A Multivariate Statistical Analysis of State Desirability

Jennifer Everson
Carthage College [*]

Melissa Hildt
College of Notre Dame of Maryland [†]

Jason Popovic
Baldwin-Wallace College [‡]

Sarah Zimmermann
Bemidji State University [§]

July 24, 2001

Abstract

State desirability can be measured by a set of several different variables. The multivariate statistical methods of factor analysis and discriminant analysis lend themselves to this issue. We used factor analysis to reduce a large number of variables to a smaller set of common factors which describe state desirability. We then used discriminant analysis to classify states according to their desirability level based upon a set of measured variables.

1 Introduction

Following the publication of statistics revealing that North Dakota's growth rate is the slowest in the United States, North Dakota state representatives are putting forth a proposal to rename the state. According to National Public Radio's *Morning Edition* (June 27, 2001), members of the North Dakota state legislature believe that their state's diminished growth rate can partially be attributed to the word *North* in its name–distorting people's perception of the state's climate and therefore causing people to leave the state and causing very few others to relocate to the state. Legislative members feel that an unappealing name further compounds the state's declining population. Therefore, state legislative members in North Dakota have proposed that if they remove *North* from the state's name, then perhaps people will not view the state as being a less desirable state to live in.

While slightly comical, this very real example has led us to an interesting series of questions whose answers could have serious ramifications for a particular state. Are certain states experiencing the reverse effect–increased immigration without increased emigration? If a state's name affects its appeal, what other elements influence people to change or keep

[*] Department of Mathematics, 2001 Alford Park Drive, Kenosha, WI 53140
[†] Department of Mathematics, 4701 North Charles Street, Baltimore, MD 21210
[‡] Department of Mathematics and Computer Science, 275 Eastland Road, Berea, OH 44017
[§] Department of Mathematics and Computer Science, 1500 Birchmont Drive NE, Bemidji, MN 56601

their existing residency? What makes a state desirable or not? Statisticians can help begin to answer some of these questions by using multivariate statistical techniques.

2 Methodology

We will use factor analysis and discriminant analysis to analyze state desirability. First, we will determine the variables that we believe have an impact on the desirability of a state. Then we will use factor analysis to find a set of underlying common factors within these variables. Using discriminant analysis and findings from factor analysis, we will attempt to build a model which can determine state desirability. We determine state desirability by the rate of population growth, assuming that current desirability causes future population growth. In order to complete factor analysis and discriminant analysis, we will use the statistical software program Minitab.

For our investigation we needed to determine variables that influence a person's decision to live in a particular state. We decided that economic, education, geographic, and demographic characteristics of a state are reasons why an individual would consider a state more or less desirable to live in. For each of these characteristics, we decided to use the following measurable variables listed in Table 2.1.

Table 2.1 Desirability Categories and Corresponding Variables

Economic state health and hospital spending per capita percent of population below poverty level percent of population unemployed general spending per capita personal income per capita sales tax	**Demographic** population density percent minorities median age crime rate
Education state education spending per capita graduation rate	**Geographic** average precipitation difference in elevation average temperature

Based on the data corresponding to the variables shown above (see Appendix A), we decided to look only at the 48 contiguous states because both the discriminant analysis and factor analysis models are greatly influenced by outliers. Based on our data, we classified Alaska and Hawaii as being outliers. In addition, we concluded that the decision to move outside the continental United States is a more involved decision than electing to move within the continental United States. Therefore, we disregarded Alaska and Hawaii in our analyses.

3 Data Collection

The data we collected came from four different sources: the *Statistical Abstract of the United States*, *The World Almanac and Book of Facts*, *Almanac of the 50 States*, and the United States census. We were able to obtain some of the variables directly while others were calculated from the information available.

The following variables were taken directly from their sources: 1990 and 2000 population counts [6], percent below poverty level [11], percent excise tax rate (general sales and gross receipts) [10], percent unemployed (as percent of civilian labor force) [17], total crime rate (offenses known to the police per 100,000 population) [20], population density (population per square mile of land area) [8], graduation rate (percent of students enrolled as a senior in high school in the fall of 1989 that graduated in the spring of 1990) [1], median age [5], general expenditures per capita [13], and personal income per capita (in dollars) [16].

The remaining variables were obtained indirectly through simple calculations. To calculate the percentage of minorities we found the difference between the total population and the non-minority population and divided by total population [6]. State health and hospital spending per capita and state education spending per capita were calculated by dividing the respective state spending by the total 1990 population [13]. Difference in elevation (feet) was calculated as the maximum point of elevation minus the minimum point of elevation [21]. Average temperature (degrees Fahrenheit) was calculated as the average of the normal daily mean temperatures for the cities listed within each state [22]. In the case where more than one city was listed for a state, we averaged the given cities. Average precipitation (inches) data was calculated in the same manner as average temperature [2].

4 Factor Analysis

4.1 Theory

The objective of factor analysis is to obtain a number of common factors less than the original number of variables and thus reduce the dimensionality of a set of measurable variables. Factor analysis groups highly correlated variables together and helps us identify the set of underlying common factors.

However, not all of the original information is accounted for in the new set of common factors because each response variable has an attribute that is unique to itself. We call this attribute the unique, or specific, factor for that variable. Using the common factors, specific factors, and data for the experimental units, we can build what is called a factor analysis model.

To obtain the factor analysis model, we must make a few assumptions. First, we must assume that $\underline{x} \sim \mathrm{N}_p(\underline{\mu}, \boldsymbol{\Sigma})$, where $\underline{x}$ is the $p \times 1$ vector $[\, x_1 \quad x_2 \quad \cdots \quad x_p \,]'$ of response variables, $\underline{\mu}$ is the $p \times 1$ vector $[\, \mu_1 \quad \mu_2 \quad \cdots \quad \mu_p \,]'$ of the means of the response variables, and

$\mathbf{\Sigma} = \begin{bmatrix} \sigma_{11} & \sigma_{12} & \cdots & \sigma_{1m} \\ \sigma_{21} & \sigma_{22} & \cdots & \sigma_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_{pm} \end{bmatrix}$ is the variance-covariance matrix of $\underset{\sim}{\mathrm{x}}$. Statistical testing has shown that factor analysis is robust to non-normally distributed variables; but since normally distributed variables provide increased confidence in our results, we must test each variable for normality.

One way to test for normality is the Quantile-Quantile (Q-Q) plot correlation coefficient test. A Q-Q plot is a discrete graph of the standardized values of a variable versus the actual values of the variable. The correlation coefficient ρ of the Q-Q plot is therefore $0 \leq \rho \leq 1$. Since we are interested in assessing the variable's normality, we test the hypothesis $\rho = 1$. If we reject the hypothesis, we reject normality of the variable.

