Where is the Third World?
A Multivariate Statistical Analysis of World Development

Virginia Ahalt - Clemson University, Clemson, SC, 29634
Ramone Gordon - Prairie View A & M University, Prairie View, TX, 77446
Dusti Nisbet - North Georgia College & State University, Dahlonega, GA, 30597
Lauren Vollmer - Middlebury College, Middlebury, VT, 05753

July 15, 2009

Abstract

We analyze a set of 13 variables on a sample of 170 countries. After normalizing the data, we perform Principal Component Analysis and Factor Analysis to reveal its themes and underlying factors. We then classify each of the 170 countries as developed, developing, or underdeveloped based on two Discriminant Analyses. The first discrimination rule uses a published list of the current world divisions and the second discrimination rule relies on divisions based on the first principal component we generate in our Principal Component Analysis. We compare the two analyses and conclude that our classification system more accurately describes the current state of world development.

1 Introduction

The term "third world country" first appeared in the French newspaper *L'Observateur* in August 1952, describing countries of South America, Asia, and Africa that adopted a policy of non-alignment during the Cold War. In the half-century since the term's coining it has come to suggest poverty and low human development, a connotation at odds with its ideological definition. Although the current usage of the term "third world country" contrasts with its original definition, the term's modern implications of underdevelopment are more relevant than its association with outdated political conflicts. Similarly, the "developed" and "developing" first and second worlds no longer denote capitalist and communist ideologies. We aim to reclassify the countries based on a holistic vision of current development, placing countries into categories that reflect economic and social realities rather than obsolete political identifications. We use the more politically correct terms "developed," "developing," and "underdeveloped" to represent the first, second, and third worlds, respectively.

In our research we analyze 170 countries and 13 variables. In order to consider a complete and current picture of development, we examine variables from the following categories: health, education, freedom, technology, economy, and living environment (refer to Table 1 for a full list of variables). We obtained data from various online compilations based on the CIA World Factbook and several UN databases.

Table 1: Variables

Number of Physicians (per 10,000 people)
GDP per Capita (PPP) (USD)
Urbanization of Population (%)
Compulsory Education (Years)
Literacy (%)
Internet Users (per 1,000 people)
Life Expectancy (at Birth: Years)
Inflation Rate (Consumer Prices) (%)
Press Freedom (Scale 100; 1 is free, 100 is not free)
Combined Gross Enrollment (P/S/T Ed) (%)
Health Spending per Capita (USD)
Mobile Phone Subscribers (per 1,000,000 people)
Energy Consumption per Capita (kWh)

Our analysis includes three multivariate statistical techniques: Principal Component Analysis, Factor Analysis, and Discriminant Analysis. Principal Component Analysis allows us to describe the themes of the data set by forming linear combinations of the original variables. Factor Analysis reveals the underlying, unobservable trends affecting each variable. Finally, Discriminant Analysis, the most significant technique our research, classifies the new observations into populations based on a discrimination function. This function uses training samples selected at random from the different populations from the original data set. We employ this discrimination technique to divide the 170 countries into the categories "developed," "developing," and "underdeveloped," thereby establishing an up-to-date, objective distribution system to replace the outdated first, second, and third world divisions.

2 Multivariate Statistical Techniques

2.1 Assessing Normality

Principal Component Analysis, Factor Analysis, and Discriminant Analysis all depend on the assumption of a multivariate normal distribution. Before proceeding with these techniques we must first verify that our data set has a multivariate normal distribution. We assess multivariate normality by testing the univariate normality of each variable.

A standard test for normality is the Quantile-Quantile plot test, or Q-Q plot test. A Q-Q plot is a graph of the observed versus the expected quantile of the data under an assumption of normality. Under normality, the Q-Q plot follows an equation of the form $x_{(i)} = \mu + \sigma z_{(i)}$, where $x_{(i)}$ is the ordered sample data point and $z_{(i)}$ is the corresponding normal quantile. This equation represents a linear relationship, so if the plot is linear upon visual inspection, there is evidence to suggest that the data set is normally distributed.

Although an apparently linear Q-Q plot points to a normal distribution, a visual inspection is not sufficient to establish normality. We perform a hypothesis test to verify the linearity. Consider, for example, a Q-Q plot for the variable Health Spending per Capita (Figure 1).

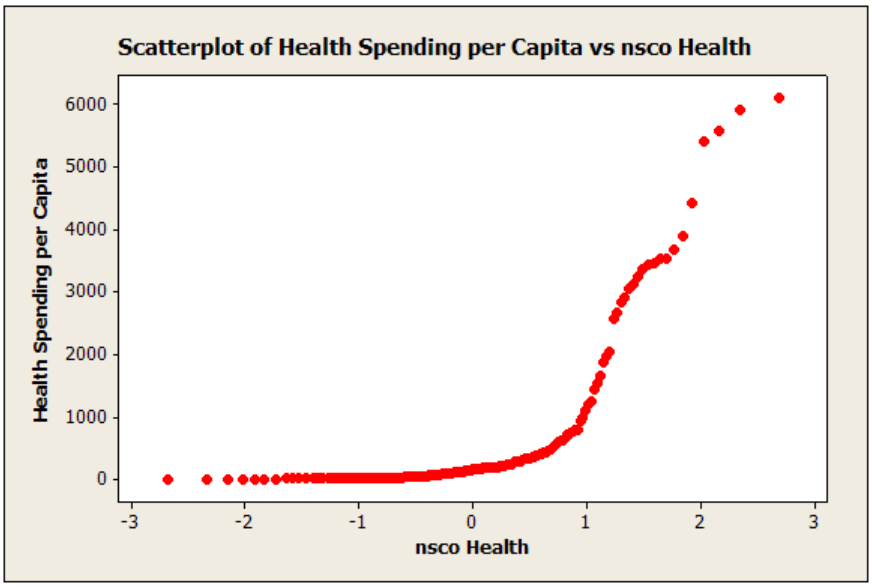

Figure 1: Sample Q-Q Plot (Health Spending per Capita)

It is evident that the plot in Figure 1 is not linear, but we must support the visual indication of non-normality with the numerical results of a hypothesis test.

