472	-9.51%	***	75.525	2.81%	**	
Interaction with other 2	Arrivals	person	117239.111				133555.800	13.92%	****	118099.688	0.73%	*
	Avg_stay	day	5.970	6.127	2.63%	**				5.844	-2.11%	**
Outcomes	Exp_pp/D	\$	293.025	272.086	-7.15%	***	271.422	-7.37%	***			

Number of * indicates the absolute value of change measured by percentage (CP):

* 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP≥10%

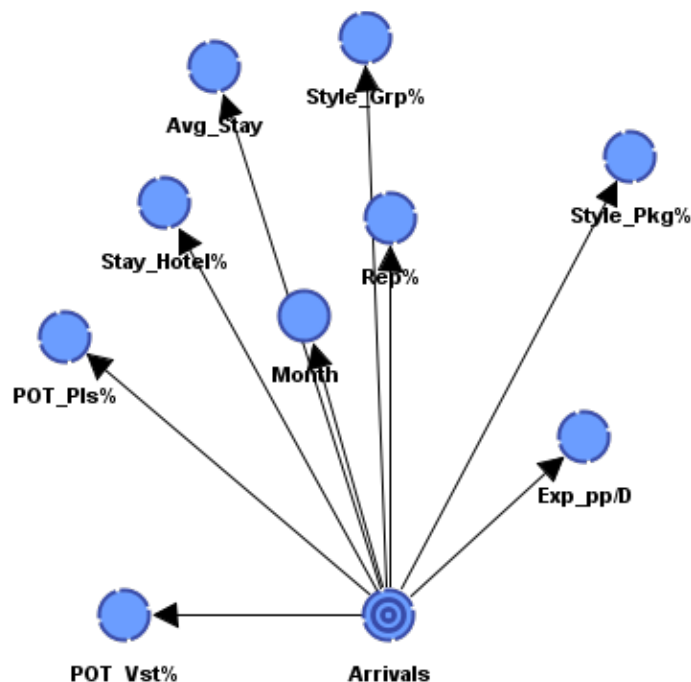


Figure 5.11 Supervised Model for Japan

5.3.2.4 Canada

Table 5.8 shows the results of supervised learning model for Canada. Winter months from December to March are most likely to have larger volumes of visitors and longer length of stay. January is also the month most likely to see higher daily personal expenditure, followed by February, June, September and November. To see a higher incoming flow of visitors, there tend to be lower percentages of visitors staying at hotel or B&B, or traveling on a package trip, but higher percentages of repeat visitors and people going to Hawaii for pleasure. Longer average length of stay is also associated with higher percentages of repeat visitors, as well as higher percentages of people who are attending a meeting or convention. Similar to arrivals, choices of stay at hotels and B&Bs and package trip travellers are negatively correlated with lengths of stay. With a

higher daily personal expenditure, the percentages of people staying at hotel, on a package trip and traveling for meeting, convention or incentives tend to be lower, and the percentages of visitors staying at B&B and having visited Hawaii before tend to be higher. Both longer lengths of stay and higher expenditure see a higher amount of arriving visitors.

Table 5.8 Posterior Probability distribution of Canada_Supervised

Target Node			Arrivals (person)				Avg_Stay (days)			Exp_pp/D (\$)		
Target State			≥ 55615				≥ 13.7			≥ 165.4		
Posterior Influence												
Month			Dec(25%), Jan(25%), Feb(25%), Mar(25%)				Jan (60%), Dec(20%), Feb(20%)			Jan(33.33%), Feb(16.67%), Jun(16.67%), Sep(16.67%), Nov(16.67%)		
Variable Name	Unit	Prior Mean	Posterior Mean	Change		Posterior Mean	Change		Posterior Mean	Change		
Stay_Hotel%	%	49.971	46.29	-7.37%	***	45.932	-8.08%	***	48.044	-3.86%	**	
Stay_B&B%	%	1.420	1.339	-5.70%	***	1.366	-3.80%	**	1.539	8.38%	***	
POT_Pls%	%	91.824	94.486	2.90%	**	94.115	2.49%	**	92.613	0.86%	*	
POT_Mtg%	%	3.633	3.596	-1.02%	**	4.041	11.23%	****	3.558	-2.06%	**	
Rep%	%	61.536	67.663	9.96%	***	67.334	9.42%	***	63.404	3.04%	**	
Style_Pkg%	%	25.853	24.418	-5.55%	***	24.117	-6.71%	***	24.113	-6.73%	***	
Interaction with other 2 Outcomes	Arrivals	person	41312.167				66373.500	60.66%	****	48602.037	17.65%	****
	Avg_stay	day	12.426	13.467	8.38%	***				12.968	4.36%	**
	Exp_pp/D	\$	156.169	155.870	-0.19%	*	161.135	3.18%	**			
Number of * indicates the absolute value of change measured by percentage (CP):												
* 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP ≥10%												

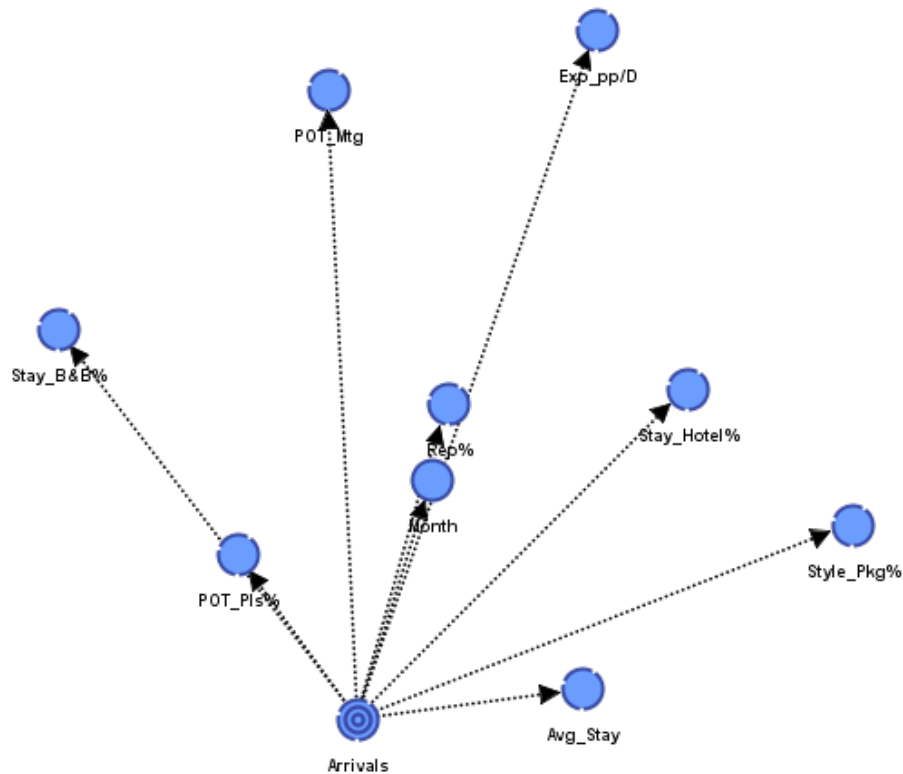


Figure 5.12 Supervised Model for Canada

5.4 Validation

In order to validate the model and avoid testing hypotheses suggested by the data, also known as Type III error (Mosteller, 2006), this research used three methods for validation: 1) cross validation using the original data set to validate the algorithms and modeling approach, 2) validation with an additional unseen set of more recent visitors data to validate the predictive inference, and 3) validate the results against existing knowledge from literature and professional opinions. This section will introduce the first two validation methods, and the third method is included in Chapter 7.

5.4.1 Cross Validation within Original Data Set

The purpose of cross validation is to evaluate a statistics analysis by assessing how well the results can be generalized to other data independent of the data used for the analysis. It is often used for model selection by comparing the prediction accuracy and sensibility of several candidates (Devijver & Kittler, 1982; Geisser, 1993; Kohavi, 1995). In general, cross validation involves partitioning the data set into a training subset and a testing subset. The training subset is used to generate the model, which is to be validated by the testing subset, also known as the unseen testing data set in the sense that it was not involved in the model building. The performance of the model is measured by the variance.

For classification problems, the fitness of a model can be measured by whether the classification is correct or incorrect – the misclassification error rate. For continuous value prediction, the variances are measured by the deviation of the predicted results. In this research, the primary objective is to find out the relationships among factors, first qualitatively via the unsupervised Bayesian network model across MMA, and then both qualitatively and quantitatively via the supervised models for each MMA. In order to test the complete research approach, both unsupervised and supervised models need to be validated.

In this analysis, the aggregate visitors data of each MMA in month is an instance (144 instances). But considering the application context, the data set used for analysis should cover all 12 months in a calendar year to model a complete pattern that can have sensible implications. Previous analysis already showed

seasonal trends on several factors. It would be biased to test the model built with data from January to June against the testing data set consisting data from July to December.

Therefore, the original data set of 36 months in 3 years was divided into 3 subsamples, each including 12 months from January to December. One important prerequisite for cross validation to be yield meaningful results is that the training data set and the testing data set are from the same population, meaning that the data structure doesn't vary within the data set. In the tourism market, some factors could cause a change in the data structure, like great events and conventions (e.g. the Olympics), natural disaster, political or military turmoil, pandemic diseases, or even financial crisis. Such factors could dramatically increase or decrease the performance in a year. Section 7.3 will discuss the treatment of outliers. Here, before running cross validation, the researcher took a quick look at the trend for each MMA over the 3 years. Take U.S. West as an example, as shown in Figure 5.13, no outlier was observed.

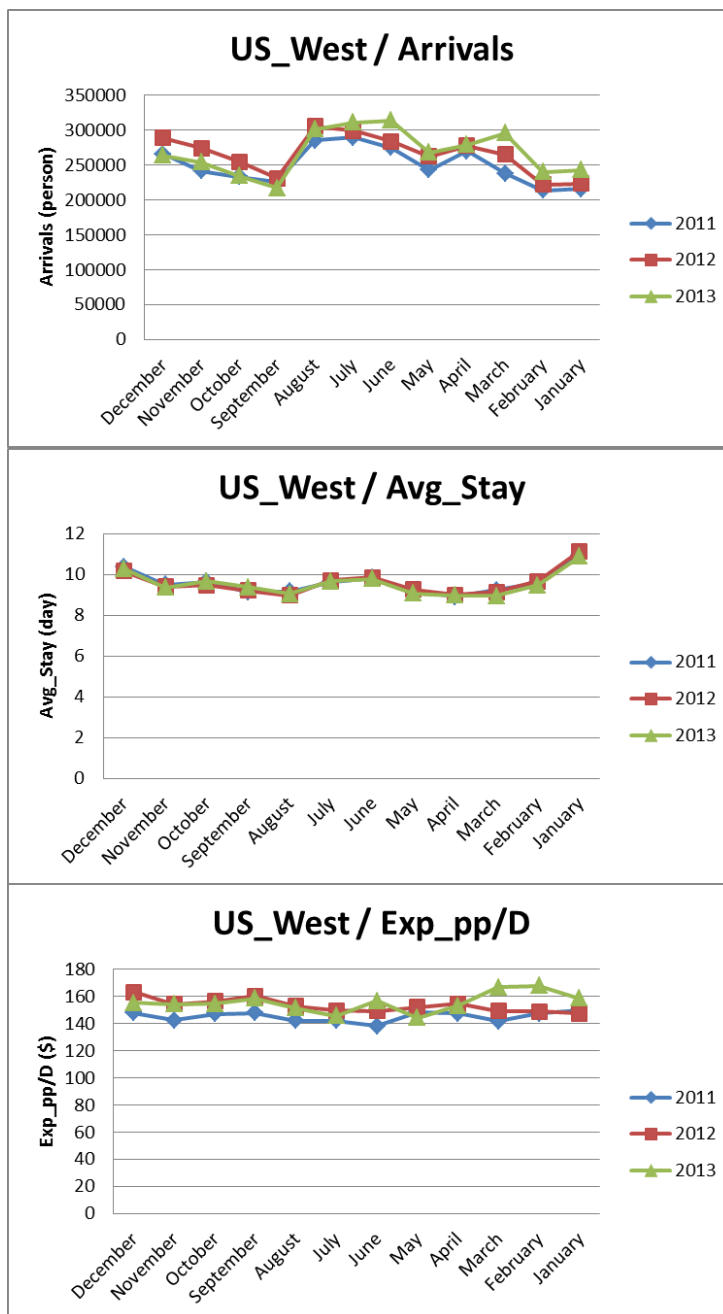


Figure 5.13 3-Year Trend_U.S. West

In a classic k-fold cross validation, the complete data set is partitioned into k subsamples of equal size. A single subsample is taken out as the testing data set, while the rest (k-1) subsamples are used as training data set. The validation is

repeated k times so that each subsample is used once and only once as the training data set. And the validation results of all k folds are averaged out to generate a single estimation of variance. By comparing the variances of multiple candidate models, or comparing the learning rate needed for each model to reach a satisfying prediction accuracy, the best model is selected. In this research, however, no prior prediction method or benchmark exists for comparison. So cross validation is employed here mainly to check against overfitting and to verify the model's algorithms. In this 3-fold cross validation, one subsample group was retained as the testing subset, and the rest two groups were used as the training subset as input to BayesiaLab, following the same analysis procedures introduced in Section 5.1 to 5.3. Such a validation process was repeated 3 times. The following sub-sections, a detailed description for the validation round using data of year 2011 as the testing data set and 2012 and 2013 together as the training data set. The other 2 rounds of rotational validation showed very similar results.

5.4.1.1 Validation of Unsupervised Model

Using the same approach, Table 5.9 shows the MI ranking for each factor across MMA for each outcome, together with the ranking by average MI of all 3 outcomes. Comparing with the MI ranking of the complete data set in Table 5.1, the ranking orders are highly consistent with only 2 differences in the ranking for daily personal expenditure: Stay_B&B% is ranked 3rd in Table 5.1 but 6th here; Rep% is slightly lower (gap < 0.01) than POT_PIs% in Table 5.1, but higher here.

But these differences in absolute values are small, so overall the average ranking across 3 outcome factors is the same as Table 5.2.

Table 5.9 Predictors Ranking by Mutual Information with Each Outcome and Average_Training Data Set

Arrivals		Avg_Stay		Exp_pp/D		Average MI Ranking	
Factors	MI	Factors	MI	Factors	MI	Factor	Average MI
MMA	1.5979	MMA	1.2972	MMA	1.5000	MMA	1.4650
Stay_F&R%	1.0979	Stay_F&R%	1.1763	Stay_Hotel%	1.4137	Stay_Hotel%	1.1509
Stay_Hotel%	1.0556	POT_Vst%	1.1185	Style_Grp%	1.2095	Stay_F&R%	1.0914
POT_Vst%	1.0348	Stay_B&B%	1.0728	Stay_F&R%	1.0000	POT_Vst%	1.0301
Rep%	0.9271	Stay_Hotel%	0.9835	POT_Vst%	0.9369	Style_Grp%	0.9896
Style_Grp%	0.8733	Style_Pkg%	0.9023	Stay_B&B%	0.9267	Stay_B&B%	0.9439
Stay_B&B%	0.8321	Style_Grp%	0.8859	Style_Pkg%	0.8535	Style_Pkg%	0.8173
POT_Pls%	0.7897	Rep%	0.5929	Rep%	0.4673	Rep%	0.6624
Style_Pkg%	0.6962	POT_Pls%	0.5467	POT_Pls%	0.3767	POT_Pls%	0.5710
POT_Mtg%	0.2060	Month	0.2821	POT_Mtg%	0.2590	Month	0.2002
Month	0.1723	POT_Mtg%	0.0945	Month	0.1462	POT_Mtg%	0.1865

Using Table 5.11 for unsupervised learning, the Bayesian network is shown in Figure 5.14, with the same meaning of legends: the lengths of arcs are inversely proportional to mutual information values, and the color and numeric labels on of the arcs indicate Pearson's Correlation coefficient values. Comparing with the unsupervised model of the complete data set in Figure 5.3, the network structures are very similar: MMA is related to most of the other factors. Figure 5.14 shows one more connection between Stay_Hotel% and Style_Grp%, which are highly positively correlated. And the posterior inference also shows consistency with Figure 5.5 to Figure 5.7, as demonstrated in Figure 5.15 to Figure 5.17: Visitors from U.S. West are most likely to have the highest arrivals, visitors from Canada are most likely to stay for the longest period, and visitors

from Japan tend to generate highest expenditures during their stay in Hawaii.

Comparing Figure 5.8 with Figure 5.18, the same contrast of visitor characteristics can be observed between Japan and U.S. West.