The second assumption we must make is that $\underset{\sim}{\mathrm{f}} \sim \mathrm{N}_m(\mathbf{0}, \mathbf{I})$, where $\mathbf{I}$ is the $m \times m$ identity matrix, and $\underset{\sim}{\mathrm{f}} = [\, f_1 \quad f_2 \quad \cdots \quad f_m \,]'$ is the $m \times 1$ vector of common factors. Third, we assume that $\underset{\sim}{\eta} \sim \mathrm{N}_p(\mathbf{0}, \mathbf{\Psi})$, where $\underset{\sim}{\eta} = [\, \eta_1 \quad \eta_2 \quad \cdots \quad \eta_p \,]'$, the $p \times 1$ vector of specific factors, and $\mathbf{\Psi}$ is the diagonal $p \times p$ matrix $\begin{bmatrix} \psi_1 & 0 & \cdots & 0 \\ 0 & \psi_2 & & \vdots \\ \vdots & & \ddots & 0 \\ 0 & \cdots & 0 & \psi_p \end{bmatrix}$, where ψ_j is the variance, or specificity, of η_j. The assumptions regarding the means and standard deviations for the previous two distributions can be made without any loss of generality. Finally, we must assume that f_k and η_j are independent for all combinations of k and j, where $k = 1, 2, ..., m$ and $j = 1, 2, ..., p$. Thus, the covariance of f_k and η_j is 0.

We can now write a linear equation to represent each of our response variables. We obtain $x_i = \lambda_{i1} f_1 + \lambda_{i2} f_2 + \cdots + \lambda_{im} f_m + \eta_i$ for $i = 1, 2, ..., p$. We can build this model because λ_{ik} is the covariance between the i_{th} response variable and the k^{th} factor, f_k. We call λ_{ik} the *factor loading* of the i^{th} response variable on the k^{th} factor. This equation can be written in matrix form as $\underset{\sim}{\mathrm{x}} = \mathbf{\Lambda} \underset{\sim}{\mathrm{f}} + \underset{\sim}{\eta}$, where $\mathbf{\Lambda} = \begin{bmatrix} \lambda_{11} & \lambda_{12} & \cdots & \lambda_{1m} \\ \lambda_{21} & \lambda_{22} & \cdots & \lambda_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \lambda_{p1} & \lambda_{p2} & \cdots & \lambda_{pm} \end{bmatrix}$, the $p \times m$ matrix of factor loadings. From this equation, it can be shown that $\mathbf{\Sigma} = \mathbf{\Lambda}\mathbf{\Lambda}' + \mathbf{\Psi}$.

The communality of a variable, $h_i^2 = \sum_{k=1}^{m} \lambda_{ik}^2$, is the percent of variance that is explained by the common factors. It can be shown that $\sigma_{ii} = h_i^2 + \psi_i$, where σ_{ii} is the variance of x_i. If we standardize the values of the response variables, σ_{ii} equals ρ_{ii}, the correlation of x_i with itself, which is 1. Then we have $\sigma_{ii} = \rho_{ii} = 1 = h_i^2 + \psi_i$, and consequently $\psi_i = 1 - h_i^2$. We use this equation to obtain the values of the specific factors.

Since we will standardize the values for all of the response variables, we replace $\mathbf{\Sigma}$ with

$\mathbf{P} = \begin{bmatrix} \rho_{11} & \rho_{12} & \cdots & \rho_{1m} \\ \rho_{21} & \rho_{22} & \cdots & \rho_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ \rho_{p1} & \rho_{p2} & \cdots & \rho_{pm} \end{bmatrix}$ and obtain the equation $\mathbf{P} = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi}$. Then λ_{ik} becomes the correlation between the i^{th} variable and the k^{th} common factor.

Using the sample correlation matrix, which is the maximum likelihood estimate of $\mathbf{P}$, we choose a number m of common factors and seek to find $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ such that $\mathbf{P} = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi}$. If a solution does exist, the variance between the variables is explained exactly by the m common factors. We find the maximum likelihood estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$, the values which are most likely to have produced the data.

However, the solution first found (whether exact or a maximum likelihood estimate) is not a unique solution because $\mathbf{\Lambda}$ may be multiplied by an $m \times m$ orthogonal matrix $\mathbf{T}$ to obtain a new matrix of factor loadings, namely $\mathbf{\Lambda T}$. Recall that $\mathbf{TT}' = \mathbf{I}$, the identity matrix, by definition of orthogonality. Then we have $(\mathbf{\Lambda T})(\mathbf{\Lambda T})' + \mathbf{\Psi} = \mathbf{\Lambda TT}'\mathbf{\Lambda}' + \mathbf{\Psi} = \mathbf{\Lambda I \Lambda}' + \mathbf{\Psi} = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi} = \mathbf{P}$ so we maintain the equality $\mathbf{P} = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi}$. Since there are infinitely many orthogonal $m \times m$ matrices for all $m > 1$, there are infinitely many solutions for $\mathbf{\Lambda}$ given a set of data with more than one common factor. Multiplying $\mathbf{\Lambda}$ by an orthogonal matrix $\mathbf{T}$ is called rotating the factors through a certain angle because each of the factor axes is rotated.

There are specific rotational methods which can be used to simplify identification of the common factors. The method by which we have elected to rotate is called the Varimax rotation technique. This method maximizes the variation between the factor loadings (makes the larger loadings larger and the smaller loadings smaller) so that we can more easily identify and label the factors.

To determine the number of common factors, we must first make an initial guess. A good number to start with is the number of eigenvalues of the correlation matrix which are greater than 1. We then find the maximum likelihood estimates of $\mathbf{\Lambda}$ and $\mathbf{\Psi}$ so that we can test the hypothesis that the number of factors chosen is sufficient using the Chi-squared (χ^2) statistic. We test the hypothesis that $\mathbf{P} = \mathbf{\Lambda\Lambda}' + \mathbf{\Psi}$. We let n, p, m, and ν represent the number of experimental units, the number of response variables, the number of factors, and the degrees of freedom $(n-1)$, respectively. We let $\hat{\mathbf{\Lambda}}$, $\hat{\mathbf{\Lambda}}'$, $\hat{\mathbf{\Psi}}$, and $\hat{\mathbf{\Sigma}}$, represent the maximum likelihood estimates of $\mathbf{\Sigma}$, $\mathbf{\Lambda}$, $\mathbf{\Lambda}'$, and $\mathbf{\Psi}$, respectively. We then obtain the following equation.

$$\chi^2_\nu = \left(n - 1 - \frac{(2p + 4m + 5)}{6}\right) \ln\left(\frac{|\hat{\mathbf{\Lambda}}\hat{\mathbf{\Lambda}}' + \hat{\mathbf{\Psi}}|}{|\hat{\mathbf{\Sigma}}|}\right)$$

Since we are standardizing the values of the response variables and using $\mathbf{P}$, the correlation matrix, instead of the variance-covariance matrix $\mathbf{\Sigma}$, we replace $\hat{\mathbf{\Sigma}}$ in the above equation with $\mathbf{R}$, the sample correlation matrix. We subjectively set a significance level α, and proceed by testing our initial guess as to the number of common factors. After computing χ^2_ν, we calculate the probability, or P-value, of obtaining a χ^2 value at least as extreme as χ^2_ν.