In hypothesis testing of normality, we take as our null hypothesis (H_0) that the population correlation coefficient ρ is equal to 1, implying normality. The alternative hypothesis (H_1) states that the population correlation coefficient ρ is less than 1, which implies non-normality. The value of ρ is the determined by the relationship between x and z and indicates the positive slope of the ordered pairs. Our test statistic r is the sample correlation coefficient, an estimate of ρ, and thus measures the strength and linearity of the relationship between $x_{(i)}$ and $z_{(i)}$. The sample correlation coefficient is given by

$$0 < r = \frac{\hat{Cov}(x_{(i)}, z_{(i)})}{\sqrt{\hat{Var}(x_{(i)})\hat{Var}(z_{(i)})}} = \frac{\sum_{i=1}^{n}(x_{(i)} - \overline{x})(z_{(i)} - \overline{z})}{\sqrt{\sum_{i=1}^{n}(x_{(i)} - \overline{x})^2}\sqrt{\sum_{i=1}^{n}(z_{(i)} - \overline{z})^2}} < 1. \quad (1)$$

We reject H_0 if $r < c$, where c is a critical value determined by the sample size n of the data set and a chosen significance level (α). We accept H_0 if $r > c$ and therefore conclude that our sample is normally distributed. Since our sample comprises 170 countries and we use a significance level of 5%, the value of c for our test is 0.991. In our example case of Health Spending per Capita, $r = 0.754$. Since $0.754 < 0.991$, we can conclude, as we suspect based on the non-linear Q-Q plot, that this variable is not normally distributed.

If a variable is not normally distributed, we must transform it to a normal distribution before proceeding with our analysis. Transformation, a standard statistical technique, is mathematically valid because it does not affect the relative order or the relationships among the data. Possible transformations include reciprocals, roots, and powers of the variable as well as logarithms. Combinations of these transformations, like roots of logarithms or vice versa, are also valid. In the case of Health Spending per Capita, we take the square root of the natural logarithm of the variable. A Q-Q plot of this transformed variable appears normal (Figure 2).

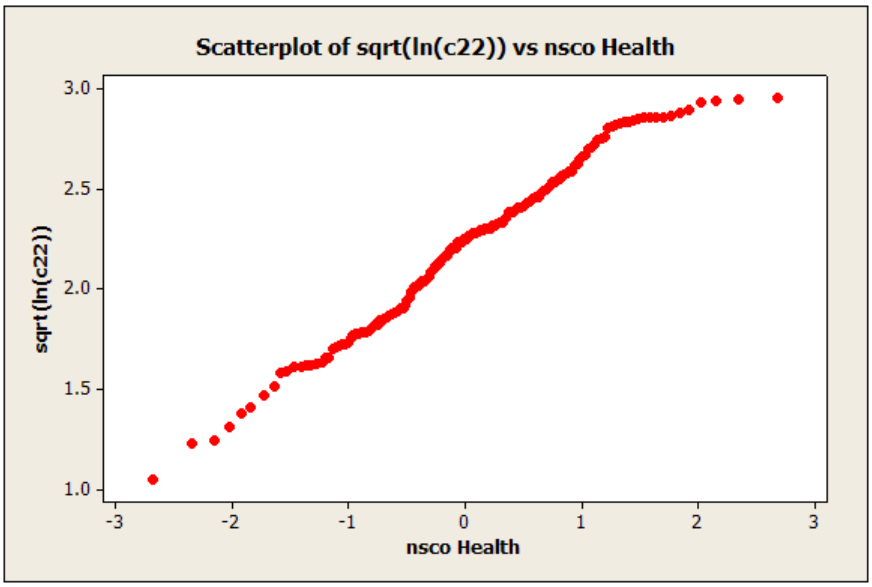

Figure 2: Sample Normalized Q-Q Plot (Health Spending per Capita)

We observe from the plot that the transformation has normalized the data, and a new correlation coefficient of $0.992 > 0.991$ confirms the visual indication of normality.

We apply tests of normality and, when necessary, transformations to each of the 13 variables we consider. A list of the variables, the transformations, and the corresponding correlation coefficients r is given in Table 2.

Table 2: r-values Before and After Transformation

Variable	r	**Transformation**	r **after Transformation**
Number of Physicians	0.967	$\sqrt{x}$	0.986
GDP per Capita	0.876	$x^{\frac{1}{6}}$	0.992
Urbanization	0.989		0.989
Years of Compulsory Education	0.969	x^2	0.996
Literacy	0.909	$(\ln(x))^{\frac{1}{5}}$	0.988
Internet Users	0.932	$x^{\frac{1}{3}}$	0.984
Life Expectancy	0.955	x^4	0.983
Inflation Rate	0.939	$x^{\frac{1}{3}}$	0.996
Press Freedom	0.98		0.98
Combined Gross Enrollment	0.985	x^2	0.993
Health Spending per Capita	0.754	$\sqrt{\ln(x)}$	0.992
Mobile Phone Subscribers	0.954	$x^{\frac{1}{3}}$	0.983
Energy Consumption	0.827	$x^{\frac{1}{6}}$	0.993

Table 2 indicates that we are unable to achieve normality, that is $r \not> 0.991$, for several of our variables. We believe that this phenomenon results from the large size of our data set and the high standard of our significance level of 5%. Although these variables do not demonstrate normality, they are essential to our understanding of national development. We note that these variables are normally distributed under a less stringent significance level, so we include them in our analysis.