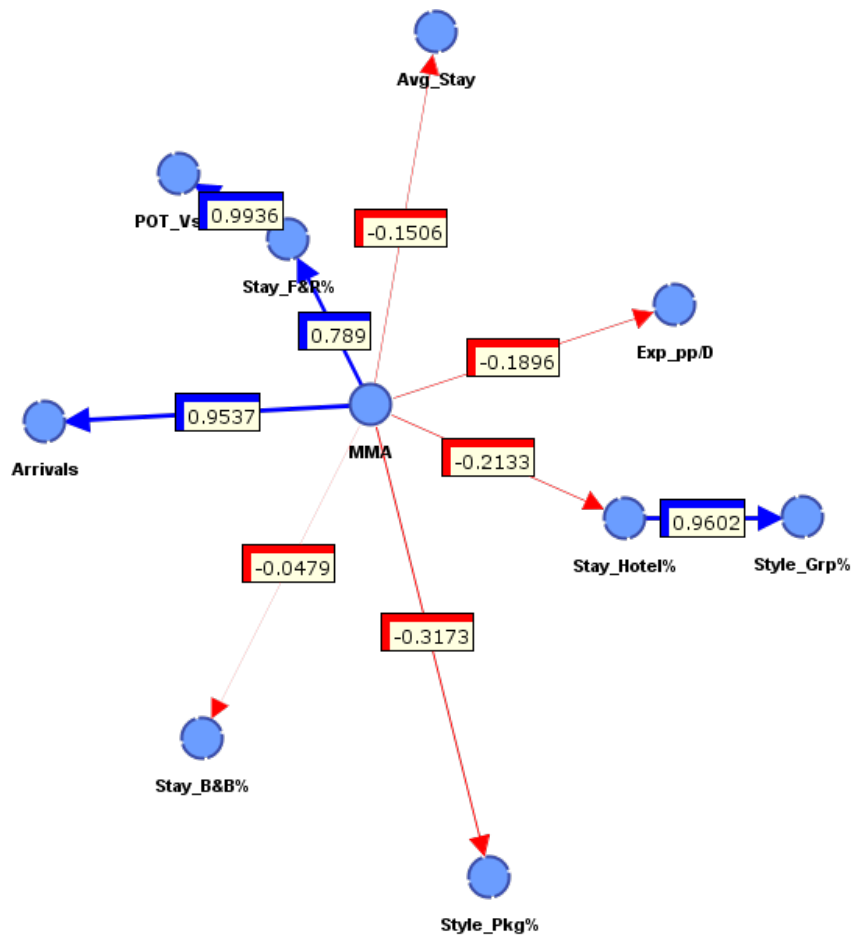


Figure 5.14 Unsupervised Learning Model_Training Data Set

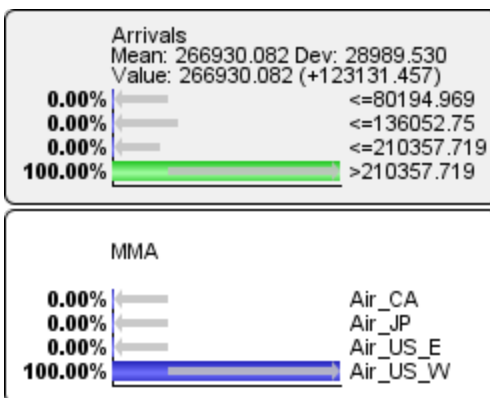


Figure 5.15 Posterior Probability distribution of MMA given Highest Arrivals_ Training Data Set

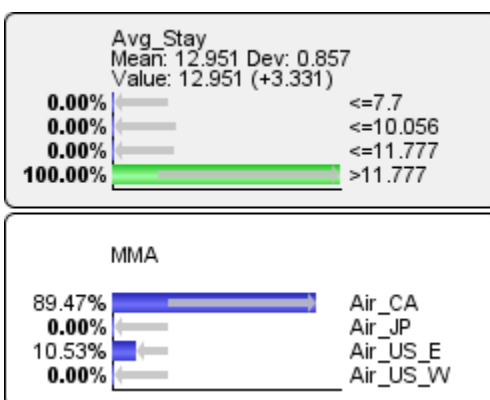


Figure 5.16 Posterior Probability distribution of MMA given Highest Avg_Stay_ Training Data Set

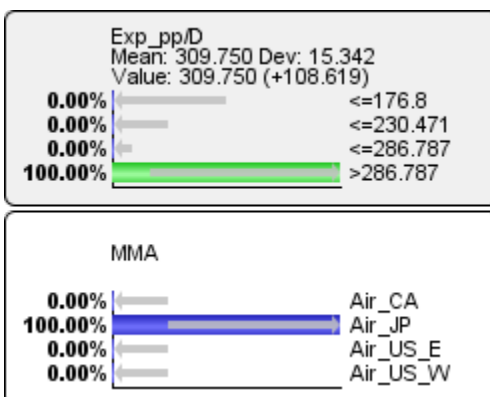


Figure 5.17 Posterior Probability distribution of MMA given Highest Exp_pp/D_ Training Data Set

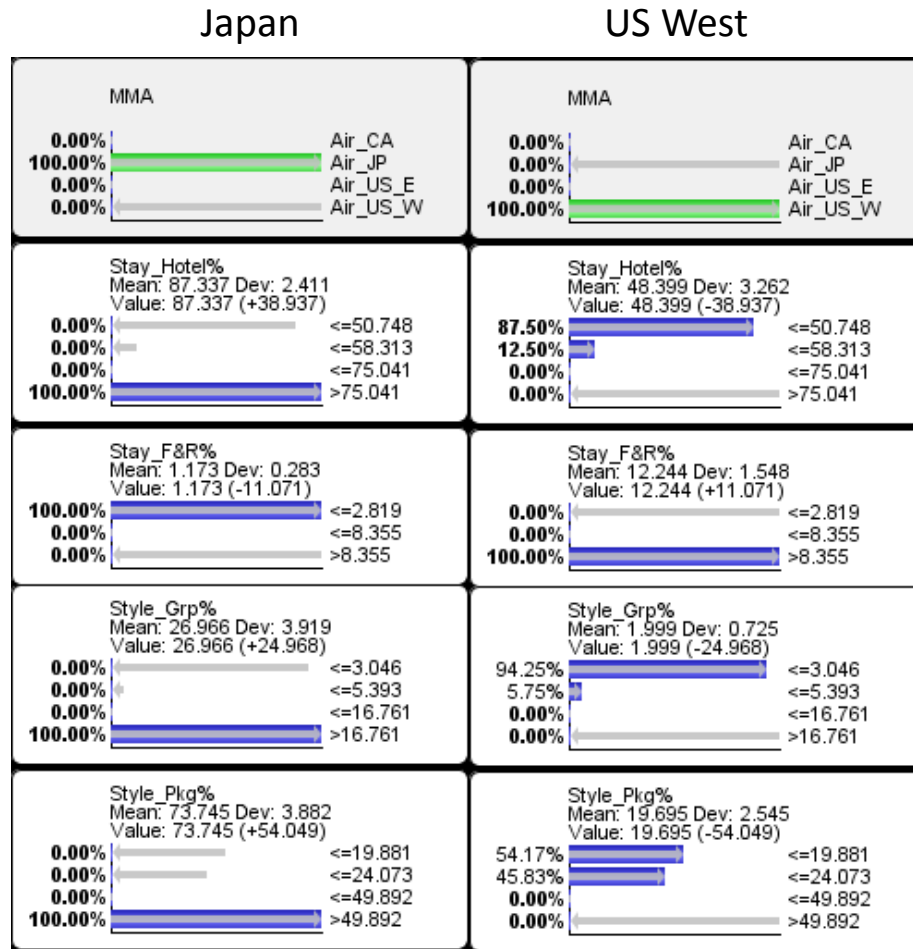


Figure 5.18 Visitor Characteristics Posterior Distribution: Japan v.s. US West_Training Data Set

At this point, the unsupervised model based on the training data set proved to be highly consistent with the unsupervised model trained with the complete data set. Next, the subsample unsupervised model is to be tested against the testing data set.

The factor ranking by MI with each outcome and the ranking by average MI value across MMA for the testing data set is shown in Table 5.10. The differences in ranking orders from the training data set and the complete data set also exist for

Exp_pp/D. The ranking by average MI is still highly consistent with the training data set and the complete data set.

Table 5.10 Predictors Ranking by Mutual Information with Each Outcome and Average _Testing Data Set

Arrivals		Avg_Stay		Exp_pp/D		Average MI Ranking	
Factors	MI	Factors	MI	Factors	MI	Factor	Average MI
MMA	1.5479	MMA	1.2449	MMA	1.4402	MMA	1.4110
Stay_F&R%	1.1568	Stay_F&R%	1.2272	Stay_Hotel%	1.3426	Stay_Hotel%	1.1959
Stay_Hotel%	1.0492	Stay_Hotel%	1.1514	Style_Grp%	1.2344	Stay_F&R%	1.1217
POT_Vst%	0.9461	POT_Vst%	1.0785	Style_Pkg%	0.9971	Style_Grp%	0.9735
Rep%	0.9171	Stay_B&B%	0.9743	Stay_F&R%	0.9811	POT_Vst%	0.9395
Style_Grp%	0.8338	Style_Pkg%	0.8643	Stay_B&B%	0.9797	Stay_B&B%	0.9159
Stay_B&B%	0.7938	Style_Grp%	0.8522	POT_Vst%	0.7940	Style_Pkg%	0.8259
POT_Pls%	0.7661	POT_Pls%	0.6546	POT_Pls%	0.6824	POT_Pls%	0.7010
Style_Pkg%	0.6162	Rep%	0.5596	Rep%	0.5635	Rep%	0.6801
Month	0.1900	Month	0.3849	POT_Mtg%	0.4615	Month	0.2740
POT_Mtg%	0.1603	POT_Mtg%	0.1743	Month	0.2470	POT_Mtg%	0.2654

To test the posterior inference of the unsupervised model in Figure 5.15 to Figure 5.18, the data of the testing data set was analyzed using the basic sorting feature of Microsoft Excel spreadsheet. The results shown in Figure 5.19 to Figure 5.22 are consistent with the posterior classification inference of Bayesian network model.

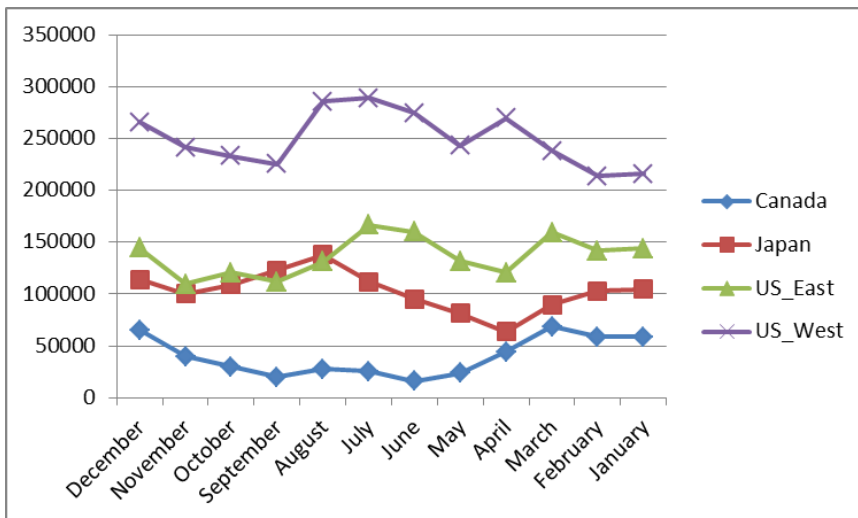


Figure 5.19 Arrivals of 4 Regions_Testing Data Set

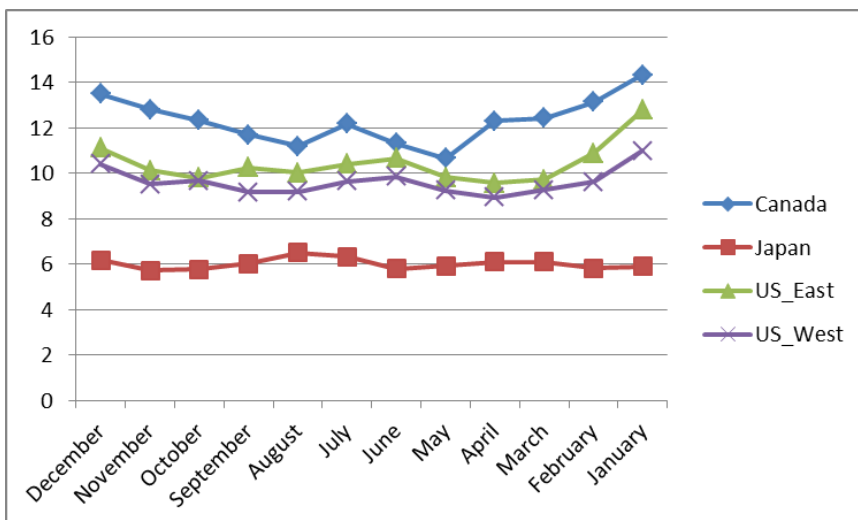


Figure 5.20 Average Lengths of Stay (day) of 4 Regions_Testing Data Set

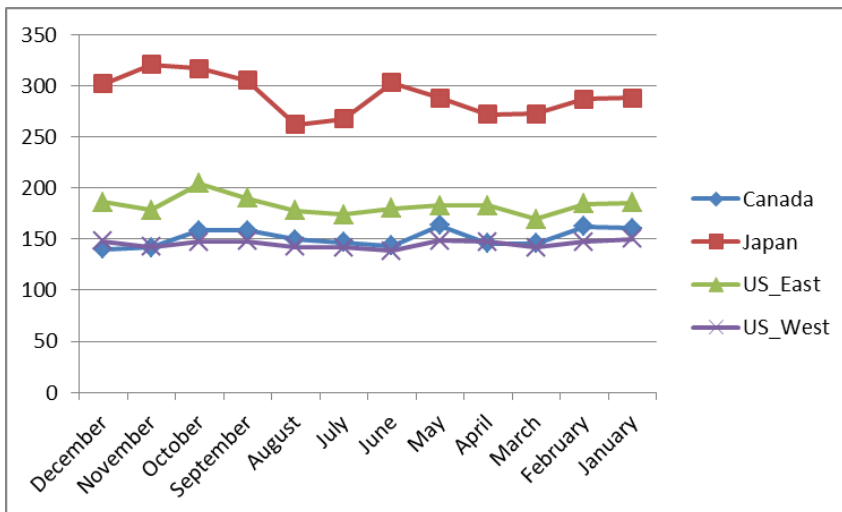


Figure 5.21 Daily Expenditure per Person (\$) of 4 Regions_Testing Data Set

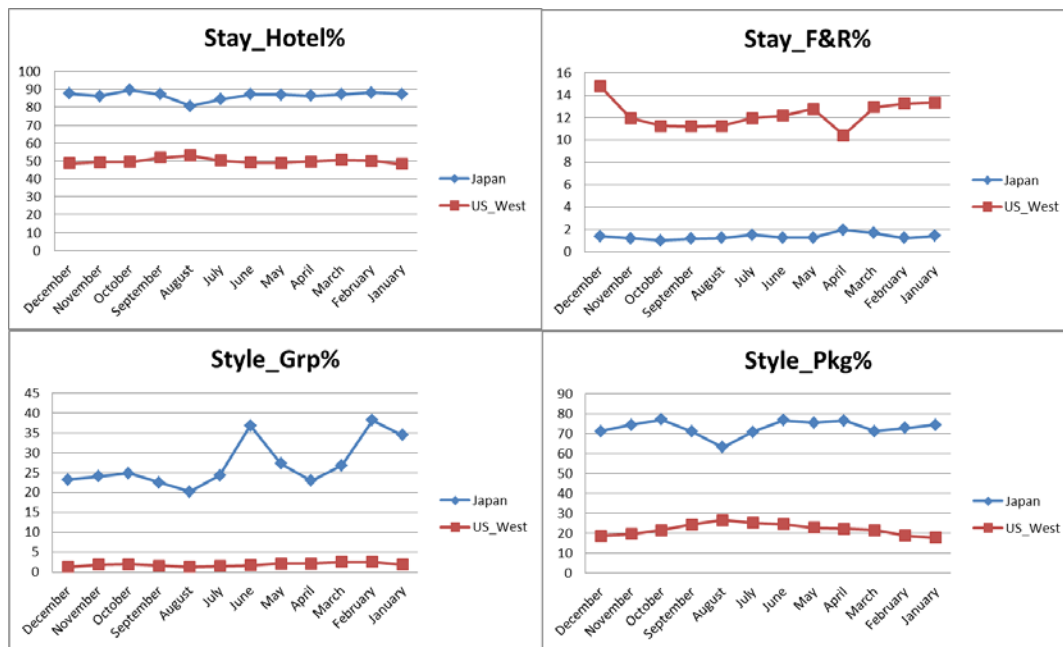


Figure 5.22 Visitors Characteristics: Japan v.s. U.S. West_Testing Data Set

5.4.1.2 Validation of Supervised Model

So far, the unsupervised model based on the training data set has been validated by the testing data set, and also proved to be consistent with the unsupervised

model trained with the complete data set. Next, the supervised model trained for each region is to be tested.