If the P-value is less than α, we reject the hypothesis that the number of factors chosen is sufficient. We then increase the number of factors by one and repeat the χ^2 test statistic until the hypothesis is accepted. If the P-value is greater than α, we accept the original hypothesis. We then decrease the number of factors by one until it is rejected. We use the number of factors previous to the number which was rejected as the final number of common factors.

Once we have determined the number of factors, we use the Varimax rotation method to find the matrix of factor loadings and communalities. Factor identification, using the factor loadings, is a very subjective process. Since it is difficult to consider and scrutinize all the factor loadings and variables for a common factor, a researcher may choose to consider only those factor loadings which are of a certain magnitude or greater (either in the positive or negative direction). For each factor we must consider both what the variable is measuring and its specific magnitude. Since the factor loadings are correlations, the closer a variable's factor loading is to 1 or -1, the greater its contribution to the factor. For example, a variable with a factor loading close to 1 or -1 may very well be the sole determiner of a factor, while the other highly correlated variables are results of this one variable. Therefore, it takes a subjective (and objective) mind to discern the factors. If several variables are highly and similarly correlated, then all contribute equally to that factor. Often, one uses a broader category name for the factor (e.g., economic factor) to encompass several variables (e.g., Gross National Product, percent unemployment, interest rate, and percent inflation).

It is also important to understand that both negative and positive correlations are significant. If two variables both contribute significantly to a factor but one is positive and another is negative, we must find a factor which would explain these correlations. A factor which has both highly positive and negative factor loadings is called a bipolar factor.

Once all of the factors have been determined, the information can be used to reduce the number of variables that must be measured in future studies. If factor analysis shows only five out of twenty variables contribute significantly to the factors, then future studies on the same topic may be done more cost- or time-effectively using only these five variables.

We can also use factor analysis to determine a set of factor scores. A factor score is a number assigned to each member of the population for each common factor. This helps to reduce dimensionality. A population with twenty characteristics to observe can quickly become a population with only four or five characteristics to compare. These are some of the major applications of factor analysis.

4.2 Results

In the case of our project, we are using factor analysis to find the common factors involved in state desirability. We first tested each of the fifteen variables under the hypothesis that the variable was normal. At the .01 significance level, the hypothesis of normality was rejected for the following variables: general spending per capita, percent sales tax, population density, median age, and difference in elevation. The other eleven variables were not rejected under the hypothesis of normality. Recall, however, that factor analysis is robust to non-normally distributed variables, and therefore we proceed to construct the factor analysis model. We begin by making an initial guess as to the number of factors. After finding the eigenvalues of the correlation matrix, we found that five of the values were greater than 1. Therefore, we hypothesized that five factors would sufficiently describe our data.

Next we tested this hypothesis. Using Minitab and Microsoft Excel, we computed the χ^2 statistic for the 5-factor model. We chose to use a significance level of .05–subjectively determined, but the common standard. The χ^2 value of the 5-factor model was 50.25. The probability, or P-value, that the 5-factor model sufficiently described the data was approximately 0.3458. We then reduced the number of factors to 4 to test the hypothesis that a 4-factor model would adequately describe the data. The χ^2 value of the 4-factor model was 80.55, which yielded a P-value of approximately .0017. Since this was clearly below our chosen level of significance, we accepted the 5-factor model, which, according to Minitab, accounts for 69.3% of the variance between the response variables.

4.3 Discussion

Using the matrix of factor loadings and communalities for the 5-factor model given by Minitab (see Table 4.1), we were able to identify the common factors. The factor loadings in boldface are those which are greater than .400 or less than -.400. Recall that a factor loading represents the correlation between a variable and a factor, so the closer a factor loading's absolute value is to one, the more we should consider it in naming a common factor.

Table 4.1 Factor Loadings for the 5-factor model

Response Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5
Health and Hospital Spending per Capita	0.154	-0.279	-0.319	**-0.601**	0.067
Percent of Below Poverty Level	0.258	-0.051	0.297	0.143	**-0.907**
Percent of Population Unemployed	0.162	-0.166	-0.014	-0.027	**-0.530**
General Spending per Capita	-0.194	0.106	-0.210	**-0.949**	0.079
Personal Income per Capita	0.071	-0.048	**-0.853**	-0.119	**0.421**
Percent Sales Tax	0.084	-0.037	-0.239	0.022	-0.146
Population Density	0.093	-0.302	**-0.691**	-0.311	0.188
Percent Minorities	**0.842**	0.148	-0.171	-0.069	-0.358
Median Age	-0.220	-0.344	-0.395	0.065	0.004
Crime Rate	**0.794**	0.216	-0.153	-0.039	0.096
Education Spending per Capita	0.058	0.232	**0.411**	**-0.703**	-0.056
Graduation Rate	**-0.782**	0.284	0.029	0.032	0.277
Average Precipitation	0.251	**-0.950**	-0.105	0.039	-0.146
Difference in Elevation	0.129	**0.700**	0.204	-0.023	0.139
Average Temperature	**0.777**	-0.203	0.180	0.185	-0.258

As shown in Table 4.1, the percent minorities, crime rate, graduation rate, and average temperature were the most highly correlated with the first factor. For this reason, while perhaps the hardest factor to interpret, we decided to label the first factor "Social Climate." The second factor is highly correlated with both average precipitation and difference in elevation, so we named it the "Geographic" factor.

We had difficulty distinguishing between Factor 3 and Factor 5 because they had one variable in common (personal income) and seemed to be measuring the same type of situation, economic. We finally decided to name Factor 3 and Factor 5 “Personal Wealth” and “Job Opportunities and Pay,” respectively. Factor 3 is correlated highly with personal income per capita and population density. We named Factor 3 based on the personal income per capita variable, assuming that population density and state education spending per capita are causes or effects of personal income. Factor 5 was largely determined by the percent of population below the poverty level, which had a factor loading of -.907. Personal income and the percent of population unemployed are, of course, also associated with job opportunities and pay. Finally, we classified the fourth factor as “State Spending” since education spending, health spending, and general spending all had high factor loadings for this factor.

Now that we have determined the common factors in our data, we attempt to create a model, using discriminant analysis, to determine whether a state is more or less desirable. We can use these findings from factor analysis to determine which variables are necessary in the discriminant analysis model.

5 Discriminant Analysis

5.1 Theory

Discriminant analysis is a multivariate statistical method used to classify observational units into the correct population, examine differences between these populations, establish the best way to differentiate between populations, and check if the assignment rule is accurate. The basic goal of discriminant analysis is to create a rule that will successfully place an unknown experimental unit into the correct population.

To begin, let $\underset{\sim}{x}$ be a p-variate vector of measured variables for an experimental unit in the population $\Pi_1 = \mathrm{N}_p(\underset{\sim}{\mu}_1, \Sigma)$ or the population $\Pi_2 = \mathrm{N}_p(\underset{\sim}{\mu}_2, \Sigma)$. We assume all observations from each of the two populations come from a random sample and each population is normally distributed, although discriminant analysis is robust to non-normality.