2.2 Principal Component Analysis

After normalizing the 13 variables, we wish to condense them into a smaller number of principal components, facilitating data analysis. We do so using Principal Component Analysis (PCA), which takes variables $x_1, x_2, ..., x_p$ and reduces the dimensionality of the data by forming linear combinations of these variables.

The goal of PCA is to form m linear combinations $Y_1, Y_2, ..., Y_m$ using the original p variables, with m much smaller than p. These $Y_1, Y_2, ..., Y_m$ will be of the form:

$$\begin{aligned} Y_1 &= a_{11}x_1 + a_{12}x_2 + ... + a_{1p}x_p = \mathbf{a_1^T x} \\ Y_2 &= a_{21}x_1 + a_{22}x_2 + ... + a_{2p}x_p = \mathbf{a_2^T x} \\ &\vdots \\ Y_m &= a_{m1}x_1 + a_{m2}x_2 + ... + a_{mp}x_p = \mathbf{a_m^T x}, \end{aligned} \tag{2}$$

where a_{ij} is a coefficient that represents the weight of the variable x_j of component Y_i. Each component Y_i is thus a weighted average of the p variables.

We generate these principal components $Y_1, Y_2, ..., Y_m$ in such a way that the components collectively account for the maximum variance of the data set. To do so, we evaluate the variance of each component and maximize it with respect to $\mathbf{a_i}$. We note that the variance of each principal component is given by:

$$Var(Y_i) = \mathbf{a_i^T \Sigma a_i}, \tag{3}$$

where Σ is a positive definite matrix in which each entry σ_{ij} represents the covariance of x_i and x_j.

The choice of $\mathbf{a_i}$ that maximizes $Var(Y_i)$ is a normalized eigenvector of Σ that corresponds to the i^{th} largest eigenvalue λ_i of Σ. The vectors $\mathbf{a_1}, \mathbf{a_2}, ..., \mathbf{a_m}$ are subject to the constraints (for $i \neq j$)

$$\mathbf{a_i^T a_i} = 1$$

$$\mathbf{a_j a_i^T} = 0. \tag{4}$$

That is, each vector $\mathbf{a_i}$ has unit length and each pair of vectors $\mathbf{a_i}$,$\mathbf{a_j}$ is orthogonal. The variance of the i^{th} principal component $Var(Y_i) = \mathbf{a_i^T \Sigma a_i}$ is equal to λ_i. To determine the contribution of each principal component to the total variance, we compare the eigenvalue λ_i to the sum of the eigenvalues, which represents the total variance. The percentage contribution of Y_i is given by

$$Percent\,of\,Total\,Variance = \frac{\lambda_i}{\lambda_1 + \lambda_2 + ... + \lambda_p} \cdot 100. \tag{5}$$

We generate five principal components, $m = 5$, accounting for 87.6% of the total variance of the data set. The individual contributions of the components are given in Table 3.

Table 3: PCA Eigenvalues and Proportions

	PC 1 (Y_1)	PC 2 (Y_2)	PC 3 (Y_3)	PC 4 (Y_4)	PC 5 (Y_5)
Eigenvalue	8.1244	1.2721	0.7878	0.6588	0.5438
Proportion	0.625	0.098	0.061	0.051	0.042
Cumulative	0.625	0.723	0.783	0.834	0.876

The weights of the variables within each principal component are listed in Table 4.

Table 4: PCA Coefficients

	x_i	PC 1	PC 2	PC 3	PC 4	PC 5
Number of Physicians	x_1	0.291	0.228	-0.282	-0.030	0.227
GDP per Capita	x_2	0.324	0.064	0.169	0.115	0.018
Urbanization	x_3	0.261	0.267	0.390	-0.034	0.465
Compulsory Education	x_4	0.213	-0.322	-0.331	-0.651	0.251
Literacy	x_5	0.277	0.099	-0.497	-0.064	-0.201
Internet Users	x_6	0.320	0.000	0.014	0.148	-0.153
Life Expectancy	x_7	0.315	0.013	0.012	0.217	0.100
Inflation Rate	x_8	-0.148	0.646	-0.392	0.142	-0.208
Press Freedom	x_9	-0.237	0.447	0.058	-0.141	0.501
Combined Gross Enrollment	x_{10}	0.309	0.047	-0.247	0.102	0.213
Health Spending per Capita	x_{11}	0.332	-0.037	0.130	0.120	0.005
Mobile Phone Subscribers	x_{12}	0.305	0.015	0.140	0.203	-0.192
Energy Consumption	x_{13}	0.201	0.372	0.359	-0.618	-0.468

Having computed these principal components, we wish to label each component. The label is intended to convey the interrelation of the variables within that component. We choose this label based on the relative importance of each variable to the component. Evaluating the relative weights of the variables within each component leads us to assign the labels in Table 5. For example, the label "Economic Strength" suits PC1 because the significant variables, excluding Inflation Rate and Press Freedom, all contribute to and indicate economic strength. Similarly in PC2, the contributing variables are Inflation Rate and Press Freedom, which both indicate government's intervention.