Table 5.11 shows the top 5 factors ranking by MI for each outcome and each MMA. As observed in Table 5.3, Month remains the universal top factor, and Rep% also appears among the top 5 for all regions and all outcomes. Besides, some regional features are observed: Stay_B&B% is among top 5 for all outcomes for U.S. West, Style_Grp% for U.S. East, Stay_Hotel for Japan and Canada, and POT_Pls% for Canada.

Table 5.11 Top 5 Factors by MI for Each Outcome and MMA Training Data Set

MMA	Arrivals		Avg_Stay		Exp_pp/D	
US_West	Month	1.5733	Month	1.5051	Month	1.3122
	POT_Pls%	0.8745	POT_Pls%	0.6192	Stay_Hotel%	0.4888
	Stay_B&B%	0.8175	Stay_B&B%	0.5512	Stay_B&B%	0.3583
	POT_Mtg%	0.5522	Rep%	0.5265	Style_Pkg%	0.3541
	Rep%	0.4688	POT_Vst%	0.4653	Rep%	0.3022
US_East	Month	1.8250	Month	1.5255	Month	1.2178
	POT_Pls%	0.7085	Rep%	0.6719	POT_Vst%	0.6549
	Stay_B&B%	0.6876	Style_Grp%	0.6089	Rep%	0.4730
	Style_Grp%	0.6594	POT_Pls%	0.5280	Style_Grp%	0.3965
	Rep%	0.6526	Stay_F&R%	0.5010	Stay_F&R%	0.3447
Japan	Month	1.6625	Month	1.5051	Month	1.3011
	Stay_Hotel%	0.8081	Stay_Hotel%	0.5674	POT_Vst%	0.5709
	Rep%	0.7179	Rep%	0.4976	Rep%	0.4917
	Style_Pkg%	0.5900	POT_Mtg%	0.4934	POT_Pls%	0.4594
	Style_Grp%	0.4917	POT_Vst%	0.4840	Stay_Hotel%	0.4425
Canada	Month	1.8657	Month	1.6991	Month	1.2663
	POT_Pls%	1.0410	Stay_Hotel%	1.0460	Stay_B&B%	0.4859
	Rep%	1.0255	Rep%	1.0093	Rep%	0.4423
	Stay_Hotel%	0.7808	POT_Pls%	0.7198	POT_Pls%	0.3593
	Stay_B&B%	0.5498	Style_Pkg%	0.6414	Stay_Hotel%	0.3234

Table 5.12 shows the factors for supervised learning for each MMA, according to their average MI with the 3 outcomes.

Table 5.12 List of Factors for Supervised Learning_Training Data Set

US West	US East	Japan	Canada
Arrivals	Arrivals	Arrivals	Arrivals
Avg_Stay	Avg_Stay	Avg_Stay	Avg_Stay
Exp_pp/D	Exp_pp/D	Exp_pp/D	Exp_pp/D
Month	Month	Month	Month
POT_Pls%	POT_Pls%	Stay_Hotel%	POT_Pls%
Stay_B&B%	Stay_B&B%	Rep%	Rep%
POT_Mtg%	Style_Grp%	Style_Pkg%	Stay_Hotel%
Rep%	Rep%	Style_Grp%	Stay_B&B%
POT_Vst%	Stay_F&R%	POT_Mtg%	Style_Pkg%
Stay_Hotel%	POT_Vst%	POT_Vst%	

Based on the training data sets, supervised Bayesian networks were built up for each MMA. Table 5.13 to Table 5.16 show the posterior inference of each factor's value change when each target node is set to the target node.

Table 5.13 Posterior Probability distribution of U.S. West_Supervised_Training Data Set

Target Node		Arrivals (person)				Avg_Stay (days)				Exp_pp/D (\$)				
Target State		≥ 292450				≥ 10.3				≥ 161.9				
Posterior Influence														
Month			Mar(16.67%), Jun(16.67%), Jul(33.33%), Aug(33.33%)				Jan (66.67%), Dec (33.33%),				Dec(33.33%), Feb(33.33%), Mar(33.33%)			
Variable Name	Unit	Prior Mean	Posterior Mean	Change		Posterior Mean	Change		Posterior Mean	Change				
Stay_Hotel%	%	49.368	50.187	1.66%	**	48.602	-1.55%	**	48.939	-0.87%	*			
Stay_B&B%	%	0.897	0.739	-17.61%	****	1.073	19.62%	****	1.023	14.05%	****			
POT_Pls%	%	83.024	85.408	2.87%	**	81.539	-1.79%	**	83.167	0.17%	*			
POT_Mtg%	%	4.198	2.865	-31.75%	****	5.281	25.80%	****	4.325	3.03%	**			
POT_Vst%	%	11.317	10.508	-7.15%	***	12.785	12.97%	****	12.479	10.27%	****			
Rep%	%	81.462	80.565	-1.10%	**	83.562	2.58%	**	82.885	1.75%	**			
Interaction with other 2 Outcomes	Arrivals	person	266930.082			240407.873	-9.94%	***	271619.824	1.76%	**			
	Avg_stay	day	9.580	9.433	-1.53%	**			9.424	-1.63%	**			
	Exp_pp/D	\$	154.312	154.392	0.05%	*	153.093	-0.79%	*					

Number of * indicates the absolute value of change measured by percentage (CP):

* 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP ≥ 10%

Table 5.14 Posterior Probability distribution of U.S. East_Supervised_Training Data Set

Target Node		Arrivals (person)				Avg_Stay (days)				Exp_pp/D (\$)				
Target State		≥ 157961				≥ 11.9				≥ 204.0				
Posterior Influence														
Month			Mar(33.33%), Jun(33.33%), Jul(33.33%)				Jan (100%)				Jan(14.29%), Apr(14.29%), Aug(14.29%), Sep(28.57%), Oct(28.57%)			
Variable Name	Unit	Prior Mean	Posterior Mean	Change		Posterior Mean	Change		Posterior Mean	Change				
Stay_B&B%	%	1.310	1.151	-12.14%	****	1.504	14.81%	****	1.382	5.50%	***			
Stay_F&R%	%	12.076	11.305	-6.38%	***	11.305	-6.38%	***	11.251	-6.83%	***			
POT_Pls%	%	81.247	88.155	8.50%	***	74.712	-8.04%	***	78.140	-3.82%	**			
POT_Vst%	%	11.968	11.348	-5.18%	***	10.892	-8.99%	***	11.022	-7.90%	***			
Rep%	%	57.892	57.028	-1.49%	**	64.000	10.55%	****	55.714	-3.76%	**			
Style_Grp%	%	4.673	4.176	-10.64%	****	6.729	44.00%	****	4.844	3.66%	**			
Interaction with other 2 Outcomes	Arrivals	person	141706.542			145609.400	2.75%	**	123536.090	-12.82%	****			
	Avg_stay	day	10.469	10.144	-3.10%	**			10.577	1.03%	**			
	Exp_pp/D	\$	197.117	192.150	-2.52%	**	196.274	-0.43%	*					

Number of * indicates the absolute value of change measured by percentage (CP):
 * 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP ≥ 10%

Table 5.15 Posterior Probability distribution of Japan_Supervised_Training Data Set

Target Node			Arrivals (person)			Avg_Stay (day)			Exp_pp/D (\$)		
Target State			≥ 146305			≥ 6.2			≥ 314.8		
Prior Status			Posterior Influence								
Month			Aug(100%)			Jun(16.67%), Jul (33.33%), Aug(33.33%), Sep(16.67%)			Jan(33.33%), Jun(16.67%), Oct(16.67%), Nov(16.67%), Dec(16.67%)		
Variable Name	Unit	Prior Mean	Posterior Mean	Change		Posterior Mean	Change		Posterior Mean	Change	
Stay_Hotel%	%	87.337	82.864	-5.12%	***	85.148	-2.51%	**	88.754	1.62%	**
POT_Mtg%	%	4.801	3.193	-33.49%	****	4.661	-2.92%	**	4.637	-3.42%	**
POT_Vst%	%	1.703	1.816	6.64%	***	1.640	-3.70%	**	1.733	1.76%	**
Rep%	%	58.567	71.800	22.59%	****	63.147	7.82%	***	57.097	-2.51%	**
Style_Grp%	%	26.966	21.032	-22.01%	****	24.697	-8.41%	***	28.831	6.92%	***
Style_Pkg%	%	73.745	65.931	-10.60%	****	70.481	-4.43%	**	75.395	2.24%	**
Interaction with other 2 Outcomes	Arrivals	person	124488.083			139485.368	12.05%	****	121166.041	-2.67%	**
	Avg_stay	day	5.951	6.290	5.70%	***			5.891	-1.01%	**
	Exp_pp/D	\$	294.442	261.029	-11.35%	****	285.777	-2.94%	**		

Number of * indicates the absolute value of change measured by percentage (CP):

* 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP ≥ 10%

Table 5.16 Posterior Probability distribution of Canada_Supervised_Training Data Set

Target Node		Arrivals (person)				Avg_Stay (days)				Exp_pp/D (\$)			
Target State		≥ 56276				≥ 13.6				≥ 165.9			
Posterior Influence													
Month			Dec(25%), Jan(25%), Feb(25%), Mar(25%)				Dec(25%), Jan (50%), Feb(25%)				Jan(33.33%), Feb(16.67%), Jun(16.67%), Sep((16.67%), Nov((16.67%)		
Variable Name	Unit	Prior Mean	Posterior Mean	Change		Posterior Mean	Change		Posterior Mean	Change			
Stay_Hotel%	%	48.714	45.278	-7.05%	***	45.956	-5.66%	***	48.091	-1.28%	**		
Stay_B&B%	%	1.424	1.354	-4.92%	**	1.354	-4.92%	**	1.561	9.62%	***		
POT_Pls%	%	91.740	94.813	3.35%	**	94.565	3.08%	**	91.656	-0.09%	*		
Rep%	%	62.063	68.484	10.35%	****	68.484	10.35%	****	63.986	3.10%	**		
Style_Pkg%	%	24.708	23.600	-4.48%	**	23.524	-4.79%	**	24.001	-2.86%	**		
Interaction with other 2 Outcomes	Arrivals	person	42069.792			68163.750	62.03%	****	47805.817	13.63%	****		
	Avg_stay	day	12.481	13.626	9.17%	***			12.916	3.49%	**		
	Exp_pp/D	\$	158.654	157.970	-0.43%	*	160.486	1.15%	**				

Number of * indicates the absolute value of change measured by percentage (CP):
 * 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP ≥ 10%

To test these results above against the testing data set, the changes of factors when the outcome factor is set to the target state were calculated as shown in Table 5.17 to Table 5.20.

Table 5.17 Target Node/State Influences of U.S. West_Testing Data Set

Target Node		Arrivals (person)				Avg_Stay (days)			Exp_pp/D (\$)			
Target State		≥ 278661				≥ 10.2			≥ 148.9			
Posterior Influence												
Month			Jul(50%), Aug(50%)				Jan (66.67%), Dec (33.33%)			Dec(33.33%), Feb(33.33%), Mar(33.33%)		
Variable Name	Unit	Prior Mean	Posterior Mean	Change		Posterior Mean	Change		Posterior Mean	Change		
Stay_Hotel%	%	49.946	51.689	3.49%	**	48.437	-3.02%	**	48.251	-3.39%	*	
Stay_B&B%	%	0.862	0.613	-28.83%	****	0.951	10.31%	****	1.105	28.18%	****	
POT_Pls%	%	82.438	84.912	3.00%	**	81.151	-1.56%	**	78.371	-4.93%	**	
POT_Mtg%	%	4.409	2.648	-39.94%	****	4.473	1.46%	**	6.756	53.24%	****	
POT_Vst%	%	11.414	10.955	-4.03%	***	12.557	10.01%	****	11.540	1.10%	**	
Rep%	%	81.483	80.400	-1.33%	**	83.550	2.54%	**	83.400	2.35%	**	
Interaction with other 2 Outcomes	Arrivals	person	249561.000				240823.500	-3.50%	**	215794.000	-13.53%	****
	Avg_stay	day	9.620	9.420	-2.08%	**				10.980	14.14%	****
	Exp_pp/D	\$	145.125	141.950	-2.19%	**	148.900	2.60%	**			

Number of * indicates the absolute value of change measured by percentage (CP):
 * 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP ≥ 10%

Table 5.18 Target Node/State Influences of U.S. East_Testing Data Set

Target Node			Arrivals (person)				Avg_Stay (days)			Exp_pp/D (\$)		
Target State			≥ 152735				≥ 11.8			≥ 194.6		
Posterior Influence												
Month			Mar(33.33%), Jun(33.33%), Jul(33.33%)				Jan (100%)			Oct(100%)		
Variable Name	Unit	Prior Mean	Posterior Mean	Change		Posterior Mean	Change		Posterior Mean	Change		
Stay_B&B%	%	1.247	1.114	-10.71%	****	1.524	22.19%	****	1.250	0.24%	*	
Stay_F&R%	%	11.831	12.172	2.88%	**	11.746	-0.72%	*	9.363	-20.86%	****	
POT_Pls%	%	71.044	56.200	-20.89%	****	75.065	5.66%	**	73.599	3.60%	**	
POT_Vst%	%	11.844	12.283	3.71%	**	10.503	-11.31%	****	9.773	-17.48%	****	
Rep%	%	58.167	57.633	-0.92%	*	65.200	12.09%	****	55.000	-5.44%	***	
Style_Grp%	%	4.837	4.512	-6.72%	***	5.478	13.24%	****	5.878	21.52%	****	
Interaction with other 2 Outcomes	Arrivals	person	136856.667				144153.000	5.33%	***	120533.000	-11.93%	****
	Avg_stay	day	10.425	10.253	-1.65%	**				9.780	-6.19%	***
	Exp_pp/D	\$	182.767	174.500	-2.52%	**	185.200	1.33%	**			

Number of * indicates the absolute value of change measured by percentage (CP):

* 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP≥10%

Table 5.19 Target Node/State Influences of Japan_Testing Data Set

Target Node			Arrivals (person)				Avg_Stay (day)			Exp_pp/D (\$)		
Target State			≥ 120952				≥ 6.3			≥ 310.8		
Prior Status			Posterior Influence									
Month			Aug(50%), Sep(50%)			Jul (50%), Aug(50%)			Oct(50%), Nov(50%)			
Variable Name	Unit	Prior Mean	Posterior Mean	Change		Posterior Mean	Change		Posterior Mean	Change		
Stay_Hotel%	%	86.603	83.967	-3.04%	***	82.532	-4.70%	**	87.89183733	1.49%	**	
POT_Mtg%	%	3.570	2.277	-36.22%	****	2.458	-31.16%	****	2.516687863	-29.51%	****	
POT_Vst%	%	1.935	1.750	-9.59%	***	1.821	-5.89%	***	1.75641808	-9.24%	***	
Rep%	%	58.342	64.950	11.33%	****	68.950	18.18%	****	54.6	-6.41%	***	
Style_Grp%	%	27.139	21.379	-21.22%	****	22.262	-17.97%	****	24.43908414	-9.95%	***	
Style_Pkg%	%	72.889	67.069	-7.99%	***	67.046	-8.02%	***	75.73463552	3.90%	**	
Interaction with other 2 Outcomes	Arrivals	person	102741.167				124714.500	21.39%	****	104769.500	1.97%	**
	Avg_stay	day	6.009	6.270	4.34%	**				5.735	-4.56%	**
	Exp_pp/D	\$	290.192	283.400	-2.34%	****	264.550	-8.84%	***			

Number of * indicates the absolute value of change measured by percentage (CP):

* 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP≥10%

Table 5.20 Target Node/State Influences of Canada_Testing Data Set

Target Node			Arrivals (person)				Avg_Stay (days)			Exp_pp/D (\$)		
Target State			≥ 53638				≥ 13.4			≥ 160.0		
Posterior Influence												
Month			Dec(25%), Jan(25%), Feb(25%), Mar(25%)				Dec(50%), Jan (50%)			Jan(33.33%), Feb(33.33%), May(33.33%)		
Variable Name	Unit	Prior Mean	Posterior Mean	Change		Posterior Mean	Change		Posterior Mean	Change		
Stay_Hotel%	%	52.487	47.695	-9.13%	***	44.820	-14.61%	****	52.722	0.45%	*	
Stay_B&B%	%	1.412	1.276	-9.65%	***	1.199	-15.11%	****	1.608	13.86%	****	
POT_Pls%	%	91.992	93.937	2.11%	**	93.753	1.91%	**	91.086	-0.98%	*	
Rep%	%	60.483	66.600	10.11%	****	67.350	11.35%	****	62.400	3.17%	**	
Style_Pkg%	%	28.143	25.424	-9.66%	***	23.557	-16.29%	****	26.411	-6.15%	***	
Interaction with other 2 Outcomes	Arrivals	person	39796.917			62109.000	56.06%	****	46987.333	18.07%	****	
	Avg_stay	day	12.315	13.343	8.34%	***			12.707	3.18%	**	
	Exp_pp/D	\$	151.200	151.975	0.51%	*	150.150	-0.69%	*			
Number of * indicates the absolute value of change measured by percentage (CP):												
* 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP≥10%												

Comparing the CP values resulted from the testing data set and the values from the Bayesian network training by the training data set, the validation results for each MMA are shown in Table 5.21 to Table 5.24.