Additionally, let Ω denote the p-dimensional space of $\underset{\sim}{x}$ which is partitioned into the two regions R_1 and R_2, where $\mathrm{R}_1 \cup \mathrm{R}_2 = \Omega$ and $\mathrm{R}_1 \cap \mathrm{R}_2 = \emptyset$. Our goal is to classify $\underset{\sim}{x}$ to Π_1 or Π_2 using the partitions R_1 and R_2. When $\underset{\sim}{x}$ is in R_1, $\underset{\sim}{x}$ will be classified into Π_1, and when $\underset{\sim}{x}$ is in R_2, $\underset{\sim}{x}$ will be classified into Π_2. Therefore, Ω must be partitioned so that the probability of misclassification is minimized.

There are two possible misclassification errors that could occur, and their probabilities are denoted by α_1 and α_2, where

$$\alpha_1 = \mathrm{P}(\underset{\sim}{x} \text{ is assigned to } \Pi_2 \text{ when it actually belongs to } \Pi_1)$$
$$\text{and}$$
$$\alpha_2 = \mathrm{P}(\underset{\sim}{x} \text{ is assigned to } \Pi_1 \text{ when it actually belongs to } \Pi_2).$$

Since there are only two populations, these are clearly the only two possible probabilities of misclassification. The sum of the probabilities of the two errors of misclassification is then equivalent to the total probability of misclassification, denoted by TPM = $\alpha_1 + \alpha_2$. We assume the severity of a certain misclassification is equal to the severity of the reverse misclassification, so the TPM is subject to the condition $\alpha_1 = \alpha_2$.

Let $\mathrm{f}_i(\underline{\mathrm{x}})$ be the probability density function of $\underline{\mathrm{x}}$ under Π_i, where i=1,2. Hence the probability density function of $\underline{\mathrm{x}}$ is $\mathrm{f}_1(\underline{\mathrm{x}})$ if it is from Π_1, and the probability density function of $\underline{\mathrm{x}}$ is $\mathrm{f}_2(\underline{\mathrm{x}})$ if it is from Π_2. Recalling the definitions of α_1 and α_2, it is clear that $\alpha_1 = \int_{R_2}$ $\mathrm{f}_1(\underline{\mathrm{x}})$ d$\underline{\mathrm{x}}$ and $\alpha_2 = \int_{R_1}$ $\mathrm{f}_2(\underline{\mathrm{x}})$ d$\underline{\mathrm{x}}$. It should be intuitively appealing that when c is a constant chosen such that $\alpha_1 = \alpha_2$, the optimal partition that will minimize α will be given by $\mathrm{R}_1 \equiv \{\underline{\mathrm{x}} \mid (\lambda = \frac{f_1(\underline{\mathrm{X}})}{f_2(\underline{\mathrm{X}})}) > c\}$ and $\mathrm{R}_2 \equiv \{\underline{\mathrm{x}} \mid (\lambda = \frac{f_1(\underline{\mathrm{X}})}{f_2(\underline{\mathrm{X}})}) \leq c\}$.

It can then be shown that $2\ln\lambda = 2\underline{\delta}'\Sigma^{-1}\underline{\mathrm{x}}$ - $(\underline{\mu_1} + \underline{\mu_2})'\Sigma^{-1}\underline{\delta}$, where $\underline{\delta} = (\underline{\mu_1}$ - $\underline{\mu_2})$. Through careful computation, this equation provides us with a way to show the optimal regions for R_1, which can be expressed as a linear function by $\underline{\mathrm{a}}'\underline{\mathrm{x}} >$ h, where h is a constant and $\underline{\mathrm{a}}' = \underline{\delta}'\Sigma^{-1}$. It then follows that the Linear Discriminant Function Rule classifies $\underline{\mathrm{x}}$ into Π_1 if $\underline{\mathrm{a}}'\underline{\mathrm{x}}$ $>$ h and classifies $\underline{\mathrm{x}}$ into Π_2 if $\underline{\mathrm{a}}'\underline{\mathrm{x}} \leq$ h. From this, the probabilities of misclassification can be simplified to:

$$\begin{gathered}\alpha_1 = \alpha_2\\ \mathrm{P}(\underline{\mathrm{x}} \in \mathrm{R}_2 \mid \underline{\mathrm{x}} \in \Pi_1) = \mathrm{P}(\underline{\mathrm{x}} \in \mathrm{R}_1 \mid \underline{\mathrm{x}} \in \Pi_2)\\ \mathrm{P}(\underline{\mathrm{a}}'\underline{\mathrm{x}} \leq \mathrm{h} \mid \underline{\mathrm{x}} \in \Pi_1) = \mathrm{P}(\underline{\mathrm{a}}'\underline{\mathrm{x}} > \mathrm{h} \mid \underline{\mathrm{x}} \in \Pi_2).\end{gathered}$$

An equivalent method to the Linear Discriminant Function Rule is the Mahalanobis Distance Rule. That is,

$$\begin{gathered}\text{assign } \underline{\mathrm{x}} \text{ to } \Pi_1 \text{ if } \underline{\mathrm{a}}'\underline{\mathrm{x}} > \mathrm{h} = \text{assign } \underline{\mathrm{x}} \text{ to } \Pi_1 \text{ if } (\underline{\mathrm{x}} - \underline{\mu_1})'\Sigma^{-1}(\underline{\mathrm{x}} - \underline{\mu_1}) < (\underline{\mathrm{x}} - \underline{\mu_2})'\Sigma^{-1}(\underline{\mathrm{x}} - \underline{\mu_2})\\ \text{and}\\ \text{assign } \underline{\mathrm{x}} \text{ to } \Pi_2 \text{ if } \underline{\mathrm{a}}'\underline{\mathrm{x}} \leq \mathrm{h} = \text{assign } \underline{\mathrm{x}} \text{ to } \Pi_2 \text{ if } (\underline{\mathrm{x}} - \underline{\mu_1})'\Sigma^{-1}(\underline{\mathrm{x}} - \underline{\mu_1}) \geq (\underline{\mathrm{x}} - \underline{\mu_2})'\Sigma^{-1}(\underline{\mathrm{x}} - \underline{\mu_2}).\end{gathered}$$

The Mahalanobis squared distance is the squared distance between $\underline{\mathrm{x}}$ and $\underline{\mu_i}$ and is denoted $\mathrm{D}_i{}^2$, i=1,2 with $\mathrm{D}_1{}^2 =(\underline{\mathrm{x}}$ - $\underline{\mu_1})'\Sigma^{-1}(\underline{\mathrm{x}}$ - $\underline{\mu_1})$ and $\mathrm{D}_2{}^2 = (\underline{\mathrm{x}}$ - $\underline{\mu_2})'\Sigma^{-1}(\underline{\mathrm{x}}$ - $\underline{\mu_2})$. The Mahalanobis distance rule then classifies $\underline{\mathrm{x}}$ into the population to which it is closest. Hence, $\underline{\mathrm{x}}$ will be classified into Π_1 if $\mathrm{D}_1{}^2 < \mathrm{D}_2{}^2$; otherwise $\underline{\mathrm{x}}$ will be classified into Π_2.