Table 5: PCA Labels

	Label
PC1	Economic Strength
PC2	Government Intervention
PC3	Urban Lifestyle
PC4	Modernization
PC5	Liberalism

2.3 Factor Analysis

Principal Component Analysis allows us to condense our data set into 5 components that collectively represent the overt themes of the data. However, we also wish to identify the underlying, unobservable factors that affect the variables. We do so using Factor Analysis, which describes each of the 13 original variables as a linear combination of m factors with $m < p$ common to the data set called *common factors*, and an additional factor ε specific to that variable, called the *unique factor*.

According to the m-factor model, one of the methods of performing Factor Analysis, we represent the variables $x_1, x_2, ..., x_p$ as

$$\begin{aligned} x_1 &= \mu_1 + \ell_{11}F_1 + \ell_{12}F_2 + \ldots + \ell_{1m}F_m + \varepsilon_1 \\ x_2 &= \mu_2 + \ell_{21}F_1 + \ell_{22}F_2 + \ldots + \ell_{2m}F_m + \varepsilon_2 \\ &\vdots \\ x_p &= \mu_p + \ell_{p1}F_1 + \ell_{p2}F_2 + \ldots + \ell_{pm}F_m + \varepsilon_p \end{aligned} \tag{6}$$

where $F_1, F_2, ..., F_m$ are the common factors and ε_i is the unique factor. Each coefficient ℓ_{ik} is the factor loading of the k^{th} factor on the i^{th} variable, reflecting the importance of factor k to variable i. In matrix form, the m-factor model appears as:

$$\underline{X} - \underline{\mu} = \mathbf{L}\underline{F} + \underline{\varepsilon} \tag{7}$$

In the m-factor model, $\underline{\mu}$ is the mean and the factors $\underline{F}$ and $\underline{\varepsilon}$ are multivariate normal vectors and subject to the constraints:

1. $E(\underline{F}) = \underline{0}$
2. $\text{Cov}(\underline{F}) = \mathbf{I}$, so F_i and F_j are independent
3. $\text{Cov}(\underline{\varepsilon}) = \mathbf{\Psi} = diag(\Psi_1, \Psi_2, ..., \Psi_p)$
4. $\underline{F}$ and $\underline{\varepsilon}$ are independent.

Given the above conditions, we can factor the population covariance matrix $\mathbf{\Sigma}$ as

$$\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\mathbf{T} + \mathbf{\Psi}, \tag{8}$$

where $\mathbf{L}$ is the $p \times m$ matrix of factor loadings. That is, $\mathbf{L}$ is the matrix in which the entry in the i^{th} row and k^{th} column is the coefficient ℓ_{ik}.

Since $\mathbf{\Sigma}$ is unknown in practice, we are unable to compute the $\mathbf{L}$ and $\mathbf{\Psi}$ matrices. Thus, assuming $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\mathbf{T} + \mathbf{\Psi}$ for a given m, we obtain maximum likelihood estimates of $\mathbf{L}$ and $\mathbf{\Psi}$, which we denote $\hat{\mathbf{L}}$ and $\hat{\mathbf{\Psi}}$. We calculate these estimates by maximizing the likelihood equation

$$L(\mathbf{L}, \mathbf{\Psi}) = \frac{1}{(2\pi)^{\frac{np}{2}} |\mathbf{L}\mathbf{L}^\mathbf{T} + \mathbf{\Psi}|^{\frac{n}{2}}} e^{-\frac{1}{2}\sum_{i=1}^{n} (\underline{x}_i - \underline{\mu})^T (\mathbf{L}\mathbf{L}^\mathbf{T} + \mathbf{\Psi})^{-1} (\underline{x}_i - \underline{\mu})} \tag{9}$$

of the sample vectors $\underline{x}_1, \underline{x}_2, ..., \underline{x}_n$.

There are no explicit solutions to the likelihood equations $\frac{\partial}{\partial L} \ln L(\mathbf{L}, \mathbf{\Psi}) = 0$ and $\frac{\partial}{\partial \Psi} \ln L(\mathbf{L}, \mathbf{\Psi}) = 0$, so we use Minitab to determine $\hat{\mathbf{L}}$ and $\hat{\mathbf{\Psi}}$ using 6 factors ($m = 6$). Before we proceed with our analysis we must ensure that our choice of m is sufficient to describe the variability of the sample, so we assess the adequacy of m using a Chi-Squared hypothesis test, a Likelihood Ratio Test. Our null hypothesis H_0 states that when $m = 6$, $\mathbf{\Sigma} = \mathbf{L}\mathbf{L}^\mathbf{T} + \mathbf{\Psi}$. The alternate hypothesis H_1 states that when $m = 6$, $\mathbf{\Sigma}$ is any other positive definite matrix. The test statistic χ^2 is given by

$$\chi^2 = [n - 1 - \frac{1}{6}(2p + 4m + 5)] \ln \frac{|\hat{\mathbf{L}}\hat{\mathbf{L}}^\mathbf{T} + \hat{\mathbf{\Psi}}|}{|\mathbf{S}|}, \tag{10}$$

with $\nu = \frac{1}{2}[(p - m)^2 - p - m] = 15$ degrees of freedom. We reject H_0 if $\chi^2 > \chi^2_{\nu,\alpha}$. Our significance level (α) is 2.5%.