Among all the factors involved here, Month is the only discrete variable. The training data set shows a wider posterior probability distribution of months compared to the testing data set. This is largely due to the fact that there are 2 years'

data used for training, but only one for testing. Indeed for all the outcomes, the resulting months in testing data set are a subset of the training data set's results. For the continuous variables: First, look at the direction of the changes, indicated by positive ("+") or negative ("-"), the error rate is 13.54% (13 errors out of 96 predictions: 5 for U.S. West, 3 for U.S. East, 2 for Japan, and 3 for Canada).

Then, in terms of accuracy, this study does not provide a function to sum up the variances of all the variables or for all the outcomes because of the different scales of factors, and the absence of information to attach weights to them. Also as discussed at the beginning of this section, the cross validation is not used to compare the variances and select a best model, but to prevent overfitting. As a reference, this research used the percentage of change (CP) to scale the level of changes. In general, analysts would pay more attention to the more significant influences. By comparing the variables with $CP \geq 10\%$ between the training and testing results, the overall error rate is 20.83% (20 errors out of 96 predictions: 5 for U.S. West, 6 for U.S. East, 5 for Japan, and 4 for Canada).

Table 5.21 Supervised Model Validation of U.S. West

US_West		Training		Testing		
Target Node		Arrivals				
Target State		≥ 292450		≥ 278661		
Month		Mar(16.67%), Jun(16.67%), Jul(33.33%), Aug(33.33%)		Jul(50%), Aug(50%)		
Variable Name	Unit	Change		Change		
Stay_Hotel%	%	1.66%	**	3.49%	**	
Stay_B&B%	%	-17.61%	****	-28.83%	****	
POT_Pls%	%	2.87%	**	3.00%	**	
POT_Vst%	%	-31.75%	****	-39.94%	****	
Rep%	%	-7.15%	***	-4.03%	***	
Style_Pkg%	%	-1.10%	**	-1.33%	**	
Interaction with other 2 Outcomes	Arrivals	person				
	Avg_stay	day	-1.53%	**	-2.08%	**
	Exp_pp/D	\$	0.05%	*	-2.19%	**
Target Node		Avg_Stay				
Target State		≥ 10.3		≥ 10.2		
Month		Jan (66.67%), Dec (33.33%)		Jan (100%)		
Variable Name	Unit	Change		Change		
Stay_Hotel%	%	-1.55%	**	-3.02%	**	
Stay_B&B%	%	19.62%	****	10.31%	****	
POT_Pls%	%	-1.79%	**	-1.56%	**	
POT_Vst%	%	25.80%	****	1.46%	**	
Rep%	%	12.97%	****	10.01%	****	
Style_Pkg%	%	2.58%	**	2.54%	**	
Interaction with other 2 Outcomes	Arrivals	person	-9.94%	***	-3.50%	**
	Avg_stay	day				
	Exp_pp/D	\$	-0.79%	*	2.60%	**
Target Node		Exp_pp/D				
Target State		≥ 161.9		≥ 148.9		
Month		Dec(33.33%), Feb(33.33%), Mar(33.33%)		Dec(100%)		
Variable Name	Unit	Change		Change		
Stay_Hotel%	%	-0.87%	*	-3.39%	*	
Stay_B&B%	%	14.05%	****	28.18%	****	
POT_Pls%	%	0.17%	*	-4.93%	**	
POT_Vst%	%	3.03%	**	53.24%	****	
Rep%	%	10.27%	****	1.10%	**	
Style_Pkg%	%	1.75%	**	2.35%	**	
Interaction with other 2 Outcomes	Arrivals	person	1.76%	**	-13.53%	****
	Avg_stay	day	-1.63%	**	14.14%	****
	Exp_pp/D	\$				
Number of * indicates the absolute value of change measured by percentage (C						
* 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP≥10%						

Table 5.22 Supervised Model Validation of U.S. East

US_East		Training		Testing	
Target Node		Arrivals			
Target State		≥ 157961		≥ 152735	
Month		Mar(33.33%), Jun(33.33%), Jul(33.33%)		Mar(33.33%), Jun(33.33%), Jul(33.33%)	
Variable Name	Unit	Change		Change	
Stay_B&B%	%	-12.14%	****	-10.71%	****
Stay_F&R%	%	-6.38%	***	2.88%	**
POT_Pls%	%	8.50%	***	20.89%	****
POT_Vst%	%	-5.18%	***	-3.71%	**
Rep%	%	-1.49%	**	-0.92%	*
Style_Grp%	%	-10.64%	****	-6.72%	***
Interaction with other 2 Outcomes	Arrivals	person			
	Avg_stay	day	-3.10%	**	-1.65%
	Exp_pp/D	\$	-2.52%	**	-2.52%
Target Node		Avg_Stay			
Target State		≥ 11.9		≥ 11.8	
Month		Jan (100%)		Jan (100%)	
Variable Name	Unit	Change		Change	
Stay_B&B%	%	14.81%	****	22.19%	****
Stay_F&R%	%	-6.38%	***	-0.72%	*
POT_Pls%	%	8.04%	***	5.66%	**
POT_Vst%	%	-8.99%	***	-11.31%	****
Rep%	%	10.55%	****	12.09%	****
Style_Grp%	%	44.00%	****	13.24%	****
Interaction with other 2 Outcomes	Arrivals	person	2.75%	**	5.33%
	Avg_stay	day			
	Exp_pp/D	\$	-0.43%	*	1.33%
Target Node		Exp_pp/D			
Target State		≥ 204.0		≥ 194.6	
Month		Jan(14.29%), Apr(14.29%), Aug(14.29%), Sep(28.57%), Oct(28.57%)		Oct(100%)	
Variable Name	Unit	Change		Change	
Stay_B&B%	%	5.50%	***	0.24%	*
Stay_F&R%	%	-6.83%	***	-20.86%	****
POT_Pls%	%	3.82%	**	3.60%	**
POT_Vst%	%	-7.90%	***	-17.48%	****
Rep%	%	-3.76%	**	-5.44%	***
Style_Grp%	%	3.66%	**	21.52%	****
Interaction with other 2 Outcomes	Arrivals	person	-12.82%	****	-11.93%
	Avg_stay	day	1.03%	**	-6.19%
	Exp_pp/D	\$			

Number of * indicates the absolute value of change measured by percentage (CP):

* 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP≥10%

Table 5.23 Supervised Model Validation of Japan

Japan		Training		Testing	
Target Node		Arrivals			
Target State		≥ 146305		≥ 120952	
Month		Aug(100%)		Aug(50%), Sep(50%)	
Variable Name	Unit	Change		Change	
Stay_Hotel%	%	-5.12%	***	-3.04%	***
POT_Mtg%	%	-33.49%	****	-36.22%	****
POT_Vst%	%	6.64%	***	-9.59%	***
Rep%	%	22.59%	****	11.33%	****
Style_Grp%	%	-22.01%	****	-21.22%	****
Style_Pkg%	%	-10.60%	****	-7.99%	***
Interaction with other 2 Outcomes	Arrivals	person			
	Avg_stay	day	5.70%	***	4.34% **
	Exp_pp/D	\$	-11.35%	****	-2.34% ****
Target Node		Avg_Stay			
Target State		≥ 6.2		≥ 6.3	
Month		Jun(16.67%), Jul(33.33%), Aug(33.33%), Sep(16.67%)		Jul (50%), Aug(50%)	
Variable Name	Unit	Change		Change	
Stay_Hotel%	%	-2.51%	**	-4.70%	**
POT_Mtg%	%	-2.92%	**	-31.16%	****
POT_Vst%	%	-3.70%	**	-5.89%	***
Rep%	%	7.82%	***	18.18%	****
Style_Grp%	%	-8.41%	***	-17.97%	****
Style_Pkg%	%	-4.43%	**	-8.02%	***
Interaction with other 2 Outcomes	Arrivals	person	12.05%	****	21.39% ****
	Avg_stay	day			
	Exp_pp/D	\$	-2.94%	**	-8.84% ***
Target Node		Exp_pp/D			
Target State		≥ 314.8		≥ 310.8	
Month		Jan(33.33%), Jun(16.67%), Oct(16.67%), Nov(16.67%), Dec(16.67)		Oct(50%), Nov(50%)	
Variable Name	Unit	Change		Change	
Stay_Hotel%	%	1.62%	**	1.49%	**
POT_Mtg%	%	-3.42%	**	-29.51%	****
POT_Vst%	%	1.76%	**	-9.24%	***
Rep%	%	-2.51%	**	-6.41%	***
Style_Grp%	%	6.92%	***	-9.95%	***
Style_Pkg%	%	2.24%	**	3.90%	**
Interaction with other 2 Outcomes	Arrivals	person	-2.67%	**	1.97% **
	Avg_stay	day	-1.01%	**	-4.56% **
	Exp_pp/D	\$			
Number of * indicates the absolute value of change measured by percentage (C					
* 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP≥10%					

Table 5.24 Supervised Model Validation of Canada

Canada			Training		Testing	
Target Node			Arrivals			
Target State			≥ 56276		≥ 53638	
Month			Dec(25%), Jan(25%), Feb(25%), Mar(25%)		Dec(25%), Jan(25%), Feb(25%), Mar(25%)	
Variable Name	Unit	Change		Change		
Stay_Hotel%	%	-7.05%	***	-9.13%	***	
Stay_B&B%	%	-4.92%	**	-9.65%	***	
POT_Pls%	%	3.35%	**	2.11%	**	
Rep%	%	10.35%	****	10.11%	****	
Style_Pkg%	%	-4.48%	**	-9.66%	***	
Interaction with other 2	Arrivals	person				
Outcomes	Avg_stay	day	9.17%	***	8.34%	***
	Exp_pp/D	\$	-0.43%	*	0.51%	*
Target Node			Avg_Stay			
Target State			≥ 13.6		≥ 13.4	
Month			Dec(25%), Jan (50%), Feb(25%)		Dec(50%), Jan (50%)	
Variable Name	Unit	Change		Change		
Stay_Hotel%	%	-5.66%	***	-14.61%	****	
Stay_B&B%	%	-4.92%	**	-15.11%	****	
POT_Pls%	%	3.08%	**	1.91%	**	
Rep%	%	10.35%	****	11.35%	****	
Style_Pkg%	%	-4.79%	**	-16.29%	****	
Interaction with other 2	Arrivals	person	62.03%	****	56.06%	****
Outcomes	Avg_stay	day				
	Exp_pp/D	\$	1.15%	**	-0.69%	*
Target Node			Exp_pp/D			
Target State			≥ 165.9		≥ 160.0	
Month			Jan(33.33%), Feb(16.67%), Jun(16.67%), Sep((16.67%), Nov((16.67%)		Jan(33.33%), Feb(33.33%), May(33.33%)	
Variable Name	Unit	Change		Change		
Stay_Hotel%	%	-1.28%	**	0.45%	*	
Stay_B&B%	%	9.62%	***	13.86%	****	
POT_Pls%	%	-0.09%	*	-0.98%	*	
Rep%	%	3.10%	**	3.17%	**	
Style_Pkg%	%	-2.86%	**	-6.15%	***	
Interaction with other 2	Arrivals	person	13.63%	****	18.07%	****
Outcomes	Avg_stay	day	3.49%	**	3.18%	**
	Exp_pp/D	\$				

Number of * indicates the absolute value of change measured by percentage (CP):

* 0≤CP<1%; ** 1%≤CP<5%; *** 5%≤CP<10%; **** CP ≥ 10%

5.4.2 Validation with An Additional Data Set

The purpose of testing with an additional data set outside the original data set is to validate the predictive results of the unsupervised model. This data set is not part of the original data set used to train the model and test the hypotheses. In this test, the additional data set includes the monthly visitor highlight data from January, 2014 to July, 2014, collected from the same public data source provided by Hawaii Tourism Authority (Hawaii Tourism Authority, 2014). Supervised model is not tested in this method due to the incomplete set of data, which only includes the first half of the year.

5.4.2.1 Validation of Unsupervised Model with Additional Data Set

Figure 5.5 to Figure 5.7 in Section 5.2.2 showed the association between MMA and each outcome: U.S. West visitors tend to be the highest arrivals group, Canadian visitors tend to stay for the longest period, and Japanese visitors tend to spend most per person per day. Same results were obtained from the additional data set, shown in Figure 5.23 to Figure 5.25.

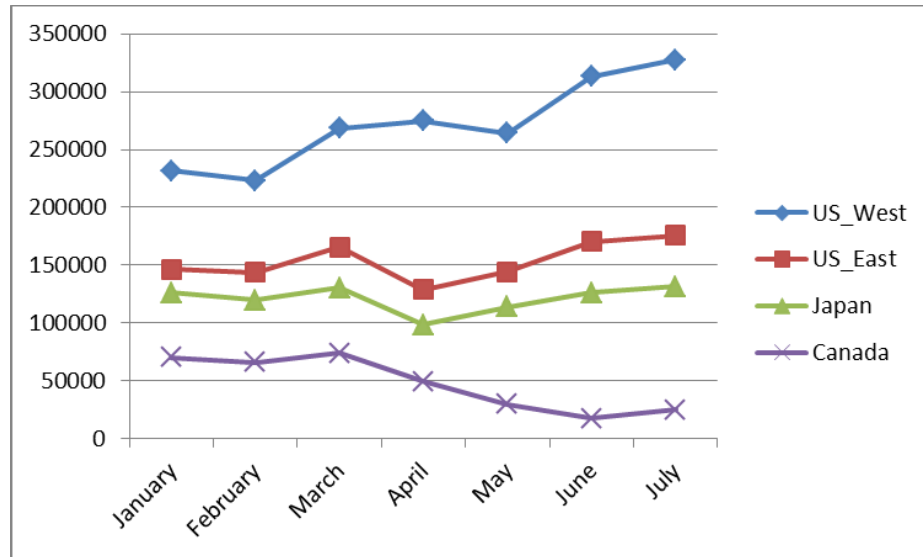


Figure 5.23 Arrivals of 4 Regions_Additional Data Set

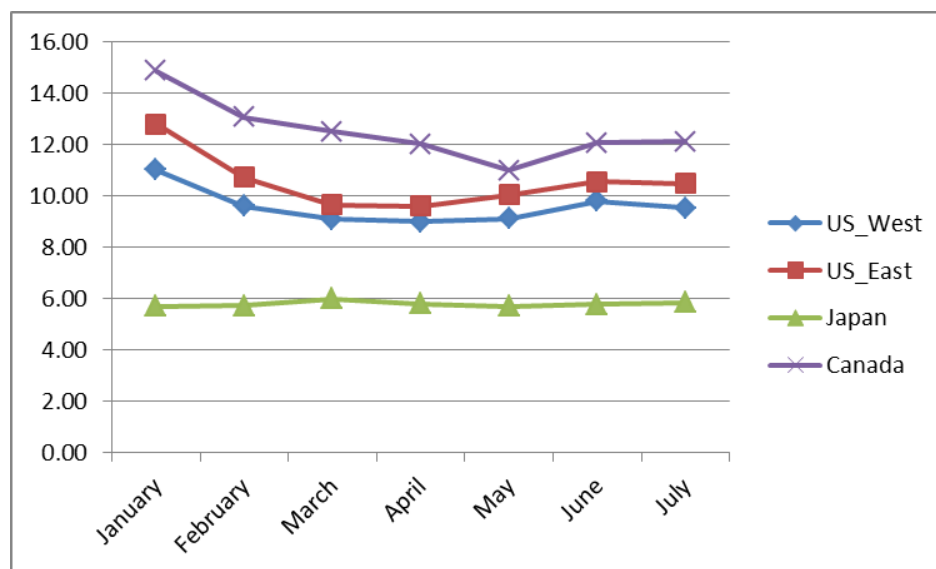


Figure 5.24 Average Lengths of Stay (day) of 4 Regions_Additional Data Set

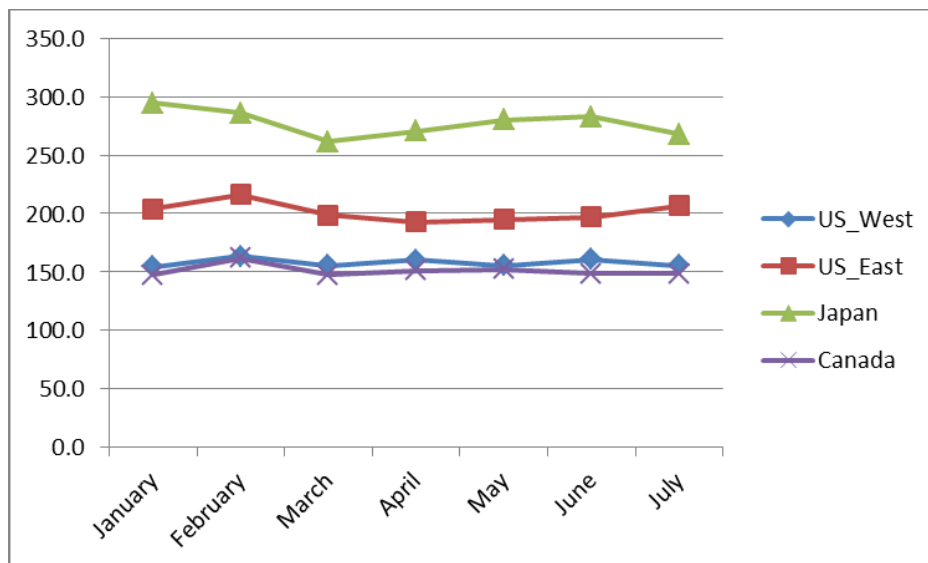


Figure 5.25 Daily Expenditure per Person (\$) of 4 Regions_Additional Data Set

Figure 5.8 showed a contrast between Japanese visitors' consuming behavior against U.S. West visitors by setting the prior probability of MMA to Japan and U.S. West and observing the influences on the other factors. The posterior inference results showed that Japanese visitors tend to stay at hotel most and stay with friends and relatives least, and they tend to take group trips and package trips; while visitors from U.S. West showed an opposite pattern. This could also be verified by the additional data set, showed in Figure 5.26.