The Mahalanobis distance can be generalized from the case with two populations into the case with k populations. We can generalize this to k groups by letting $\Pi_1, \Pi_2, ..., \Pi_k$ be k p-variate normal populations with mean vectors $\underline{\mu_1}$, $\underline{\mu_2}$, . . . , $\underline{\mu_k}$ and common variance-covariance matrix Σ. We can then in the same manner assign $\underline{\mathrm{x}}$ to Π_i if $\mathrm{D}_i{}^2 = \min\{\mathrm{D}_1{}^2, \mathrm{D}_2{}^2$, . . . , $\mathrm{D}_k{}^2\}$.

5.2 Results

State desirability can be seen as dependent upon what a state can economically, socially, educationally, and environmentally provide to its citizens. Moreover, intuition tells us that a state with an accelerated growth rate or, conversely, a moderate growth rate is more or less appealing to its inhabitants. This idea lends itself to the multivariate statistical technique of discriminant analysis because each state can be classified into one of two populations which are determined by the population growth rate and based on the measured variables.

In creating our discriminant model, we used the variables listed in Table 2.1, excluding sales tax. When the factor analysis model was constructed, the sales tax variable did not have correlations of a high magnitude with any of the factors. This implies that sales tax does not have much effect on desirability, and hence we removed it as one of the variables used to create our discriminant model.

First we established a training sample, which is a set of known observations; in our case these are the 48 states. Using Minitab, we created a discriminant function which classifies our training sample into one of two populations. The two populations were separated based upon a state's population growth rate from 1990 to 2000. Because we assume that a state with a high growth rate is more desirable to live in than a state with a low growth rate, we classified a state with a growth rate above the the national growth rate as belonging to the "Above" group, and we classified a state with a growth rate below the national growth rate as belonging to the "Below" group. Therefore, we consider the states in the "Above" group to be more desirable to live in than the states in the "Below" group. Table 5.1 is a summary of how the discriminant function classified the training sample.

Table 5.1 Summary of Classification

	Above (True Group)	Below (True Group)
Classified as Above	16	0
Classified as Below	2	30
Total States	18	30
States correct	16	30
Proportion Correct	0.899	1.000
States = 48	States Correct = 46	Proportion Correct = 0.958
		Squared Distance Between Groups = 9.00237

As illustrated in Table 5.1, our training sample contained eighteen states in the "Above" group and thirty states in the "Below" group. Two states from the "Above" group were misclassified as coming from the "Below" group.

The squared distance between groups in Table 5.1, denoted $D_p{}^2$, is the distance between the means of each population and determines if the function discriminates well. When the squared distance is large, the two groups are further apart, thus reducing their overlap. A large squared distance between groups also implies that it is less likely that an observation will be misclassified. Similarly, when the squared distance is small, the two groups are closer together, which increases their area of intersection and creates a greater likelihood that an observation will be misclassified. We can also apply the notion of the squared distance between groups to further justify the exclusion of sales tax from the discriminant model. When we included sales tax in the model, the squared distance between the groups was 9.10322. When sales tax was removed from the model, the squared distance between groups was 9.00237. The model exclusive of sales tax is an appropriate model since the change in the squared distance is very minimal in exchange for taking away one of the variables. Therefore, these two models are basically equal, so we chose to use the model exclusive of sales tax.

The apparent error rate (AER) is the percent of misclassified observations from the training sample and gives a crude estimate for the total probability of misclassification (TPM). The AER = 1 - Proportion Correct = 1 - 0.958 = 0.042, indicating that an observation of unknown group assignment has a misclassification rate of 4.2%. The minimum TPM for a population based on the training sample can be determined when the squared distance between groups in the training sample is known. Recalling that for N(0, 1) the cumulative distribution function is $\Phi(\mathcal{Z}*) = \mathrm{P}(\mathrm{Z} < \mathcal{Z}*)$, we can then calculate the minimum total probability of misclassification by using TPM $= \alpha = 2\Phi(-\frac{1}{2}D_p)$. From this method, the minimum TPM of a population based on our model is 13.36%.

With the sales tax variable in the model, the AER remained the same, as did the number of misclassifications. Using the apparent error rate and the squared distance between groups, we were able to justify that our model discriminates well. That is, our model will misclassify states with an unknown desirability level only 4.2% of the time.

Our discriminant model successfully assigns states with an unknown population growth rate over a ten year period into either the “Above” group or the “Below” group based upon the fifteen response variables used for factor analysis, excluding sales tax. After creating the discriminant function with the training sample of 48 states, we applied our model to 1996 data (see Appendix B) from 12 states. Using our discriminant model, we then predicted desirability, and hence growth rate, for each state in 2006. Table 5.2 shows which group each state was assigned to, the squared distance between the observation and each population, and the corresponding probabilities.

Table 5.2 Predictions for Test Observations

Observation	State	Predicted Group	From Group	Squared Distance	Probability
1	California	Below	Above	63.814	0.109
			Below	59.610	0.891
2	Georgia	Above	Above	52.952	0.987
			Below	61.643	0.013
3	Illinois	Below	Above	67.140	0.000
			Below	45.716	1.000
4	Louisiana	Below	Above	49.987	0.013
			Below	41.297	0.987
5	Maryland	Below	Above	36.704	0.020
			Below	28.952	0.980
6	Minnesota	Below	Above	98.143	0.000
			Below	78.657	1.000
7	North Carolina	Above	Above	34.171	0.988
			Below	42.986	0.012
8	Ohio	Below	Above	52.599	0.001
			Below	37.960	0.999
9	Oregon	Above	Above	53.115	0.935
			Below	58.446	0.065
10	Texas	Above	Above	33.145	0.989
			Below	42.206	0.011
11	West Virginia	Below	Above	74.597	0.000
			Below	59.010	1.000
12	Wisconsin	Below	Above	53.648	0.000
			Below	38.105	1.000

Minitab uses the Mahalanobis Squared Distance Rule to classify the test observations into the appropriate group. For each test observation, the discriminant function measures two Mahalanobis squared distances, one for the “Below” group and one for the “Above” group. These measurements are in the column named “Squared Distance” in Table 5.2. The states are classified according to which Mahalanobis squared distance is the smallest. The smaller the Mahalanobis squared distance, the closer the experimental unit is to the group mean, therefore, it is more likely to belong to that group. For example, the eighth state, Ohio, was classified to the “Below” group since the Mahalanobis squared distance from the “Below” group, 37.960, is smaller than the Mahalanobis squared distance from the “Above” group, 52.599. Minitab also gives the probability that each experimental unit actually belongs to the predicted group. For example, Ohio has a probability of 0.001 of belonging to the “Above” group and a probability of 0.999 of belonging to the “Below” group.

Therefore, based on our discriminant model, we predict that Georgia, North Carolina, Oregon, and Texas will all have population growth rates above the national growth rate from 1996 to 2006, and hence are more desirable states. California, Illinois, Louisiana, Maryland, Minnesota, Ohio, West Virginia, and Wisconsin are predicted to have growth rates below that of the nation’s from 1996 to 2006 and are therefore classified to be seen as less desirable states.