Using $\hat{\mathbf{L}}$ and $\hat{\mathbf{\Psi}}$, we compute χ^2 for $m = 6$ and obtain a value of 26.928 with $p - value = 0.0293$. Since $26.928 < 27.490 = \chi^2_{15,0.025}$, we do not reject H_0 and thus conclude that 6 factors are adequate.

It is important to note that if the number of factors is not adequate, we would generate a new set of factor loadings using Minitab and retest our hypothesis with a larger value of m.

Next, we label each factor according to its contribution to the variables. We do so by examining our $\hat{\mathbf{L}}$ matrix and assigning an appropriate label to each factor. However, often the original $\hat{\mathbf{L}}$ matrix does not exhibit strong trends, so we rotate the matrix to exaggerate the contrast between large and small factor loadings. We use a Varimax rotation to enlarge the variation of the squared factor loadings for each factor, thus shifting the moderately large loadings close to one and the moderately small loadings close to zero to facilitate labeling. Our rotated $\hat{\mathbf{L}}$ matrix entries, including the $\hat{\Psi}$ values, are given in Table 6.

Table 6: Factor Loadings

	F_1	F_2	F_3	F_4	F_5	F_6	Ψ_i
Number of Physicians	0.708	-0.420	0.109	-0.216	-0.244	-0.076	0.199
GDP per Capita	0.404	-0.512	0.345	-0.457	-0.234	-0.430	0.007
Urbanization	0.217	-0.935	0.167	-0.147	-0.145	0.003	0.008
Compulsory Education	0.439	-0.160	0.476	-0.101	0.027	-0.015	0.544
Literacy	0.828	-0.174	0.179	-0.256	-0.117	-0.082	0.165
Internet Users	0.464	-0.354	0.360	-0.555	-0.292	-0.063	0.133
Life Expectancy	0.467	-0.381	0.358	-0.309	-0.627	-0.088	0.012
Inflation Rate	-0.016	0.081	-0.682	0.051	0.111	0.066	0.508
Press Freedom	-0.314	0.061	-0.685	0.340	0.089	-0.026	0.305
Combined Gross Enrollment	0.652	-0.422	0.331	-0.255	-0.187	-0.030	0.186
Health Spending	0.429	-0.475	0.479	-0.440	-0.218	-0.242	0.062
Mobile Phone Subscribers	0.362	-0.413	0.332	-0.653	-0.136	-0.049	0.142
Energy Consumption	0.295	-0.456	0.005	-0.276	-0.044	-0.118	0.613

Since rotation alters the $\hat{\mathbf{L}}$ matrix, we repeat our test of the adequacy of m. We do not reject the null hypothesis that the $m = 6$ factor model is adequate since our test statistic after rotation $\chi^2 = 27.059$ remains less than $27.49 = \chi^2_{15,0.025}$ with $p - value = 0.028$.

We label the factors as:

Table 7: FA Labels

Factor	**Label**
F_1	Education
F_2	Rurality
F_3	Government Involvement
F_4	Government Control
F_5	Poor Health
F_6	Poverty

2.4 Discriminant Analysis

After examining trends in the data we turn to the main objective of our research: classifying the countries in our data set according to their level of development. We do so using Discriminant Analysis, which enables us to categorize our data into three subpopulations, Π_1, Π_2, Π_3, representing the developed, developing, and underdeveloped sectors.

To illustrate the theory, we first consider the case of two populations. Suppose we seek to assign a sample vector $\underline{x}$ to one of two multivariate normal populations, Π_1 or Π_2, while minimizing classification errors. To do so, we establish a discrimination rule based on a linear combination of the components of $\underline{x}$, given by $\underline{a}^T\underline{x} = a_1x_1 + a_2x_2 + ... + a_px_p$. We classify $\underline{x}$ into $\Pi_1$if $\underline{a}^T\underline{x} \geq c$, and classify $\underline{x}$ into Π_2 if $\underline{a}^T\underline{x} < c$, where c is a critical value. This classification rule partitions the p-space Ω of $\underline{x}$, so the partitions R_1, R_2 corresponding to the populations Π_1, Π_2 are disjoint and cover the $p-$space Ω. Thus there is no overlap between the two populations and there is no vector $\underline{x}$ such that $\underline{x}$ does not belong to one of the populations. Since there are infinitely many possible partitions, we seek to determine an optimal partition for the classification of $\underline{x}$. We want an optimal partition such that the total probability of misclassification (TPM) is minimized. The total probability of misclassification has two components: α_1 and α_2. The probability that the discrimination rule assigns a vector $\underline{x}$ belonging to Π_1 to Π_2 is given by α_1. Conversely, α_2 represents the probability that the discrimination rule assigns a vector $\underline{x}$ belonging to Π_2 to Π_1. Thus,

$$TPM = \alpha_1 + \alpha_2 = \int_{R_2} f_1(\underline{x})d\underline{x} + \int_{R_1} f_2(\underline{x})d\underline{x}, \tag{11}$$

where $f_1(\underline{x})$ and $f_2(\underline{x})$ are the probability density functions of $\underline{x}$ under Π_1 and Π_2, respectively. The optimal partition R_1 and R_2 is given by

$$R_1 = \{\underline{x} : \frac{f_1(\underline{x})}{f_2(\underline{x})} \geq c\}$$

$$R_2 = \{\underline{x} : \frac{f_1(\underline{x})}{f_2(\underline{x})} < c\}. \tag{12}$$