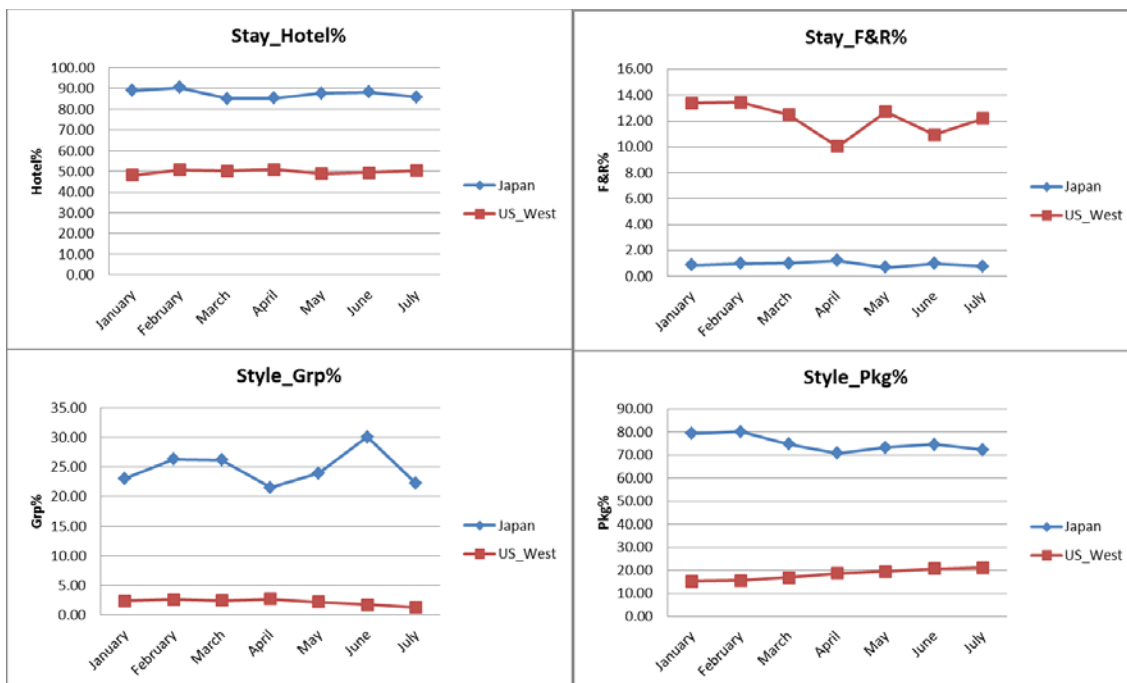


Figure 5.26 Visitors Characteristics: Japan v.s. U.S. West_Additional Data Set

5.4.3 Validation Summary

In summary, through cross validation, the classification results of the unsupervised Bayesian network proved to be consistent with the testing data set, the supervised Bayesian network had an error rate of 13.54% in predicting the trend of influences, and an error rate of 20.83% in predicting the influences with a percentage of change equal to or larger than 10%. Through validation with an unseen additional data set, the classification results of the unsupervised Bayesian network was validated to be accurate. Due to the lack of benchmark data for comparison reference, the accuracy of the supervised models can't be disclaimed. But overall, no overfitting was observed in the models resulted from the proposed research approach.

During the validation, while processing the testing data set in Excel spreadsheet, the author estimated that the time needed for the same analysis using Excel spreadsheet is 2 to 3 times of the time needed when using BayesiaLab. Yet it was just for the results validation. If Excel spreadsheet was used to analyze the unknown system from scratch, it would take significantly more time (more than double) to analyze the relationship of any two factors to understand where the valuable information exists. In addition, as discussed in Chapter 2 and 3, Excel spreadsheet is not capable of multivariate analysis involving a relationship network. It does not have the capability to see the relationships among more than three factors at one glance and in such an intuitive and graphical way.

5.5 Results Summary

In the previous sections of Chapter 5, the results of each step was presented at the end of the section. It was structured this way to help the reader understand the approach. In this section, the results are summarized into the key findings below.

Abbreviations of the variables will be mentioned frequently in this section. To ease the readers in understanding, the list of variables and definition from Chapter 4 is presented here again.

Table 5.25 Complete List of Variables and Definition

Type	Name	Definition	Unit
Outcome	Arrivals	Number of visitors arriving in Hawaii	person
	Avg_Stay	Average length of stay by days	day
	Exp_pp/D	Per person per day spending by USD	USD
Predictor	MMA	Major Market Area, the original country/region visitors came from by air	
	Month	The month when the data were collected	
	Stay_Hotel%	Percentage of visitors who plan to stay at hotel during their stay in Hawaii (including hotel only and hotel + other accommodations)	
	Stay_B&B%	Percentage of visitors who plan to stay at Bed & Breakfast during their stay in Hawaii	
	Stay_F&R%	Percentage of visitors who plan to stay with Friends/Relatives during their stay in Hawaii	
	POT_Pls%	Percentage of visitors whose Purpose of Travel was Pleasure, including Pleasure/Vacation, Wedding and Honeymoon.	
	POT_Mtg%	Percentage of visitors whose Purpose of Travel was Corporate Meeting, Convention or Incentive	
	POT_Vst%	Percentage of visitors whose Purpose of Travel was to Visit Friends or Relatives.	
	Rep%	Percentage of Repeaters whose recorded visits were not their first trips to Hawaii.	
	Style_Grp%	Percentage of visitors who traveled with a group.	
	Style_Pkg%	Percentage of visitors who traveled on a purchased package trip.	

5.5.1 MMA

MMA is an effect modifier for the Hawaii tourism market. It has strong relationships with both the outcomes and the visitor characteristics. Knowing the origin of a visitor will help predict his or her purpose of travel, choice of

accommodation, travel styles (package trip, group trip), traveling season preference, and the possible range of length of stay, daily expenditures, and the overall volume of visitors arriving in Hawaii from this region.

Posterior probability distribution of the unsupervised model shown in Figure 5.5 to Figure 5.7 shows: The highest volume of visitors are more likely to come from U.S. West, and least likely from Canada. Japanese visitors tend to spend much more than people from the other regions, with its mean value more than 50% higher than the second highest region U.S. East (\$292.0 v.s. \$192.3). But visitors from Japan tend to stay for the shortest period, 6 days on average, while Canadian visitors are likely to stay for the longest, averaging 12.4 days.

5.5.2 Travelling Season

Table 5.1 shows that, for the entire body of visitors from the 4 MMAs, Month has little influence on the outcomes. But when separated by MMA, Months stands out as a strong influencer. Visitors from different regions show different preferences in travel months: People from U.S. mainland (West and East) tend to visit Hawaii in summer months like June, July and August, while Japanese visitors are more likely to travel to Hawaii in August and the following months through December, and Canadian visitors prefer winter months from December to March.

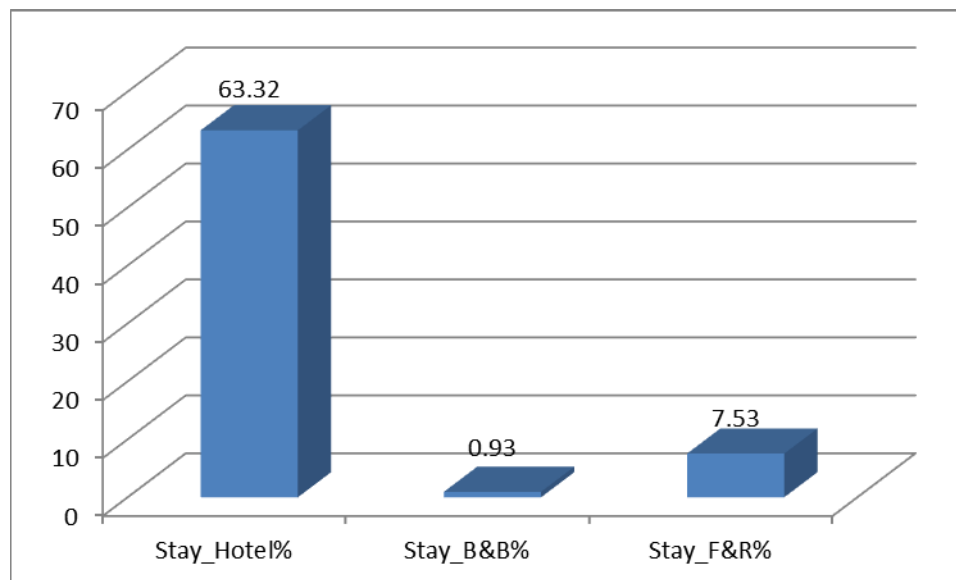
Visitors from specific regions also show certain seasonal patterns in terms of average lengths of stay and expenditure during their stay in Hawaii. Domestic visitors have a strong tendency to stay for longer during December and January.

A similar trend is observed on visitor from Canada, with the peak of lengths of stay in January, along with December and February. Japanese visitors have a different pattern: Their lengths of stay tend to reach the high in August (the same month of highest regional visitor volume), accompanied by the neighboring months July and September.

In terms of daily expenditure per person, visitors from U.S. West and U.S. East are both likely to spend more in September, but the western visitors are mostly likely to have the highest level of expenditure in March, while for eastern visitors it is November. Japanese visitors and Canadian visitors both tend to spend more during the winter months (October to February) and June.

5.5.3 Choice of Accommodation

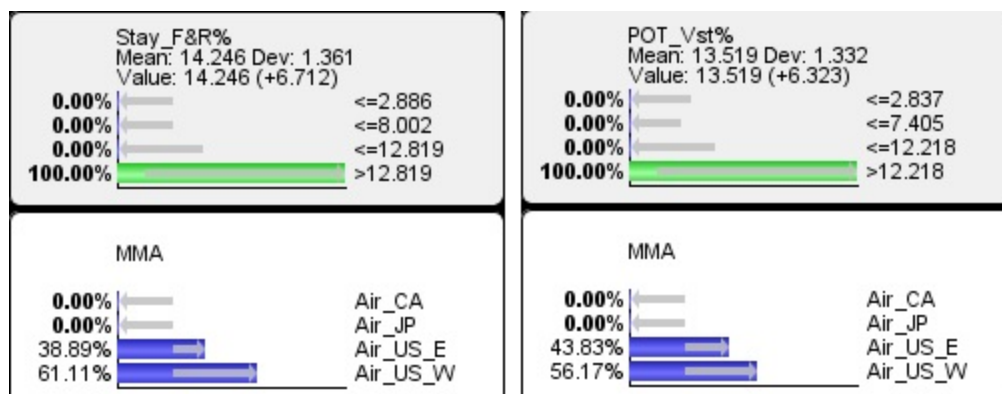
Three types of accommodation choice were included in this analysis: hotel, B&B and the home of friends/relatives. Overall, hotel is the top choice. Figure 5.27 is based on the mean values of the percentages of visitors choosing a certain accommodation type in each MMA from the monthly visitor highlight reports.



- * Stay_Hotel%: Percentage of visitors who plan to stay at Hotel during their stay in Hawaii
- Stay_B&B%: Percentage of visitors who plan to stay at Bed & Breakfast during their stay in Hawaii
- Stay_F&R%: Percentage of visitors who plan to stay with Friends/Relatives during their stay in Hawaii

Figure 5.27 Choices of Accommodation by Percentage

The last type was not included in the final supervised model due to its relatively weaker relationships with the outcomes. But Figure 5.3 shows that it is closely related to MMA and POT_Vst%, and Figure 5.8 shows that visitors from U.S. West are much more likely to stay with relatives and friends than visitors from Japan. In fact, U.S. domestic visitors are more likely to stay with family and friends than foreign visitors. It can be illustrated by the posterior inference shown in Figure 5.28.



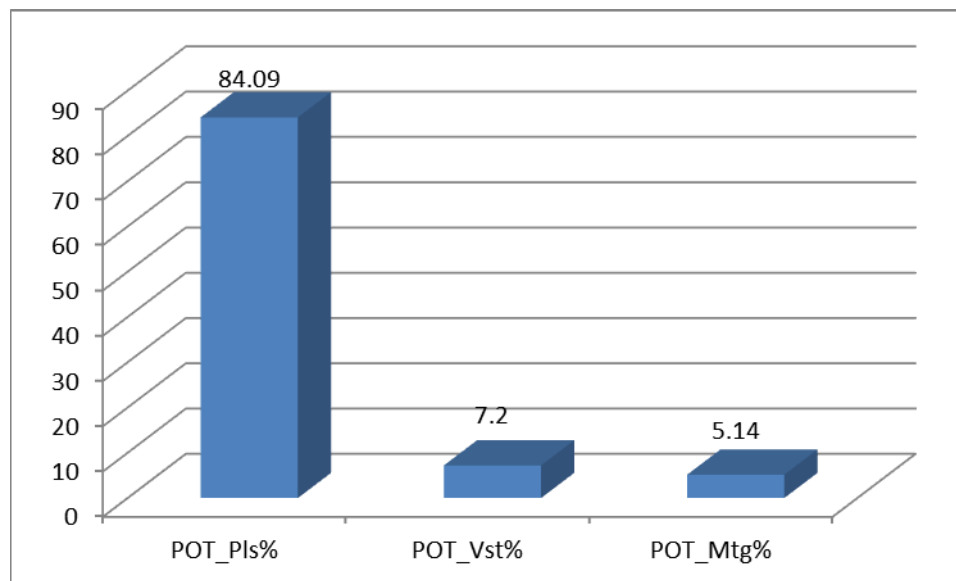
* Stay_F&R%: Percentage of visitors who plan to stay with Friends/Relatives during their stay in Hawaii
 POT_Vst%: Percentage of visitors whose Purpose of Travel was to Visit Friends or Relatives

Figure 5.28 MMA Likelihood Given High Stay_F&R% and POT_Vst%

For the other 2 types of accommodation, domestic visitors show an association between lower percentage of people choosing B&B and high arrivals, but a positive correlation with average lengths of stay and personal daily expenditure. Visitors from Canada share the same pattern except for average lengths of stay. Visitors from U.S. East, Japan and Canada all share the commonality that higher arrivals and longer lengths of stay tend to indicate smaller percentages of visitors staying at hotel. This association is especially strong for Canadian visitors.

5.5.4 Purpose of Travel

Three types of purpose of travel were included in this analysis: for pleasure (including pleasure/vacation, wedding and honeymoon), for corporate meeting, convention or incentive, and for visiting friends or relatives. Overall, pleasure is the major motivation for Hawaii visitors. Figure 5.29 is based on the mean values of the percentages of visitors with a certain purpose of travel in each MMA from the monthly visitor highlight reports.



* POT_Pls%: Percentage of visitors whose Purpose of Travel was Pleasure
 POT_Vst%: Percentage of visitors whose Purpose of Travel was to Visit Friends or Relatives
 POT_Mtg%: Percentage of visitors whose Purpose of Travel was Corporate Meeting, Convention or Incentive

Figure 5.29 Purpose of Travel by Percentage

For all the regions, higher percentage of visitors travelling for pleasure is found to be related to high arrivals. But for U.S. domestic visitors, when the lengths of stay is high, visitors travelling for pleasure tend to take a smaller portion, but it's the opposite for overseas visitors.

In addition to its relationship with the accommodation choice of staying with friends and relatives, POT_Vst% is found to be negatively correlated with arrivals and lengths of stay for both U.S. West and Japan visitors. But these 2 MMAs differ in the direction of association between POT_Vst% and daily expenditure per person: positive for U.S. West and negative for Japan.