5.3 Discussion

Our model used fourteen variables in discriminant analysis to determine state desirability. However, some of these variables allow for discrepancies to occur. While most variables would remain within the same range for any given year, such as average temperature and difference in elevation, there are others such as personal income per capita that would vary due to inflation, and therefore would not be accurate indicators of desirability. To compensate for this in future models, it would be necessary for an inflation coefficient to be determined and applied to the training sample before the discriminant model can be applied to any unknown observations. Another issue to consider is that although we were looking at desirability based on how much a population was increasing compared to the national growth rate, the national growth rate will also change each year. Therefore, the model would be more accurate if it is changed to account for the variance in the national growth rate.

6 Conclusion

Using these same methods, state desirability can be analyzed in the future. While the same variables could be used, it would be interesting to see if different variables could be used to create a more efficient model. In 2006, data could be collected to determine if we were correct in predicting the desirability level of the twelve states we classified. Also, using data from the 2000 census, another study of the states could be performed to predict state desirability for the first decade of the new millennium. Another application of these methods would be to describe and predict desirability of different countries. Certainly some new variables would have to be considered, but the process and application would be similar.

7 Appendix A

Appendix A.1 Economic Variables

State	State Health and Hospital Spending per Capita [9]	Percent of Population Below Poverty Level [11]	Percent of Population Unemployed[17]
Alabama	221.01	19.2	6.8
Alaska	319.97	11.4	6.9
Arizona	86.22	13.7	5.3
Arkansas	139.53	19.6	6.9
California	185.48	13.9	5.6
Colorado	112.92	13.7	4.9
Connecticut	290.53	6.0	5.1
Delaware	220.67	6.9	5.1
Florida	138.04	14.4	5.9
Georgia	147.42	15.8	5.4
Hawaii	280.63	11.0	2.8
Idaho	79.46	14.9	5.8
Illinois	121.87	13.7	6.2
Indiana	128.78	13.0	5.3
Iowa	191.59	10.4	4.2
Kansas	153.38	10.3	4.4
Kentucky	119.66	17.3	5.8
Louisiana	200.95	23.6	6.2
Maine	131.93	13.1	5.1
Maryland	173.80	9.9	4.6
Massachusetts	271.76	10.7	6.0
Michigan	267.98	14.3	7.1
Minnesota	186.28	12.0	4.8
Mississippi	132.13	25.7	7.5
Missouri	132.11	13.4	5.7
Montana	121.39	16.3	5.8
Nebraska	182.46	10.3	2.2
Nevada	78.21	9.8	4.9
New Hampshire	128.92	6.3	5.6
New Jersey	160.02	9.2	5.0
New Mexico	210.55	20.9	6.3
New York	278.93	14.3	5.2
North Carolina	144.83	13.0	4.1
North Dakota	134.63	13.7	3.9
Ohio	146.12	11.5	5.7
Oklahoma	168.81	15.6	5.6
Oregon	176.62	9.2	5.5
Pennsylvania	140.30	11.0	5.4
Rhode Island	255.12	7.5	6.7
South Carolina	225.71	16.2	4.7
South Dakota	129.31	13.3	3.7
Tennessee	132.45	16.9	5.2
Texas	107.03	15.9	6.2
Utah	193.86	8.2	4.3
Vermont	117.28	10.9	5.0
Virginia	204.45	11.1	4.3
Washington	163.97	8.9	4.9
West Virginia	107.05	18.1	8.3
Wisconsin	141.46	9.3	4.4
Wyoming	205.03	11.0	5.4

Appendix A.1 continued Economic Variables Continued

State	General Spending per Capita [9]	Personal Income per Capita [3]	Percent Sales Tax [10]
Alabama	1831	14998	4.00
Alaska	7790	21646	0.00
Arizona	2056	16006	5.00
Arkansas	1672	14176	4.00
California	2359	20689	5.00
Colorado	1708	18860	3.00
Connecticut	2702	25395	8.00
Delaware	2994	20095	0.00
Florida	1589	18539	6.00
Georgia	1759	17045	4.00
Hawaii	3201	20361	4.00
Idaho	1818	15250	5.00
Illinois	1754	20433	6.25
Indiana	1802	16921	5.00
Iowa	2137	17301	4.00
Kansas	1747	18104	4.25
Kentucky	1927	14992	6.00
Louisiana	2020	14528	4.00
Maine	2234	17183	5.00
Maryland	2057	21857	5.00
Massachusetts	2832	22555	5.00
Michigan	2104	18378	4.00
Minnesota	2379	18731	6.00
Mississippi	1708	12830	6.00
Missouri	1505	17479	4.23
Montana	2066	15304	0.00
Nebraska	1784	17490	5.00
Nevada	1968	19049	5.75
New Hampshire	1512	20773	0.00
New Jersey	2334	24881	7.00
New Mexico	2568	14254	5.00
New York	2763	22129	4.00
North Carolina	1894	16266	3.00
North Dakota	2483	15355	5.00
Ohio	1889	17568	5.00
Oklahoma	1784	15451	4.50
Oregon	1957	17182	0.00
Pennsylvania	1787	18679	6.00
Rhode Island	2741	18809	7.00
South Carolina	1943	15141	5.00
South Dakota	1841	15890	4.00
Tennessee	1616	15868	5.50
Texas	1391	16717	6.25
Utah	2014	13985	5.00
Vermont	2603	14506	4.00
Virginia	1920	19701	3.50
Washington	2340	18777	6.50
West Virginia	1969	13744	6.00
Wisconsin	2146	17590	5.00
Wyoming	3270	16283	3.00

Appendix A.2 Demographic Variables

State	Population Density [8]	Percent Minorities [6]	Median Age [3]	Crime Rate [19]
Alabama	79.6	27	33.0	4915
Alaska	1.0	26	29.4	5153
Arizona	32.3	28	32.2	7889
Arkansas	45.1	18	33.8	4867
California	190.8	43	31.5	6604
Colorado	31.8	19	32.5	6054
Connecticut	678.4	16	34.4	5387
Delaware	340.8	21	32.9	5360
Florida	239.6	27	36.4	8811
Georgia	111.9	30	31.6	6764
Hawaii	172.5	69	32.6	6107
Idaho	12.2	8	31.5	4057
Illinois	205.6	25	32.8	5935
Indiana	154.6	10	32.8	4683
Iowa	49.7	4	34.0	4101
Kansas	30.3	12	32.9	5193
Kentucky	92.8	8	33.0	3299
Louisiana	96.9	34	31.0	6487
Maine	39.8	2	33.9	3698
Maryland	489.2	30	33.0	5830
Massachusetts	767.6	12	33.6	5298
Michigan	163.6	18	32.6	5995
Minnesota	55.0	6	32.5	4539
Mississippi	54.9	37	31.2	3869
Missouri	74.3	13	33.5	5121
Montana	5.5	8	33.8	4502
Nebraska	20.5	7	33.0	4213
Nevada	10.9	21	33.3	6064
New Hampshire	123.7	3	32.8	3645
New Jersey	1042.0	26	34.5	5447
New Mexico	12.5	50	31.3	6684
New York	381.0	31	33.9	6364
North Carolina	136.1	25	33.1	5486
North Dakota	9.3	6	32.4	2922
Ohio	264.9	13	33.3	4843
Oklahoma	45.8	19	33.2	5599
Oregon	29.6	9	34.5	5646
Pennsylvania	265.1	12	35.0	3476
Rhode Island	960.3	11	34.0	5353
South Carolina	115.8	31	32.0	6045
South Dakota	9.2	9	32.5	2909
Tennessee	118.3	17	33.6	5051
Texas	64.9	39	30.8	7827
Utah	21.0	9	26.2	5660
Vermont	60.8	2	33.0	4341
Virginia	156.3	24	32.6	4441
Washington	73.1	13	33.1	6223
West Virginia	74.5	4	35.4	2503
Wisconsin	90.1	9	32.9	4395
Wyoming	4.7	9	32.0	4211