The constant c is chosen so that $\alpha_1 = \alpha_2$. Labeling $\lambda = \frac{f_1(\underline{x})}{f_2(\underline{x})}$, we then write $R_1 : \lambda \geq c$ as $\underline{a}^T\underline{x} \geq c*$, where $\underline{a} = \Sigma^{-1}(\underline{\mu}_1 - \underline{\mu}_2)$, and $\Sigma, \mu_1,$and μ_2 represent the common covariance matrix and the two population means. In practice these population parameters are unknown, so we use training samples of sizes n_1 and n_2 from Π_1 and Π_2 to generate unbiased estimates of the unknown parameters. We estimate μ_1 and μ_2 using $\overline{\underline{x}}_1$ and $\overline{\underline{x}}_2$, the mean vectors for each subpopulation. We then take sample covariance matrices S_1 and S_2 and calculate an unbiased estimate, S_{pooled}, of the common covariance matrix Σ according to the following equation

$$S_{pooled} = S_p = \frac{1}{n_1 + n_2 - 2}\sum_{i=1}^{2}(n_i - 1)S_i. \tag{13}$$

Since we select our training samples from our normalized data, the sample quantities $\overline{\underline{x}}_1, \overline{\underline{x}}_2$ and S_p are appropriate approximations of the population parameters. Further, these estimates allow us to compute $\hat{\underline{a}}$, giving us Fisher's Linear Discriminant Function:

$$\hat{\underline{a}}^T\underline{x} = (\overline{\underline{x}}_1 - \overline{\underline{x}}_2)^T S_p^{-1}\underline{x}. \tag{14}$$

We assign $\underline{x}$ to Π_1 if $\hat{\underline{a}}^T\underline{x} \geq \hat{h} = \frac{1}{2}(\overline{\underline{x}}_1 - \overline{\underline{x}}_2)^T S_p^{-1}(\overline{\underline{x}}_1 + \overline{\underline{x}}_2)$. Conversely, we assign $\underline{x}$ to Π_2 if $\hat{\underline{a}}^T\underline{x} < \hat{h}$. The total probability of misclassification is a function of $\hat{\Delta}_p$, which represents the distance between the two populations Π_1 and Π_2. The Mahalanobis squared distance is given by

$$\hat{\Delta}_p^2 = (\overline{\underline{x}}_1 - \overline{\underline{x}}_2)^T S_p^{-1}(\overline{\underline{x}}_1 - \overline{\underline{x}}_2). \tag{15}$$

Considering the squared distance between the populations gives rise to another discrimination rule. Under this rule, we classify $\underline{x}$ into Π_1 if D_1^2, the squared distance from $\underline{x}$ to Π_1 given by $(\underline{x} - \overline{\underline{x}}_1)^T S_p^{-1}(\underline{x} - \overline{\underline{x}}_2)$, is less than D_2^2, the squared distance from $\underline{x}$ to Π_2. Mathematically, these two discrimination rules are equivalent; that is, $\hat{\underline{a}}^T\underline{x} \geq h \iff D_1^2 < D_2^2$. This result not only facilitates computation − computer programs rely on squared distance rather than Fisher's Linear Discriminant Function to perform Discriminant Analysis − but also allows us to generalize the two-population classification to k multivariate normal populations $\Pi_1, \Pi_2, ..., \Pi_k$. We expand the discrimination rule to k populations of size $n_1, n_2, ..., n_k$ by assigning $\underline{x}$ to Π_i if $D_i^2 = min\{D_1^2, D_2^2, ..., D_k^2\}$.

When evaluating squared distances according to this discrimination rule we assume that the covariance matrices for all populations are equal, so before we proceed we must confirm the equality of the Σ matrices using a χ^2 hypothesis test. Our H_0 assumes the k covariance matrices $\Sigma_1, \Sigma_2, ..., \Sigma_k$ are equal, while the alternate hypothesis H_1 assumes that $\Sigma_i \neq \Sigma_j$ for some populations i, j with $i \neq j$. We reject H_0 if our test statistic $\frac{M}{c} \geq \chi^2_{\nu,\alpha}$, where

$$\nu = \frac{1}{2}p(p+1), \tag{16}$$

$$M = \{\sum_{i=1}^{k}(n_i - 1)\}[\ln|S_p|] - \sum_{i=1}^{k}(n_i - 1)\ln|S_i|, \tag{17}$$

and

$$\frac{1}{c} = 1 - \frac{2p^2 + 3p - 1}{6(p+1)}[(\frac{1}{n_1 - 1} + \frac{1}{n_2 - 1} + ... + \frac{1}{n_k - 1}) - \frac{1}{n_1 + n_2 + ... + n_k - k}]. \quad (18)$$

For our development classification we perform two discriminant analyses, the first based on a current listing of the first, second, and third worlds and the second using a ranking generated by our first principal component (see Table 4). In both analyses, our data fails the test for Σ equality with test statistics of 431.9 and 154.7 compared to a χ^2 value of 114.27. However, failing this test with a large sample size and many variables is not unusual, so we proceed with the analyses.

In our first discriminant analysis, DA 1, we evaluate an existing classification of first, second, and third world countries published by *Nations Online*. We use the *Nations Online* classifications as the groupings for our training samples. We generate a random training sample of 50 countries from our 170-country sample, which we divide into subpopulations of 10, 11, and 29 countries, representing the first, second, and third worlds, respectively. We choose these subpopulation sizes based on the proportional size of each world to the total sample size. Since the *Nations Online* ranking is based on Cold War political affiliation, DA 1 evaluates the outdated system we seek to correct. In this regard, DA 1 is a control analysis in which we pair outdated rankings with our modern data and variable choices.