Travelling for meeting, convention or incentive is not a significant factor for Japanese visitors. For the other 3 MMAs, a lower percentage of people who travel for this purpose is related to high arrivals, but high lengths of stay is related to higher percentages of meeting/convention/incentive travellers. For U.S. domestic visitors, especially visitors from U.S. East, the posterior influence is significant in POT_Mtg% is dramatic.

5.5.5 Repeat Visitor

The overall average percentage value of repeat visitors arriving in Hawaii by air from the 4 MMAs is 64.87%, with the highest from U.S. West (81.47%) and lowest in U.S. East (57.98%). It is also a significant factor in the final supervised model for all regions. All 4 MMAs show that long lengths of stay is related to higher percentages of repeat visitors, especially for Japan. But when it comes to arrivals, the posterior inference shows a split: for domestic visitors, high volumes of arrivals indicate a slightly lower percentage of repeat visitors, while for overseas visitors, this means the percentage is likely to increase by about 10%.

5.5.6 Travel Style

In this analysis, “travel style” includes 2 factors: whether or not to purchase a package trip, and whether or not to travel with a group. They are not exclusive of each other. A visitor can choose to travel with an agency on a package trip. The cross-MMA mean value of Style_Grp% is 9.05%, and for Style_Pkg% it is 36.04%.

For all the regions, traveling on a package trip is associated with shorter lengths of stay, and for Japanese and Canadian visitors, this also tend to be connected with lower arrivals – but this is different for U.S. West. Visitors from U.S. East and Japan both show that when the arrivals are high, the percentages of group travellers tend to be lower.

CHAPTER 6. CONCLUSION

This chapter will review the hypotheses brought up at the beginning, and link the results to them. Consequently, this study will be concluded.

6.1 Hypothesis Validation

Based on the analysis results, the two hypotheses proposed in Section 1.5 can be validated:

1. Aggregate data can be used as input to Bayesian networks to analyze complex system and provide valuable insights on the relationships among multiple factors.

Validation: This analysis used data aggregated from individual visitor information, presented as a group sample of visitors from a specific region to Hawaii in each month. Starting from raw data without prior knowledge or experience of the system, following the analysis procedures, this study has revealed new knowledge of practical values. The approach developed in this analysis can be extended to applications in other domains.

2. In the travel and tourism section, visitors from different regions have different behaviors which will affect the outcomes evaluated by measurable metrics, such as arrivals, length of stay, expenditure.

Validation: This hypothesis has been well proved throughout the analysis and results summary. MMA is an effect modifier for the Hawaii tourism market with strong influences on both the outcomes and the visitor characteristics. When separated by MMA, the characteristics of each regional visitors group and their interactions were revealed. Without realizing the significance of visitor original region, the analysis could be much less meaningful, and even misleading. For example, Month did not stand out as a significant factor except in regional analysis.

6.2 Conclusions

Through a hybrid research approach with unsupervised and supervised modeling using Bayesian networks, analysis with aggregate data produced valuable findings on the omnidirectional relationships in a multi-factor consumer service system. The approach used in this study provided an opportunity to get information with aggregate data, which are usually already available, or can be easily obtained without conducting additional survey on individuals, and the findings are directly linked to DMO and service providers' decision-making and interests. The data visualization feature of Bayesian network enabled an intuitive presentation of the results. The analysis of Hawaii tourism market confirmed that original region is the most information-rich factor in the network. Knowing visitors' origin can significantly reduce the uncertainties of their behavior and the outcomes of the service supply chain. The awareness of the influences of the

regional factor justifies conducting consumer research by region, which reveals more meaningful and accurate knowledge than region-blind analysis.

CHAPTER 7. DISCUSSION

Today's consumer industry is among the most data-driven businesses. As most organizations recognize that being a successful, data-driven company requires skilled developers and analysts, fewer grasp how to use data to tell a meaningful story (Waisberg, 2014). When the analysis uncovers hidden unknown connections in a network, the story telling becomes more interesting, and requires more skills and theoretical support. The directed arcs themselves in a learning Bayesian network are no more than statistical relationships without interpretation in an application environment. This chapter will interpret the relationships from the results in Section 5.4 and verify the causalities, in order to tell the stories in Hawaii tourism market.

Although the desired target of each outcome is the maximum value, they are not necessarily positively correlated, and the correlation differs from one region to another. The same factor can have positive influence on one outcome and negative impact on another. For example, for visitors from U.S. West, when there are more visitors choose to stay at B&B, it's more likely that it's a time when the visitors numbers are high, but their length of stay is shorter.

7.1 Visitor Origin

It is not surprising that visitors' origins play such a big role in their consuming behavior. In the globalized consumer market, national and regional cultures' influences on consumer behavior have been widely recognized (De Mooij, 2010; Gopaldas & Fischer, 2012; Kacen & Lee, 2002; Luna & Gupta, 2001; Singh & Appiah-Adu, 2008). In the tourism industry, tourist's decision-making and demand pattern, how the tourist is influenced by relation groups, the tourist's buying roles and preferences and perception of purchase and travel risk, and emotions and feelings leading to the tourist's experience and level of satisfaction are all related to the national or regional culture (Reisinger, 2009b).

As one of the key results of the analysis of Hawaii tourism market, MMA has strong influence on visitor arrivals, average length of stay and daily expenditure per person. Visitor arrivals is the direct result of the choice of destination, which is usually the very foremost decision made for a trip. The most direct factor could be travel distance. Global transportation has made it technically possible too travel to almost every spot of the world. But in practice, travel distances can affect the decision making or destination and travel pattern in many ways. In such cases, they are often represented as the perception of the distance to a destination, rather than purely physical distance. However, past studies have shown that tourists' cognitive perceptions of the distance to destinations are often highly inaccurate and that this inaccuracy is not necessarily related to actual

distance, but rather more directly related to perceptions of cost of travelling to the destination (Harrison-Hill, 2000).

Economic factors like travel costs (e.g. airfare, luggage fee), considerations over convenience and comfort including the needs for passport, Customs check, flight transfer, jet lag, language, currency and culture all contribute to the barriers of travelling to destinations far from home. They also establish an emotional distance perception of the destination, which would in turn further intensify the factual considerations.

On the other side, in the tourism industry, unknown and unfamiliarity are often the motivation of travel. A word often used as the synonym of vacation trip is “escape”. It perfectly tells the expected characteristics of such a trip: new, unknown, far away from the daily norms, and enchanting. These factors are known as “pull” and “push” factors in tourist motivations. The push factors for a vacation are socio-psychological motives. The pull factors are motives aroused by the destination rather than emerging exclusively from within the traveler himself, also termed “cultural” (Crompton, 1979). Literature has identified that whether it is labeled enchantment, novelty, luxury or far-off allure, the assertion is that the attractiveness of a destination increases with distance for some travelers (Harrison-Hill, 2000).

Considering physical distance, U.S. West is the closest to Hawaii, followed by Japan. The “pull” factors of Japanese visitors towards Hawaii are probably

stronger than domestic U.S. visitors, given the nature of international travel and cultural differences. Based on literature, Japanese put a great emphasis on the group, the family, and belonging and loyalty. When on vacation, Japanese tourists are activity-oriented unlike the Western tourists who travel to do nothing. Shopping is very important to them (Reisinger, 2009a). From the model analysis, visitors from Japan are more likely to travel on a group tour and have higher individual daily expenditure. The statistical relationships have found theoretical and empirical support.

Another aspect of the influences of origin is demonstrated by the accommodation choices. Compared to visitors from U.S. West, the percentage of Japanese visitors who tend to stay with relatives and friends is lower. According to the 2008-2012 American Community Survey 5-Year Estimates (U.S. Census Bureau, 2013), in the total population of 1,362,730 in Hawaii, 186,988 people reported their race as Japanese (13.72% of the state population), while the population of White Americans, Black or African Americans, and American Indians and Alaska Natives totaled 399,194 (29.29% of the state population) . These numbers don't directly translate to the amounts of relatives and friends that domestic visitors and Japanese visitors have in Hawaii. But they support the deduction that Japanese visitors may not have as many relatives and friends to stay with.

7.2 Purpose of Travel

Although not shown with the strongest relationships, during the modeling analysis, the purpose of travel has been noticed with ties to origin, as well as with travel patterns such as group or package tour.

By the definition of the Hawaii Tourism Authority, Pleasure as POT includes Pleasure/Vacation, Honeymoon and Get Married – with the common intention for the visitor to be pleased. For vacation-oriented visitors, they can enjoy the pleasure by traveling independently or with a group, and each type of tour mode has its unique attributes to satisfy tourists' special needs. However, tourists have various needs, and they need to choose a tour mode that can satisfy the most of their needs in order to maximize their satisfaction (U, 2007).

While independent travelers enjoy the fun from Do-It-Myself and independence, people who prefer a package group trip may have different reasons. For some package tour tourists, especially for those tourists who enjoy being served and escorted during the tour, travel in a comfortable and convenient way can allow them to enjoy the pleasure tour more and have a safer tour overseas (U, 2007). These needs match the characteristics of Japanese tourists, to whom trust and relationship-building are vital, and a high standard of services is critical to their satisfaction (Reisinger, 2009a). And their preference of shopping for gift-giving during trips contributes to the higher expenditure. Besides, for people who go for honeymoon or wedding, it is reasonable that they are prepared to spend more on the significant event of their life.

For visitors whose POT is to visit relatives and friend (VFR), it's plausible that these visitors are likely to stay at the homes of people they visit, which consequently reduces the likelihood to choose hotel or B&B, and to arrange a group or package trip. Past studies also show that some differentiating features of VFR visitors include most often travelling with children and fewest adults, using more public transport, and spending less than the total tourist body (Seaton & Palmer, 1997).

But it must be noted that although the data came from the real market, the observations were based on probabilistic theories, algorithms and inference. As mentioned in previous sections, the arcs in the learned Bayesian model is more of a statistical relationship rather than causal relationship. Further research and analysis are needed to verify the reasoning behind the arcs.

From the discussion above, the findings related to Purpose of Travel found explanation to support the statistic relationships to become causal relationships.

The analysis of visitor origins and purpose of travel demonstrates the opportunities brought up from the Bayesian network model. The findings and knowledge resulted from the networks help filter out noises and insignificant factors, and inspire further studies in a more focused and oriented manner.

Many meaningful action plans can be developed from the mined knowledge. An example is for "firsttimers". Knowing that there are likely to be more first-time

visitors from Japan and U.S. East, this is a chance to impress them with outstanding service, so that these visitors will become repeaters. Since Japanese visitors tend to take package trips with a group, travel agencies, airlines, hotel and restaurants can develop package products that are customized for the Japan market. Hiring Japanese-speaking staff and adding Japanese language menus, labels, greetings or instructions are a few other examples.

7.3 Consideration of Cross Validation

Section 5.4 presented the cross validation method and the results, and talked about some restrictions and differences from the classic cross validation technique. In this section, more will be discussed.

7.3.1 Sample Size

The original data set includes 36 months for 4 MMAs, 144 instances. But to avoid biased data selection due to missing seasonal pattern, the 144 instances were grouped into 3 calendar years. Essentially, in the analysis for each MMA, it is to use 1 year's data to test the model trained with 2 years' data. The sample size is too small to average out the year-to-year fluctuation. Although the available data set was able to test the unsupervised model and rule out the risk of overfitting, validation of the prediction results in the supervised model need to be improved with more data.

7.3.2 Consideration of Outlier

At the beginning of Section 5.4.1, the data set was checked against outlier to validate that the data set meets the pre-requisite of running cross validation: all the data came from the same population. In practical application, it is recommended to take an initial check to detect any possible outlier. Knowledge of the existence of unusual events that caused outstanding changes in the tourism market performance in a certain period of the year should be taken into consideration. Outliers should be made aware of and excluded from the analysis.

7.4 Limitations and Future Work

There are several limitations of this study that need to be considered or addressed in the future:

1. Tool limitation: The study began with a free trial version of the software in which some features were limited. For example, the number of nodes in a model is limited to 10. However, it was compensated by the carefully thought factor ranking and screening method. The author argues that this method does not only serve the purpose of reducing the variable list, but also provide additional thinking and observations through the analysis procedure. In fact, prioritizing significant factors based on mutual information and research interest has been demonstrated and recommended in relationship analysis of systems with a number of factors (Conrady & Jouffe, 2013a, 2013c). In addition, an elastic priced license

was purchased and used to validate the final models. Same results were received to validate the hypothesis and support the key findings.

2. In-depth analysis needed: This is an exploratory study for a non-specialist to get an initial picture of the issues. Some results of the study are good as guidance for making policies and strategies, but not accurate enough to achieve delicate plans. This study does serve as a filter to screen out the weak relationships, and to bring efforts and attention to the most noteworthy areas.

Future work:

1. Suggestion for HTA: Market refining for visitor sectors from different regions is recommended to enhance consumer satisfaction and loyalty.
2. Continuous model improvement with onward data collection: As all the machine learning techniques, the more data is used to train the model, the more stable and accurate the resulting Bayesian network model is. As the models evolve, cross validation also has more data to compare the updated model with the older ones to select the best. It will also be possible to evaluate the learning rate through time, and to understand when the model is mature enough. As no similar prediction method is known for the Hawaii tourism market, this research also sets a baseline for future comparison.
3. Improvement of prediction accuracy: Figure 5.2 shows that with more intervals in discretization, the posterior inference results will fall into

narrower segments in the continuous data range. In this research, the selection of K in K-Means clustering was determined in the unsupervised learning stage where $K=4$ was enough to provide enough information to guide the next step in analysis. More work is needed to test the selection of different numbers of bins in supervised learning to find out the optimum accuracy. In practice, the setting of K and the desired level of accuracy may also be determined based on the user requirement.

4. Feedback and cooperation with service suppliers: This study aims at the consumer service industry, so the feedback from people who actually work in the related areas is of great value. A questionnaire (see Appendix 1) has been designed and sent out to organizations identified as representative service suppliers in Hawaii to gather their feedback on the key findings of this study. This survey has been approved by the Institutional Review Board (IRB) of Purdue University (see Appendix 2). The survey response confirmed that the results of this study is helpful in business decision making and achieving higher customer satisfaction and loyalty. One authority's feedback shown in Figure 7.1 suggested that the higher visitors arrivals from U.S. mainland to Hawaii is from June to August (same as the results from this study), and that the high in August is related to Labor Day vacation. This agrees with the concept raised in this study, that the statistical relationships learned from the Bayesian networks need to be interpreted with expert knowledge, experience or literature to be verified as causal relationships.

Leisure visitors to Hawaii have definite seasonal travel patterns which I think also are similar to the U.S. Mainland:
U.S. visitors' peak travel is summer months: June 15- Labor Day
Japan visitors' peak travel is August and also Late December - January.
Some recent increase in September travel

Figure 7.1 Quote from Survey Feedback

The response also suggested that other than the results obtained from this study, the responder wants to know about how much the visitor spend while travelling and on what. The author of this thesis also agreed that a detailed expenditure pattern analysis could be done, given the support from the service suppliers.

In addition, the validation results can be better measured given inputs from the service suppliers. Knowing the factor they are most concerned about and the variances' influences projected in real business operation helps develop a metric meaningful for decision making.

5. Application in other areas: Using the research approach proposed in this study, some exploratory efforts in other consumer service areas have been done (L. Zhang et al., 2014). It's recommended that the research approach to be further examined in other service industry sectors involving customer behavior characteristics and potential difficulty in data collection, such as health care and education.

The tourism industry is a field of intricacies and financial interests, yet lacking a thorough understanding. The nature of constantly changes and uncertainties, sensitivities to various factors, known or unknown, the heavy dependencies on

consumer experience and behavior, and the needs of decision making in complicated settings, all make tourism an ideal area of application of Bayesian network methodology. Bayesian networks as a data mining technique, allows comprehensive and visual analysis of a complex system. The research approach proposed in this study adds to the literature of Bayesian networks application, and provides valuable practical recommendation for service suppliers.