Appendix A.3 Education Variables

State	State Education Spending per Capita [9]	Graduation Rate (Percent) [1]
Alabama	836.51	64.7
Alaska	1921.66	68.4
Arizona	752.75	72.5
Arkansas	717.23	76.7
California	904.10	68.7
Colorado	756.13	74.1
Connecticut	662.59	74.9
Delaware	1064.30	68.5
Florida	605.12	61.1
Georgia	779.23	62.7
Hawaii	1004.31	86.8
Idaho	745.97	79.4
Illinois	567.60	76.6
Indiana	763.87	75.0
Iowa	870.80	87.5
Kansas	744.68	82.0
Kentucky	799.12	69.0
Louisiana	752.85	56.7
Maine	767.15	77.6
Maryland	599.19	72.8
Massachusetts	581.08	77.0
Michigan	690.46	70.1
Minnesota	862.61	89.7
Mississippi	731.77	64.2
Missouri	639.82	77.2
Montana	720.84	83.3
Nebraska	574.64	85.5
Nevada	703.92	77.2
New Hampshire	361.50	74.0
New Jersey	697.27	79.8
New Mexico	1109.52	68.2
New York	792.98	65.1
North Carolina	900.03	68.0
North Dakota	926.74	88.1
Ohio	711.71	74.0
Oklahoma	753.12	77.5
Oregon	608.66	71.6
Pennsylvania	587.04	79.1
Rhode Island	779.30	69.6
South Carolina	815.38	58.5
South Dakota	541.66	85.7
Tennessee	578.41	67.9
Texas	645.98	64.1
Utah	965.26	83.1
Vermont	916.91	91.6
Virginia	763.33	73.6
Washington	1044.24	77.2
West Virginia	810.72	77.3
Wisconsin	753.31	84.2
Wyoming	1155.23	78.6

Appendix A.4 Geographic Variables

State	Average Precipitation [2]	Difference in Elevation [21]	Average Temperature [22]
Alabama	56.90	2405	67.50
Alaska	53.15	20320	40.00
Arizona	7.11	12563	71.20
Arkansas	49.20	2698	61.90
California	17.28	14776	60.90
Colorado	15.31	11083	50.30
Connecticut	44.39	2380	49.80
Delaware	41.38	442	54.00
Florida	49.91	345	71.80
Georgia	48.61	4784	61.20
Hawaii	23.47	13796	77.00
Idaho	11.71	11952	51.10
Illinois	33.34	956	49.80
Indiana	39.12	937	52.10
Iowa	34.71	1190	49.70
Kansas	28.61	3360	56.40
Kentucky	43.56	3882	56.20
Louisiana	59.74	543	68.20
Maine	43.52	5267	45.00
Maryland	41.84	3360	55.10
Massachusetts	43.84	3487	51.50
Michigan	32.23	1408	44.15
Minnesota	26.36	1701	41.45
Mississippi	52.82	806	64.60
Missouri	33.91	1542	54.75
Montana	11.37	10999	44.70
Nebraska	30.34	4584	51.10
Nevada	4.19	12661	49.40
New Hampshire	36.53	6288	45.30
New Jersey	41.93	1803	53.10
New Mexico	8.91	10319	56.20
New York	39.28	5344	49.80
North Carolina	42.46	6684	59.50
North Dakota	15.36	2756	41.30
Ohio	37.77	1094	51.57
Oklahoma	30.89	4684	59.90
Oregon	37.39	11239	53.00
Pennsylvania	40.26	3213	52.30
Rhode Island	41.91	812	50.30
South Carolina	51.59	3560	63.30
South Dakota	17.47	6276	45.30
Tennessee	48.49	6465	60.50
Texas	34.70	8749	65.90
Utah	15.31	11528	51.70
Vermont	33.69	4298	44.10
Virginia	45.22	5729	57.70
Washington	27.66	14410	49.30
West Virginia	40.74	4621	54.80
Wisconsin	30.89	1372	46.10
Wyoming	13.31	10705	45.70

8 Appendix B

Appendix B.1 Economic Variables

State	State Health and Hospital Spending per Capita [13]	Percent of Population Below Poverty Level [15]	Percent of Population Unemployed [14]
California	261.75	16.9	7.2
Georgia	175.30	14.8	4.6
Illinois	194.99	12.1	5.3
Louisiana	343.37	20.5	6.7
Maryland	211.36	10.3	4.9
Minnesota	239.80	9.8	4.0
North Carolina	215.35	12.2	4.3
Ohio	202.99	12.7	4.9
Oregon	252.50	11.8	5.9
Texas	160.97	16.6	5.6
West Virginia	111.17	18.5	7.5
Wisconsin	187.79	8.8	3.5

Appendix B.1 continued Economic Variables Continued

State	General Spending per Capita [13]	Personal Income per Capita [16]
California	3101	25368
Georgia	2535	23028
Illinois	2540	26855
Louisiana	2901	19709
Maryland	2639	27676
Minnesota	3388	25699
North Carolina	2656	22244
Ohio	2552	23493
Oregon	3014	23111
Texas	2175	22324
West Virginia	3089	18225
Wisconsin	2964	23390

Appendix B.2 Demographic Variables

State	Population Density [12]	Percent Minorities [7]	Median Age [4]	Crime Rate [20]
California	204.4	48	33	5208
Georgia	127.0	32	33	6310
Illinois	213.1	28	34	5316
Louisiana	99.9	36	33	6839
Maryland	518.8	34	35	6062
Minnesota	58.5	08	35	4463
North Carolina	150.3	25	35	5526
Ohio	272.8	14	35	4456
Oregon	33.4	11	36	5997
Texas	73.0	43	33	5709
West Virginia	75.8	05	38	2483
Wisconsin	95.0	10	35	3821

Appendix B.3 Education Variables

State	State Education Spending per Capita [13]	Graduation Rate (Percent) [18]
California	1091.03	65.3
Georgia	1078.88	55.0
Illinois	740.53	80.0
Louisiana	999.31	57.9
Maryland	798.50	73.9
Minnesota	1241.52	85.3
North Carolina	1099.41	62.4
Ohio	923.21	70.6
Oregon	1083.96	66.6
Texas	891.31	58.4
West Virginia	1138.55	76.1
Wisconsin	1031.78	80.4