From DA 1 we obtain an apparent error rate (AER) of 10% based on the training sample classification. Using this rule, we classify the remaining 120 countries as developed, developing, and underdeveloped. Before analysis the list contains 25 first world countries, 29 second world countries, and 116 third world countries. After analysis we have 44 developed, 33 developing, and 93 underdeveloped countries, with a discrepancy rate of 25% in the predicted classifications. That is, of the countries for which the discrimination rule predicted the classification, 30 out of 120 were assigned to populations that differed from the existing classification. The most significant contrast between the existing list and the discriminant classification involves the third world. Many supposedly "third world" countries became developed countries after analysis; the Bahamas, Brazil, Iraq, Mexico, and Singapore are notable examples. Other countries regressed to underdeveloped status, including Estonia, Lithuania, and Turkey. A complete summary of developmental changes is listed below in Table 8.

Table 8: DA 1 Changes

Nations Online Classification	DA 1 Classification	Country
Third World	Developed	Bahamas, Brazil, Cape Verde, Dominica, Fiji, Finland, Iraq, Lebanon, Libya, Malta, Mexico, Panama, Peru, Saint Kitts and Nevis, Singapore, Solomon Islands, Tonga, Uruguay
Third World	Developing	Burma, Comoros, Cuba, Jordan, Maldives, Paraguay, Saint Lucia Sri Lanka, Syria, Zimbabwe
Second World	Developed	Hungary, Slovenia
Second World	Underdeveloped	Czech Republic, Estonia, Latvia, Lithuania
First World	Developing	
First World	Underdeveloped	Turkey

Since the primary objective of our study is to evaluate the outdated system and improve it using modern data, we perform a second discriminant analysis, DA 2. This analysis uses both a modern ranking paired with modern data and variables. To obtain this modern ranking, we sort the countries according to a development score generated by our first principal component, PC 1. A country's development score is the dot product of PC 1 coefficients and country data. For example, consider the PC 1 weights and the data for the United States in Table 9.

Table 9: Sample PC 1 Score Data

Variable	PC 1	United States
Number of Physicians	0.291	26
GDP per Capita	0.324	47,000
Urbanization	0.261	0.820
Compulsory Education	0.213	12
Literacy	0.277	0.990
Internet Users	0.320	659.680
Life Expectancy	0.315	77
Inflation Rate	-0.148	0.042
Press Freedom	-0.237	17
Combined Gross Enrollment	0.309	0.933
Health Spending	0.332	6,096.200
Mobile Subscribers	0.305	680.307
Energy Consumption	0.201	12,924.220

We calculate the development score for the United States as

$$Score = 0.291(26) + 0.324(47,000) + ... + 0.201(12,924.220) = 20,310.940. \tag{19}$$

We repeat the calculation for all 170 countries, giving us a general picture of each country's economic strength. After sorting the countries from highest to lowest development score, we observe that the United States' development score of 20,310.940 places it sixth in the overall rankings. The development scores range from 141.861 to 37,477.659. Next, we perform a visual inspection of a scatterplot comparing each country's rank to its score. After inspecting trends present in the graph, we observe three distinct groupings of points, so we separate our sample into subpopulations according to the natural divisions. Before analysis the sample contains 30 developed countries, 50 developing countries, and 90 underdeveloped countries. We use proportional training samples of 10, 15, and 25 countries to develop our second discrimination rule, which has an AER of 2%. The second discriminant analysis produces a population of 32 developed, 62 developing, and 76 underdeveloped countries, with a discrepancy rate of 24.2%. Unlike DA 1, DA 2 produces no extreme classification shifts. It is also interesting to note that the relative sizes of the developmental categories in DA 2 more closely align with the modern era. Specifically, many previously underdeveloped countries transition to "developing" status. Developing countries account for a more significant proportion of the total population, reflecting today's advanced technological society which encourages and facilitates developmental progress. Table 10 summarizes the developmental changes of DA 2.

Table 10: DA 2 Changes

PC 1 Classification	DA 2 Classification	Country
Underdeveloped	Developed	
Underdeveloped	Developing	Bolivia, Bosnia and Herzegovina, Colombia, Cuba, Dominican Republic, Ecuador, Egypt, Jamaica, Jordan, Mongolia, Pakistan, Peru, Syria, Ukraine, Vietnam
Developing	Developed	Cyprus, Czech Republic, Dominica, Malta, Portugal, Saint Lucia, Slovakia, Uruguay
Developing	Underdeveloped	Djibouti
Developed	Developing	Bahrain, Kuwait, Qatar, Singapore, Slovenia, United Arab Emirates
Developed	Underdeveloped	

Comparing the two analyses, we observe that the two discrimination rules produce the same development classification for only 100 countries. In only 3 countries, Cuba, Jordan, and Syria, did the two analyses predict the same transition from underdeveloped to developing status. Conversely, there are only two countries, the Bahamas and Finland, for which all three predictions - the discriminant analysis based on a published list, the ranking based on our first principal component, and the second discriminant analysis - yield the same classifications, which contradicts the existing classification. Exactly half of the 170 countries remain stable in their development classifications. Of these 85 countries, the vast majority remain either in the first or third worlds, emphasizing the importance of the second world as a locus of developmental change. In Table 11 we list the final world divisions obtained from the second discriminant analysis.