REFERENCES

REFERENCES

- Alpaydin, E. (2004). *Introduction to machine learning*. MIT press.
- Bakeman, R., & Quera, V. (2011). *Sequential Analysis And Observational Methods Behavioral Sciences*. Cambridge University Press. Retrieved from <http://www.cambridge.org/us/academic/subjects/psychology/social-psychology/sequential-analysis-and-observational-methods-behavioral-sciences>
- Ball, G. H., & Hall, D. J. (1967). A clustering technique for summarizing multivariate data. *Behavioral Science*, 12(2), 153–155.
- Bayesia. (n.d.-a). About Us. Retrieved November 07, 2013, from <http://www.bayesia.com/en/about-us/index.php>
- Bayesia. (n.d.-b). BayesiaLab. Retrieved November 07, 2013, from <http://www.bayesia.com/en/products/bayesialab.php>
- Bayesia. (n.d.-c). Features. Retrieved November 07, 2013, from <http://www.bayesia.com/en/products/bayesialab/features.php>
- Bayesia. (2012). Characterization of the Target Node. *BayesiaLab Library*. Retrieved August 20, 2014, from <http://library.bayesia.com/display/BlabC/Characterization+of+the+Target+Node>
- Bayesia. (2013a). Association Discovering. *BayesiaLab Library*. Retrieved August 19, 2014, from <http://library.bayesia.com/display/BlabC/Association+Discovering>
- Bayesia. (2013b). K-Means Discretization. *BayesiaLab Library*. Retrieved October 06, 2014, from <http://library.bayesia.com/display/FAQ/K-Means+Discretization>
- Bayesia. (2014). Score-Based Learning Algorithms. *BayesiaLab Library*. Retrieved August 18, 2014, from <http://library.bayesia.com/display/FAQ/Score-Based+Learning+Algorithms>

- Bower, B. (2013). Science & society: Bias seen in behavioral studies: Analysis suggests researchers too often find what they seek. *Science News*, 184(7), 10–10. doi:10.1002/scin.5591840707
- Bradley, P. S., & Fayyad, U. M. (1998). Refining Initial Points for K-Means Clustering. In *ICML* (Vol. 98, pp. 91–99). Citeseer.
- BusinessDictionary. (n.d.). What is service industry? definition and meaning. Retrieved August 05, 2014, from <http://www.businessdictionary.com/definition/service-industry.html>
- Castillo-Manzano, J. I., López-Valpuesta, L., & Gonzalez-Laxe, F. (2013). Profiling the Purpose of Travel: New Empirical Evidence. *Annals of Tourism Research*, 42, 425–428. doi:10.1016/j.annals.2013.02.004
- Center for Tobacco Research and Intervention. (2002). *INSIGHTS: SMOKING IN WISCONSIN*.
- Centers for Disease Control and Prevention. (2013). Risk Factors Exposures at Home and Work That May Cause Lung Cancer. Retrieved from http://www.cdc.gov/cancer/lung/basic_info/risk_factors.htm
- Chandoo. (2010). How to Select Right Chart for your Data | Chandoo.org - Learn Microsoft Excel Online. Retrieved August 16, 2014, from <http://chandoo.org/wp/2010/04/19/chart-selection-process/>
- Chase, R. B. (2010). Revisiting “Where Does the Customer Fit in a Service Operation?” In P. P. Maglio, C. A. Kieliszewski, & J. C. Spohrer (Eds.), *Handbook of Service Science* (pp. 11–17). Boston, MA: Springer US. doi:10.1007/978-1-4419-1628-0
- Coates, A., & Ng, A. Y. (2012). Learning feature representations with k-means. In G. B. Orr & K.-R. Müller (Eds.), *Neural Networks: Tricks of the Trade* (pp. 561–580). London: Springer-Verlag.
- Conrady, S., & Jouffe, L. (2013a). Driver Analysis and Product Optimization - White Papers - BayesiaLab's Library. Retrieved from <http://library.bayesia.com/display/whitepapers/Driver+Analysis+and+Product+Optimization>
- Conrady, S., & Jouffe, L. (2013b). Introduction to Bayesian Networks & BayesiaLab. Franklin, TN: BayesiaLab. Retrieved from <http://library.bayesia.com/display/whitepapers/Introduction+to+Bayesian+Networks+and+BayesiaLab>

- Conrady, S., & Jouffe, L. (2013c). Vehicle Size , Weight , and Injury Risk High-Dimensional Modeling and Causal Inference with Bayesian Networks Table of Contents. Franklin, TN: BayesiaLab. Retrieved from <http://library.bayesia.com/display/whitepapers/Vehicle+Size%2C+Weight%2C+and+Injury+Risk>
- Cook, W., & IIE Annual Conference & Expo. (2014). IIE Annual Conference & Expo 2014-Q&A with William Cook. Retrieved August 05, 2014, from <http://www.iienet2.org/Annual2/details.aspx?id=37067>
- Cover, T. M., & Thomas, J. A. (2006). ENTROPY, RELATIVE ENTROPY, AND MUTUAL INFORMATION. In *Elements of Information Theory* (2nd Editio., pp. 13–55). Hoboken, NJ, USA: Wiley-Interscience. Retrieved from <http://onlinelibrary.wiley.com.ezproxy.lib.purdue.edu/book/10.1002/047174882X;jsessionid=CFDDA55DA201485F6A2092EC7A278955.f02t03>
- Cranfield, J. A. L. (1999). *Aggregating non-linear consumer demands: A maximum entropy approach*. Purdue University.
- Crompton, J. L. (1979). Motivations for Pleasure Vacation. *Annals of Tourism*, 6(4), 408–424.
- De Mooij, M. K. (2010). *Consumer behavior and culture: Consequences for global marketing and advertising*. Thousand Oaks, Calif.: Sage Publications.
- Department of Business Economic Development & Tourism. (2012). *2012 Annual Report*. Retrieved from <http://dbedt.hawaii.gov/overview/annual-reports-reports-to-the-legislature>
- Department of Business Economic Development & Tourism. (2013). Income, Expenditures, and Wealth. In *2013 State of Hawaii Data Book*. Retrieved from <http://dbedt.hawaii.gov/economic/databook/db2013/>
- Department of Economic Development & Tourism. (2014). What are the major industries in the State of Hawaii? Retrieved August 17, 2014, from <http://dbedt.hawaii.gov/economic/library/faq/faq08/>
- Department of Electronics Information and Bioengineering. (n.d.). K-means Clustering. Retrieved August 19, 2014, from http://home.deib.polimi.it/matteucc/Clustering/tutorial_html/kmeans.html
- Devijver, P. A., & Kittler, J. (1982). *Pattern recognition: A statistical approach* (Vol. 761). Englewood Cliffs, N.J.: Prentice/Hall International.

- Donaldson, S. I., & Grant-Vallone, E. J. (2002). Understanding Self-Report Bias in Organizational Behavior Research. *Journal of Business and Psychology*, 32(2), 245–260. doi:10.1023/A:1019637632584
- Fayyad, U., Piatetsky-shapiro, G., & Smyth, P. (1996). From Data Mining to Knowledge Discovery in Databases, 17(3), 37–54.
- Feixas, M., Bardera, A., Rigau, J., & Xu, Q. (2014). Information Theory Basics. In *Information Theory Tools for Image Processing* (pp. 1–24). San Rafael: Morgan & Claypool Publishers.
- Felthous, A. R. (2014). Bias in behavioral study and analysis of international and domestic terrorism: an editorial introduction. *Behavioral Sciences & the Law*, 32(3), 263–8. doi:10.1002/bsl.2125
- Fenton, N. (2013). Norman Fenton Student Project Suggestions. Retrieved September 29, 2014, from https://www.eecs.qmul.ac.uk/~norman/student_projects.html
- Fenton, N., & Neil, M. (2012). *Risk assessment and decision analysis with Bayesian networks*. Boca Raton, FL: CRC Press.
- Fishe, W. W., Groff, R. A., & Roane, H. S. (2011). Applied Behavior Analysis History, Philosophy, Principles and Basic Methods. In W. W. Fishe, C. C. Piazza, & H. S. Roane (Eds.), *Handbook of Applied Behavior Analysis* (pp. 3–13). New York: Guilford Publications.
- Geisser, S. (1993). Introduction. In *Predictive inference* (Vol. 55, pp. 1–5). New York: Chapman & Hall.
- Gopaldas, A., & Fischer, E. (2012). Beyond Gender: Intersectionality, Culture, and Consumer Behavior. In C. C. Otnes & L. T. Zayer (Eds.), *Gender, Culture, and Consumer Behavior* (pp. 393–410). Hoboken: Taylor and Francis. Retrieved from [http://reader.ebilib.com.ezproxy.lib.purdue.edu/\(S\(3vqcprnxc3mkjwaoelwgobwt\)\)/Reader.aspx?p=957774&o=58&u=99AnZLJ1mdw%3d&t=1409298566&h=811FDCF8FEE8BD7CE15334E46F4A6FB93481F32C&s=25856402&ut=130&pg=426&r=img&c=-1&pat=n&cms=-1](http://reader.ebilib.com.ezproxy.lib.purdue.edu/(S(3vqcprnxc3mkjwaoelwgobwt))/Reader.aspx?p=957774&o=58&u=99AnZLJ1mdw%3d&t=1409298566&h=811FDCF8FEE8BD7CE15334E46F4A6FB93481F32C&s=25856402&ut=130&pg=426&r=img&c=-1&pat=n&cms=-1)
- Han, J., Kamber, M., & Pei, J. (2012a). 1 - Introduction. In *Data Mining: Concepts and Techniques* (3rd Ed., pp. 1–38). Burlington : Elsevier Science. doi:10.1016/B978-0-12-381479-1.00001-0

- Han, J., Kamber, M., & Pei, J. (2012b). 8 - Classification: Basic Concepts. In J. Han, M. Kamber, & J. Pei (Eds.), *Data Mining Concepts and Techniques* (Third Edit., pp. 327–391). Burlington : Elsevier Science. doi:10.1016/B978-0-12-381479-1.00008-3
- Han, J., Kamber, M., & Pei, J. (2012c). 9 – Classification: Advanced Methods. In *Data Mining: Concepts and Techniques* (3rd Ed., pp. 393–442). Burlington : Elsevier Science. doi:10.1016/B978-0-12-381479-1.00009-5
- Harrison-Hill, T. (2000). *IMPLICATIONS OF LONG HAUL TRAVEL ON THE MARKETING TOURISM*. Griffith University.
- Hartigan, J. A., & Wong, M. A. (1979). Algorithm AS 136: A k-means clustering algorithm. *Applied Statistics*, 100–108.
- Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2009a). Overview of Supervised Learning. In *The elements of statistical learning: data mining, inference and prediction* (2nd Ed., Vol. 27, pp. 9–41). New York: Springer.
- Hastie, T., Tibshirani, R., Friedman, J., & Franklin, J. (2009b). Unsupervised Learning. In *The elements of statistical learning: data mining, inference and prediction* (2nd Ed., Vol. 27, pp. 485–585). New York: Springer.
- Hawaii Tourism Authority. (2014). Visitor Highlights. Retrieved September 20, 2014, from <http://www.hawaiitourismauthority.org/research/research/visitor-highlights/>
- Heckerman, D., Mamdani, A., & Wellman, M. P. (1995). Real-world applications of Bayesian networks. *Communications of the ACM*, 38(3), 24–26.
- Herskovits, E. H., & Gerring, J. P. (2003). Application of a data-mining method based on Bayesian networks to lesion-deficit analysis. *NeuroImage*, 19(4), 1664–1673. doi:10.1016/S1053-8119(03)00231-3
- Heskett, J. L., & Sasser, W. E. (2010). The Service Profit Chain. In P. P. Maglio, C. A. Kieliszewski, & J. C. Spohrer (Eds.), *Handbook of Service Science* (pp. 19–29). Boston, MA: Springer US. doi:10.1007/978-1-4419-1628-0
- Hibbard, D., & Salbosa, A. (2006). The Close of An Era. In *Designing Paradise : The Allure of the Hawaiian Resort* (pp. 191–201). New York: Princeton Architectural Press. Retrieved from <http://site.ebrary.com/lib/purdue/docDetail.action?docID=10470257>

- Holmes, D. E., & Jain, L. C. (Eds.). (2008). *Innovations in Bayesian Networks* (Vol. 156). Berlin, Heidelberg: Springer Berlin Heidelberg. doi:10.1007/978-3-540-85066-3
- Huang, T.-M., Kecman, V., & Kopriva, I. (2006a). Introduction. In *Kernel based algorithms for mining huge data sets: Supervised, semi-supervised, and unsupervised learning* (Vol. 17, pp. 1–9). Berlin: Springer.
- Huang, T.-M., Kecman, V., & Kopriva, I. (2006b). Support Vector Machines in Classification and Regression — An Introduction. In *Kernel Based Algorithms for Mining Huge Data Sets* (Vol. 17, pp. 11–60). Berlin: Springer. doi:10.1007/3-540-31689-2
- Huang, T.-M., Kecman, V., & Kopriva, I. (2006c). Unsupervised Learning by Principal and Independent Component Analysis. In *Kernel Based Algorithms for Mining Huge Data Sets* (Vol. 208, pp. 175–208). Berlin: SpringerLink. Retrieved from <http://link.springer.com.ezproxy.lib.purdue.edu/book/10.1007%2F3-540-31689-2>
- Husmeier, D., Dybowski, R., & Roberts, S. (Eds.). (2005). *Probabilistic Modeling in Bioinformatics and Medical Informatics*. London: Springer-Verlag. doi:10.1007/b138794
- Jain, A. K. (2010). Data clustering: 50 years beyond K-means. *Pattern Recognition Letters*, 31(8), 651–666.
- Kacen, J. J., & Lee, J. A. (2002). The influence of culture on consumer impulsive buying behavior. *Journal of Consumer Psychology*, 12(2), 163–176.
- Kahng, S., Ingvarsson, E. T., Quigg, A. M., Kimberly E. Seckinger, & Teichman, H. M. (2011). Defining and Measuring Behavior. In W. W. FISHE, C. C. Piazza, & H. S. Roane (Eds.), *Handbook of Applied Behavior Analysis* (pp. 113–131). New York: Guilford Publications. Retrieved from [http://reader.ebilib.com.ezproxy.lib.purdue.edu/\(S\(mxq1voi0dzeljxyowae3u4a4\)\)/Reader.aspx?p=735602&o=58&u=99AnZLJ1mdw%3d&t=1407391171&h=81024776E8098AA87EFF681B71DBE1667F8FD530&s=24932357&ut=130&pg=1&r=img&c=-1&pat=n&cms=-1#](http://reader.ebilib.com.ezproxy.lib.purdue.edu/(S(mxq1voi0dzeljxyowae3u4a4))/Reader.aspx?p=735602&o=58&u=99AnZLJ1mdw%3d&t=1407391171&h=81024776E8098AA87EFF681B71DBE1667F8FD530&s=24932357&ut=130&pg=1&r=img&c=-1&pat=n&cms=-1#)
- Kanungo, T., Mount, D. M., Netanyahu, N. S., Piatko, C. D., Silverman, R., & Wu, A. Y. (2002). An efficient k-means clustering algorithm: Analysis and implementation. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 24(7), 881–892.

- Karayiannis, N. B., & Mi, G. W. (1997). Growing radial basis neural networks: merging supervised and unsupervised learning with network growth techniques. *Neural Networks, IEEE Transactions on*, 8(6), 1492–1506.
- Kasabov, N. (2001). Evolving fuzzy neural networks for supervised/unsupervised online knowledge-based learning. *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, 31(6), 902–918.
- Kersting, K., & De Raedt, L. (2001, November 23). Bayesian Logic Programs. Artificial Intelligence; Logic in Computer Science. Retrieved from <http://arxiv.org/abs/cs/0111058>
- Kohavi, R. (1995). A study of cross-validation and bootstrap for accuracy estimation and model selection. In *IJCAI* (Vol. 14, pp. 1137–1145).
- Lohrding, R. K., Johnson, M. M., & Whiteman, D. E. (1978). Computer graphics for extracting information from data. In *11th Annual Symposium on the Interface of Computer Science and Statistics*. Los Alamos: Office of Scientific and Technical Information, U.S. Dept. of Energy. Retrieved from <http://www.osti.gov/scitech/biblio/5197589>
- Luna, D., & Gupta, S. F. (2001). An integrative framework for cross-cultural consumer behavior. *International Marketing Review*, 18(1), 45–69.
- MacQueen, J. (1967). Some methods for classification and analysis of multivariate observations. In *Proceedings of the fifth Berkeley symposium on mathematical statistics and probability* (Vol. 1, pp. 281–297). California, USA.
- Maglio, P. P., Vargo, S. L., Caswell, N., & Spohrer, J. (2009). The service system is the basic abstraction of service science. *Information Systems and E-Business Management*, 7(4), 395–406. doi:10.1007/s10257-008-0105-1
- Maimon, O., & Rokach, L. (2010). Introduction to Knowledge Discovery and Data Mining. In O. Maimon & L. Rokach (Eds.), *Data Mining and Knowledge Discovery Handbook* (2nd ed., pp. 1–15). Boston, MA: Springer US. doi:10.1007/978-0-387-09823-4
- Mak, J. (2008). Tourism in Hawaii: An Overview. In *Developing a Dream Destination : Tourism and Tourism Policy Planning in Hawai'i* (pp. 13–45). Honolulu: University of Hawai'i Press. Retrieved from <http://site.ebrary.com/lib/purdue/docDetail.action?docID=10386623>

- Mccann, R. K., Marcot, B. G., & Ellis, R. (2006). INTRODUCTION / INTRODUCTION Bayesian belief networks : applications in ecology and natural resource management 1, (Reckhow 1999), 3053–3062. doi:10.1139/X06-238
- Moorman, R. H., & Podsakoff, P. M. (1992). A meta-analytic review and empirical test of the potential confounding effects of social desirability response sets in organizational behaviour research. *Journal of Occupational and Organizational Psychology*, 65(2), 131–149. doi:10.1111/j.2044-8325.1992.tb00490.x
- Mosteller, F. (2006). A k-sample slippage test for an extreme population. In *Selected Papers of Frederick Mosteller* (pp. 101–109). Springer.
- Moutinho, L. (2000). Tourism Marketing Research. In *Strategic Management in Tourism* (pp. 79–120). Wallingford, Oxon, GBR: CABI Publishing. Retrieved from <http://site.ebrary.com/lib/purdue/docDetail.action?docID=10257002>
- Munteanu, P., & Bendou, M. (2001). The EQ framework for learning equivalence classes of Bayesian networks. In *Proceedings 2001 IEEE International Conference on Data Mining* (pp. 417–424). IEEE Comput. Soc. doi:10.1109/ICDM.2001.989547
- Pearl, J. (1988). *Probabilistic reasoning in intelligent systems: networks of plausible inference*. San Mateo, Calif.: Morgan Kaufmann Publishers.
- Pearson, K. (1895). Note on Regression and Inheritance in the Case of Two Parents, Vol. 58 (1895), pp. 240-242. *Proceedings of the Royal Society of London*, 58, 240–242. Retrieved from <http://www.jstor.org.ezproxy.lib.purdue.edu/stable/115794?seq=2>
- Rajesh, R. (2013). Impact of Tourist Perceptions, Destination Image and Tourist Satisfaction on Destination Loyalty: A Conceptual Model. *PASOS. Revista de Turismo Y Patrimonio Cultural*, 11(3), 67–78.
- Rawlins, M. (2011). Bayes rules: the legacy of Thomas Bayes. *The Lancet*, 378(9804), 1692. doi:10.1016/S0140-6736(11)61728-5
- Ray, S., & Turi, R. H. (1999). Determination of number of clusters in k-means clustering and application in colour image segmentation. In *Proceedings of the 4th international conference on advances in pattern recognition and digital techniques* (pp. 137–143).