Appendix B.4 Geographic Variables

State	Average Precipitation [2]	Difference in Elevation [21]	Average Temperature [22]
California	15.31	11083	50.3
Georgia	48.61	4784	61.2
Illinois	33.34	956	49.8
Louisiana	59.74	543	68.2
Maryland	41.84	3360	55.1
Minnesota	32.23	1408	44.2
North Carolina	42.46	6684	59.5
Ohio	37.77	1094	51.6
Oregon	37.39	11239	53.0
Texas	34.70	8749	65.9
West Virginia	40.74	4621	54.8
Wisconsin	30.89	1372	46.1

Acknowledgements

We would like to thank Dr. Vasant Waikar for his classroom instruction and assistance throughout all stages of this project. We would also like to thank our graduate assistant Lindsay Kellam for helping us learn and understand multivariate analysis and Minitab, for helping us revise this paper, and for her support throughout the entire program. Special thanks to Ms. Moira Miller for her help throughout the writing process and Dr. Patrick Dowling for his assistance in formatting this final paper. Thank you to all SUMSRI faculty, staff, and participants for making this summer a fun and educational experience. Finally, we wish to express our gratitude to Miami University, the National Science Foundation, and the National Security Agency for sponsoring this program.

References

[1] M. S. Hoffman (ed.), National Center for Educational Statistics: U.S. Department of Education, *Public High School Graduation and Dropout Rates, 1990*, in *The World Almanac and Book of Facts 1993 Edition*, Pharos Books, New York, (1992), 196.

[2] M. S. Hoffman (ed.), National Climatic Data Center, NESDIS, NOAA, U.S. Department of Commerce, *Normal Temperatures, Highs, Lows, Precipitation*, in *The World Almanac and Book of Facts 1992 Edition*, Pharos Books, New York, (1991), 208.

[3] E. R. Hornor (ed.), *Almanac of the 50 States 1993 Edition*, Information Publications, CA, (1993).

[4] E. R. Hornor (ed.), *Almanac of the 50 States 1999 Edition*, Information Publications, CA, (1999).

[5] K. A. Hovey and H. A. Hovey, *A-8 Median Age, 1996*, *CQ's State Fact Finder*, Congressional Quarterly Inc., Washington, D.C., (1998).

[6] U.S. Census Bureau , *Demographic Profiles: Census 2000*, in *United States Census 2000*, found http://www.census.gov/Press-Release/www/2001/demoprofile.html.

[7] U.S. Bureau of the Census, *(ST-99-29) Population Estimates for States by Race and Hispanic Origin: July 1, 1996*, http://www.census.gov/population/estimates/state/srh/srh96.txt.

[8] U.S. Bureau of the Census, *No. 31. Resident Population- States: 1970 to 1992*, in *Statistical Abstract of the United States 112th edition* (compiled by U.S. Department of Commerce *et al*), U.S. Government Printing Office, Washington, D.C., (1993), 28-29.

[9] U.S. Bureau of the Census, *No. 460. State Governments- Revenue, Debt, and Expenditure 1970-1990, and by State 1990*, in *Statistical Abstract of the United States 112th edition* (compiled by U.S. Department of Commerce *et al*), U.S. Government Printing Office, Washington, D.C., (1992), 286-287.

[10] U.S. Bureau of the Census, *No. 463. State Government Tax Collections and Excise Taxes, 1970-90, and by State, 1990*, in *Statistical Abstract of the United States 112th edition* (compiled by U.S. Department of Commerce *et al*), U.S. Government Printing Office, Washington, D.C., (1992), 290.

[11] U.S. Bureau of the Census, *No. 723. Percent of Persons Below Poverty Level, by State: 1984 to 1990*, in *Statistical Abstract of the United States 112th edition* (compiled by U.S. Department of Commerce *et al*), U.S. Government Printing Office, Washington, D.C., (1992), 458.

[12] U.S. Bureau of the Census, *No. 26. Resident Population- States: 1970 to 1996*, in *Statistical Abstract of the United States 117th edition* (compiled by U.S. Department of Commerce *et al*), U.S. Government Printing Office, Washington, D.C., (1997), 28.

[13] U.S. Bureau of the Census, *No. 516. State Governments- Expenditure and Debt, by State: 1996*, in *Statistical Abstract of the United States 118th edition* (compiled by U.S. Department of Commerce *et al*), U.S. Government Printing Office, Washington, D.C., (1998), 320.

[14] U.S. Bureau of the Census, *No. 649. Characteristics of the Civilian Labor Force, by State: 1996*, in *Statistical Abstract of the United States 118th edition* (compiled by U.S. Department of Commerce *et al*), U.S. Government Printing Office, Washington, D.C., (1998), 406.

[15] U.S. Bureau of the Census, *No. 761. Persons Below Poverty Level, by State: 1980 to 1996*, in *Statistical Abstract of the United States 118th edition* (compiled by U.S. Department of Commerce *et al*), U.S. Government Printing Office, Washington, D.C., (1998), 479.

[16] U.S. Bureau of Economic Analysis, *No. 727. Personal Income per Capita in Current and Constant (1992) Dollars, by State: 1990 to 1997*, in *Statistical Abstract of the United States 118th edition* (compiled by U.S. Department of Commerce *et al*), U.S. Government Printing Office, Washington, D.C., (1998), 460.

[17] U.S. Bureau of Labor Statistics, *No.641. Total Unemployed and Insured Unemployed-States: 1980-1990*, in *Statistical Abstract of the United States 112th edition* (compiled by U.S. Department of Commerce *et al*), U.S. Government Printing Office, Washington, D.C., (1992), 402.

[18] U.S. Department of Education, *Public High School Graduates, 1995-1996*, in *The World Almanac and Book of Facts 2001*, World Almanac Books, New Jersey, (2001), 242.

[19] U.S. Federal Bureau of Investigation, *No. 289. Crime Rates by State, 1985 to 1990, and by Type, 1990*, in *Statistical Abstract of the United States 112th edition* (compiled by U.S. Department of Commerce *et al*), U.S. Government Printing Office, Washington, D.C., (1992), 181.

[20] U.S. Federal Bureau of Investigation, *No. 337. Crime Rates, by State, 1994 to 1996, and by Type, 1996*, in *Statistical Abstract of the United States 118th edition* (compiled by U.S. Department of Commerce *et al*), U.S. Government Printing Office, Washington, D.C., (1998), 211.

[21] U.S. Geological Survey, *No. 345. Extreme and Mean Elevations States and Other Areas*, in *Statistical Abstract of the United States 112th edition* (compiled by U.S. Department of Commerce *et al*), U.S. Government Printing Office, Washington, D.C., (1998), 320.

[22] U.S. National Oceanic and Atmospheric Administration, *No. 367. Normal Daily Mean Temperatures- Selected Cities*, in *Statistical Abstract of the United States 112th edition* (compiled by U.S. Department of Commerce *et al*), U.S. Government Printing Office, Washington, D.C., (1992), 219.