Table 11: World Classifications

Final Classification (DA 2)	Countries
Developed	Australia, Austria, Bahamas, Belgium, Canada, Cyprus, Czech Republic, Denmark, Dominica, Finland, France, Germany, Greece, Iceland, Ireland, Israel, Italy, Japan, Luxembourg, Malta, Netherlands, New Zealand, Norway, Portugal, Saint Lucia, Slovakia, Spain, Sweden, Switzerland, United Kingdom, United States, Uruguay
Developing	Antigua and Barbuda, Argentina, Azerbaijan, Bahrain, Barbados, Belarus, Bolivia, Bosnia and Herzegovina, Botswana, Brazil, Bulgaria, Chile, Colombia, Costa Rica, Croatia, Cuba, Dominican Republic, Ecuador, Egypt, Equatorial Guinea, Estonia, Gabon, Grenada, Hungary, Iran, Jamaica, Jordan, Kazakhstan, Kuwait, Latvia, Lebanon, Libya, Lithuania, Macedonia, Malaysia, Mauritius, Mexico, Mongolia, Oman, Pakistan, Panama, Peru, Poland, Qatar, Romania, Russia, Saint Kitts and Nevis, Saint Vincent and the Grenadines, Saudi Arabia, Seychelles, Singapore, Slovenia, South Africa, South Korea, Suriname, Syria, Trinidad and Tobago, Turkey, Ukraine, United Arab Emirates, Venezuela, Vietnam
Underdeveloped	Afghanistan, Albania, Algeria, Angola, Armenia, Bangladesh, Belize, Bhutan, Burkina Faso, Burma, Burundi, Cambodia, Cameroon, Cape Verde, Central African Republican, Chad, China, Comoros, Côte d'Ivoire, Democratic Republic of the Congo, Djibouti, El Salvador, Eritrea, Ethiopia, Fiji, Gambia, Georgia, Ghana, Guatemala, Guinea, Guyana, Haiti, Honduras, India, Indonesia, Iraq, Kenya, Kyrgyzstan, Liberia, Madagascar, Malawi, Maldives, Mali, Mauritania, Moldova, Morocco, Mozambique, Namibia, Nepal, Nicaragua, Niger, Nigeria, Papua New Guinea, Paraguay, Philippines, Republic of the Congo, Rwanda, Samoa, Senegal, Sierra Leone, Solomon Islands, Sri Lanka, Sudan, Swaziland, Tajikistan, Tanzania, Thailand, Togo, Tonga, Tunisia, Turkmenistan, Uganda, Uzbekistan, Yemen, Zambia, Zimbabwe

We notice several interesting results that arise from our final classification system, DA 2, and observe that these results are supported by current world events and conditions. First, many oil-rich Middle Eastern countries, such as Kuwait, Qatar, Saudi Arabia, and United Arab Emirates, were ranked in the developed world in our PC 1 ranking, which strongly emphasize their economic strength. In fact, Qatar was the overall highest-ranked country, with a PC score of 37,478. All of these Middle Eastern countries were assigned to the developing world, which we feel more accurately reflects their overall state of development. Second, the communist second world countries of the Cold War were redistributed between the developing and underdeveloped worlds. Third, despite much media commentary on China and India's emerging economies, both of these countries remain classified as underdeveloped. Due to similar media attention, we initially hoped to compare North and South Korea's development levels, but we were unfortunately unable to obtain data for North Korea. Finally, as current political turmoil suggests, Afghanistan, Haiti, and most of sub-

Saharan Africa remain underdeveloped.

3 Conclusion

We constructed a new model of world development using the multivariate statistical techniques Principal Component Analysis, Factor Analysis, and Discriminant Analysis. Unlike the ideological classifications dating from the Cold War, our model takes aspects of modernization into account and thus better represents the current country classifications. We consider not only modern data but also modern variables such as technology and energy usage, encompassing a wide variety of developmental indicators to provide an accurate picture of each country's development status. We expected the developing world to grow in size as a reflection of overall global progress post-Cold War, and were pleased to see that the discriminant analysis based on our first principal component produced this anticipated result. Initially we hoped to include measures of gender equity and civil liberties among our variables, but sufficient information was not available. The same holds true for crime data in order to obtain a measure of internal violence and unrest. Incorporating these variables into our study would enhance the picture of development by reducing the emphasis on the economy. With access to more data, we could provide a more comprehensive depiction of world development.

4 Acknowledgments

We would like to extend our gratitude to Dr. Vasant Waikar, our seminar director, for his teaching and guidance throughout our research. We would also like to thank Dr. Tom Farmer for his assistance with the technical writing aspect of the research. In addition, we thank Ashley Swandby, our Graduate Assistant, for her support and encouragement with our project. We are also grateful to the SUMSRI Program and its sponsors, the National Security Agency, National Science Foundation, and Miami University, for providing us with this wonderful opportunity.

References

1. R. Johnson and D. Wichern, *Applied Multivariate Statistical Analysis*, Pearson Prentice Hall, Upper Saddle River, New Jersey, 6th edition, 2007.

2. Central Intelligence Agency, *The World Factbook*, 2009.

3. United Nations, *UNdata*, United Nations Statistics Division, 2009.

4. Freedom House, *Press Freedom Survey 2008*, 2008.

5. Heritage Foundation, *2009 Index of Economic Freedom*, 2009.

6. One World - Nations Online, *The First, Second, and Third World*, 2008.