- Reisinger, Y. (2008). Globalization, tourism and culture. In *International Tourism Cultures and Behavior* (pp. 1–29). Amsterdam ; London: Elsevier. doi:10.1016/B978-0-7506-7897-1.00001-7
- Reisinger, Y. (2009a). Cultural Differences Among International Societies. In *International Tourism Cultures and Behavior* (pp. 347–373). Amsterdam ; London: Butterworth-Heinemann. doi:10.1016/B978-0-7506-7897-1.00015-7
- Reisinger, Y. (2009b). Cultural influences on tourist buying behavior. In *International Tourism Cultures and Behavior* (pp. 321–345). Amsterdam ; London: Butterworth-Heinemann. doi:10.1016/B978-0-7506-7897-1.00014-5
- Rencher, A. C., & Christensen, W. F. (2012). Introduction. In *Methods of Multivariate Analysis* (pp. 1–5). Hoboken, NJ, USA: John Wiley & Sons, Inc.
- Rish, I. (2001). An empirical study of the naive Bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence* (Vol. 3, pp. 41–46).
- Sabelhaus, J. (1990). Testing Neoclassical Consumer Theory with Aggregate and Household Data. *Applied Economics*, 22(11), 1471–1478. doi:10.1080/00036849000000117
- Sampson, S. E. (2010). The Unified Service Theory. In P. P. Maglio, C. A. Kieliszewski, & J. C. Spohrer (Eds.), *Handbook of Service Science* (pp. 107–131). SpringerLink. Retrieved from http://link.springer.com.ezproxy.lib.purdue.edu/chapter/10.1007/978-1-4419-1628-0_7/fulltext.html
- Schneider, B., & Bowen, D. E. (2010). Winning the Service Game. In P. P. Maglio, C. A. Kieliszewski, & J. C. Spohrer (Eds.), *Handbook of Service Science* (pp. 31–59). Boston, MA: Springer US. doi:10.1007/978-1-4419-1628-0
- Seaton, A., & Palmer, C. (1997). Understanding VFR tourism behaviour: the first five years of the United Kingdom tourism survey. *Tourism Management*, 18(6), 345–355. doi:10.1016/S0261-5177(97)00033-2
- Sebastiani, P., Abad, M. M., & Ramoni, M. F. (2010). *Data Mining and Knowledge Discovery Handbook*. (O. Maimon & L. Rokach, Eds.). Boston, MA: Springer US. doi:10.1007/978-0-387-09823-4
- Sebe, N., Cohen, I., Garg, A., & Huang, T. S. (2005). Introduction. In *Machine Learning in Computer Vision* (Vol. 29, pp. 1–13). Berlin/Heidelberg: Springer-Verlag. doi:10.1007/1-4020-3275-7

- Singh, S., & Appiah-Adu, K. (2008). Culture, Creativity, and Advertising. In S. Singh (Ed.), *Business Practices in Emerging and Re-emerging Markets* (pp. 133–150). New York: Palgrave Macmillan.
- Stassopoulou, A., & Petrou, M. (1998). Obtaining the correspondence between Bayesian and Neural Networks. *International Journal of Pattern Recognition and Artificial Intelligence*, 12(7), 901–920.
- State of Hawaii, & Hawaii Tourism Authority. (2004). *Hawai'i Tourism Strategic Plan, 2005-2015*. Retrieved from http://www.hawaii tourism authority.org/default/assets/file/about/tsp2005_2015_final.pdf
- Tourism Research Division. (2014). Research & Reports. Retrieved November 07, 2013, from <http://www.hawaii tourism authority.org/research/>
- Tourism Research Division of Hawaii Tourism Authority. (2014). Annual Visitor Research Report. Retrieved August 10, 2014, from <http://www.hawaii tourism authority.org/research/reports/annual-visitor-research/>
- Tsamardinos, I., Aliferis, C. F., Statnikov, A. R., & Statnikov, E. (2003). Algorithms for Large Scale Markov Blanket Discovery. In *FLAIRS Conference* (Vol. 2003, pp. 376–381).
- U, I. M. (2007). *A Comparative Study of Tourist Behaviour in Pleasure Travel: Group Package Tours and Free Independent Tours*. University of South Australia.
- U.S. Census Bureau. (2013). *2008-2012 American Community Survey 5-Year Estimates State of Hawaii Data Profiles DP05* (pp. 1–3). Retrieved from http://files.hawaii.gov/dbedt/census/acs/ACS2012/ACS2012_5_Year/acs_hi_2012_geographic_5_yr/acs12_hi_5yr.pdf
- Votolato, G. (2007). INTRODUCTION. In *Transport Design* (pp. 7–20). London: Reaktion Books Ltd. Retrieved from <http://search.proquest.com.ezproxy.lib.purdue.edu/docview/905227402/985B1D31E5CE4886PQ/3?accountid=13360>
- Waisberg, D. (2014). Tell a Meaningful Story With Data – Think with Google. Google. Retrieved from <http://www.thinkwithgoogle.com/articles/tell-meaningful-stories-with-data.html>

- Weber, P., Medina-Oliva, G., Simon, C., & Lung, B. (2012). Overview on Bayesian networks applications for dependability, risk analysis and maintenance areas. *Engineering Applications of Artificial Intelligence*, 25(4), 671–682. doi:10.1016/j.engappai.2010.06.002
- Wikipedia. (2013). Category:Service industries - Wikipedia, the free encyclopedia. Retrieved August 05, 2014, from http://en.wikipedia.org/wiki/Category:Service_industries
- Wikipedia. (2014a). k-means clustering. Retrieved August 01, 2014, from http://en.wikipedia.org/wiki/K-means_clustering
- Wikipedia. (2014b). Pearson product-moment correlation coefficient. Retrieved August 16, 2014, from http://en.wikipedia.org/wiki/Pearson_product-moment_correlation_coefficient
- Witten, I. H., & Frank, E. (2005a). Implementations: Real Machine Learning Schemes. In *Data Mining: Practical machine learning tools and techniques* (2nd Ed., pp. 187–284). Amsterdam ; Boston: Morgan Kaufmann.
- Witten, I. H., & Frank, E. (2005b). What's It All About. In *Data Mining: Practical machine learning tools and techniques* (2nd ed., pp. 3–40). Amsterdam ; Boston: Morgan Kaufmann.
- Woo, I. (2012). *Information-assisted data exploration, analysis and visualization techniques - ProQuest*. Purdue University. Retrieved from <http://search.proquest.com.ezproxy.lib.purdue.edu/docview/1328395135>
- Xu, R., & Wunsch, D. C. (2008). Partitional Clustering. In *Clustering* (pp. 63–110). Hoboken, NJ, USA: John Wiley & Sons, Inc. doi:10.1002/9780470382776
- Zhang, H. (2004). The optimality of naive Bayes. *AA*, 1(2), 3.
- Zhang, L., Gao, Y., Bidassie, B., & Duffy, V. G. (2014). Application of Bayesian Networks in Consumer Service Industry and Healthcare. In *Digital Human Modeling. Applications in Health, Safety, Ergonomics and Risk Management* (pp. 484–495). Springer.
- Zhao, Z., & Liu, H. (2007). Spectral feature selection for supervised and unsupervised learning. In *Proceedings of the 24th international conference on Machine learning* (pp. 1151–1157). ACM.
- Zhu, X., & Goldberg, A. B. (2009). Introduction to Semi-Supervised Learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 3(1), 1–130. doi:10.2200/S00196ED1V01Y200906AIM006

APPENDICES

Appendix A Permission for the Use of Data

The permission was given by email. The appendix shows the original email theme. For the purpose of privacy protection, the names and contact information of the related persons at Hawaii Tourism Authority masked out.

7/8/2014

myMail

myMail

gao186@purdue.edu

[± Font Size](#) ▾

RE: Permission of referencing tourist statistics data in research thesis

From : [REDACTED] <[REDACTED]@gohta.net> Wed, Nov 06, 2013 08:03 PM
Subject : RE: Permission of referencing tourist statistics data in research thesis
To : gao186@purdue.edu
Cc : Daniel Nahoopii <Daniel@gohta.net>

Aloha Yuan Gao,

Thank you for your e-mail and interest in our visitor data.

What we publish on the Hawai'i Tourism Authority website can be used for your research thesis.

If you are using statistics: Visitor Days, visitor arrivals, visitor spending, Visitor Satisfaction, Visitor Plant Inventory stats etc. Please site HTA as the source.

If you are using hotel occupancy, Average Daily Room Rates, or Revenues per available room statistics, I.e. show on pages 146-148 in our 2012 Annual Visitor Research Report, please site, Smith Travel Research, Hospitality Advisors, LLC, as the source of this data.

Aloha,

[REDACTED]

Aloha,

[REDACTED]

Tourism Research Manager
 Hawai'i Tourism Authority
 Hawai'i Convention Center
 1801 Kalakaua Avenue, Honolulu, Hawai'i 96815

web hawaii tourism authority.org

Please update your record. Here is my new email address: [REDACTED]

tel [REDACTED] **fax** [REDACTED]

From: [REDACTED] [mailto:[REDACTED]@hawaii tourism authority.org] **On Behalf Of** &InfoHTA

Sent: Wednesday, November 06, 2013 2:40 PM

To: [REDACTED]

Subject: Fw: Permission of referencing tourist statistics data in research thesis

Aloha [REDACTED],

Can you please assist with the following inquiry and respond?

Thanks,

[REDACTED]

Hawai'i Tourism Authority
 1801 Kalakaua Avenue, 1st Floor
 Honolulu, HI 96815

[REDACTED] fax

www.hawaii tourism authority.org

----- Forwarded by [REDACTED] /HTA/DBEDT on 11/06/2013 02:38 PM -----

From: Yuan Gao <gao186@purdue.edu>

To: info@hawaii tourism authority.org,

Date: 11/03/2013 12:09 PM

Subject: Permission of referencing tourist statistics data in research thesis

<https://my mail.purdue.edu/h/printmessage?id=12086&1>

1/2

Figure A.1 Permission of Data Use by Hawaii Tourism Authority_1

7/8/2014

myMail

Dear Hawaii Tourism Authority,

My name is Yuan Gao. I'm a graduate student at the School of Industrial Engineering at Purdue University, Indiana. I am working on my Master's thesis concerning causal relationship analysis. I have a personal interest in traveling, and after coming across your website (<http://www.hawaiitourismauthority.org>), I think that the data published under the "Research and Reports" section can make a very good source material for the application of the methodology discussed in my thesis. I've read through the Term of Use and see no conflict or violation with my purpose of use. But I still would like to confirm with you to be sure.

I will use data including tourists origin, arrivals, expenditure, purpose of travel, length of stay and types of trip (group, package) and analyze the relationship among these factors. A full citation will be included in the thesis to indicate your website as the original source of these data. The thesis is for research purpose only and not of any commercial interest. No human subject identification is involved.

I'd appreciate your response with consent of the use of data as stated above. Should there be any concern or question, please do not hesitate to let me know.

Thanks and Best Regards!

Yuan Gao (Gill)

Figure A.2 Permission of Data Use by Hawaii Tourism Authority_2

Appendix B Purdue IRB Approval

HUMAN RESEARCH PROTECTION PROGRAM
INSTITUTIONAL REVIEW BOARDS

To: VINCENT DUFFY
GRIS 236

From: JEANNIE DICLEMENTI, Chair
Social Science IRB

Date: 07/24/2014

Committee Action: **Exemption Granted**

IRB Action Date: 07/24/2014

IRB Protocol #: 1407015022

Study Title: Master thesis: Application of Bayesian Networks in Consumer Service Industry: A Case Study of Hawaii Tourism Market

The Institutional Review Board (IRB) has reviewed the above-referenced study application and has determined that it meets the criteria for exemption under 45 CFR 46.101(b)(4) .

If you wish to make changes to this study, please refer to our guidance "**Minor Changes Not Requiring Review**" located on our website at <http://www.irb.purdue.edu/policies.php>. For changes requiring IRB review, please submit an **Amendment to Approved Study** form or **Personnel Amendment to Study** form, whichever is applicable, located on the forms page of our website www.irb.purdue.edu/forms.php. Please contact our office if you have any questions.

Below is a list of best practices that we request you use when conducting your research. The list contains both general items as well as those specific to the different exemption categories.

General

- To recruit from Purdue University classrooms, the instructor and all others associated with conduct of the course (e.g., teaching assistants) must not be present during announcement of the research opportunity or any recruitment activity. This may be accomplished by announcing, in advance, that class will either start later than usual or end earlier than usual so this activity may occur. It should be emphasized that attendance at the announcement and recruitment are voluntary and the student's attendance and enrollment decision will not be shared with those administering the course.
- If students earn extra credit towards their course grade through participation in a research project conducted by someone other than the course instructor(s), such as in the example above, the students participation should only be shared with the course instructor(s) at the end of the semester. Additionally, instructors who allow extra credit to be earned through participation in research must also provide an opportunity for students to earn comparable extra credit through a non-research activity requiring an amount of time and effort comparable to the research option.
- When conducting human subjects research at a non-Purdue college/university, investigators are urged to contact that institution's IRB to determine requirements for conducting research at that institution.
- When human subjects research will be conducted in schools or places of business, investigators must obtain written permission from an appropriate authority within the organization. If the written permission was not

Figure B.3 IRB Approval for Conducting Survey with Service Providers_1

submitted with the study application at the time of IRB review (e.g., the school would not issue the letter without proof of IRB approval, etc.), the investigator must submit the written permission to the IRB prior to engaging in the research activities (e.g., recruitment, study procedures, etc.). This is an institutional requirement.

Category 1

- When human subjects research will be conducted in schools or places of business, investigators must obtain written permission from an appropriate authority within the organization. If the written permission was not submitted with the study application at the time of IRB review (e.g., the school would not issue the letter without proof of IRB approval, etc.), the investigator must submit the written permission to the IRB prior to engaging in the research activities (e.g., recruitment, study procedures, etc.). This is an institutional requirement.

Categories 2 and 3

- Surveys and questionnaires should indicate
 - only participants 18 years of age and over are eligible to participate in the research; and
 - that participation is voluntary; and
 - that any questions may be skipped; and
 - include the investigator's name and contact information.
- Investigators should explain to participants the amount of time required to participate. Additionally, they should explain to participants how confidentiality will be maintained or if it will not be maintained.
- When conducting focus group research, investigators cannot guarantee that all participants in the focus group will maintain the confidentiality of other group participants. The investigator should make participants aware of this potential for breach of confidentiality.
- When human subjects research will be conducted in schools or places of business, investigators must obtain written permission from an appropriate authority within the organization. If the written permission was not submitted with the study application at the time of IRB review (e.g., the school would not issue the letter without proof of IRB approval, etc.), the investigator must submit the written permission to the IRB prior to engaging in the research activities (e.g., recruitment, study procedures, etc.). This is an institutional requirement.

Category 6

- Surveys and data collection instruments should note that participation is voluntary.
- Surveys and data collection instruments should note that participants may skip any questions.
- When taste testing foods which are highly allergenic (e.g., peanuts, milk, etc.) investigators should disclose the possibility of a reaction to potential subjects.

Figure B.4 IRB Approval for Conducting Survey with Service Provider_2