DATA SCIENCE IN DEVELOPMENT ECONOMICS: USING CLUSTER ANALYSIS TO GENERATE A MULTIVARIATE DEVELOPMENT TAXONOMY

by

Andrew Gross

A thesis submitted to the Faculty of the University of Delaware in partial fulfillment of the requirements for the degree of Honors Bachelor of Science in Mathematics and Economics with Distinction

Spring 2018

DATA SCIENCE IN DEVELOPMENT ECONOMICS: USING CLUSTER ANALYSIS TO GENERATE A MULTIVARIATE DEVELOPMENT TAXONOMY

by

Andrew Gross

I certify that I have read this thesis and that in my opinion it meets the academic and professional standard required by the University as a thesis for the degree of Bachelor of Science.

Signed: ____________________________________
Dr. Evangelos Falaris, Ph.D.
Professor in charge of thesis

Approved: ____________________________________
Dr. Dominique Guillot, Ph.D.
Committee member from the Department of Mathematical Sciences

Approved: ____________________________________
Dr. Meryl Gardner, Ph.D.
Committee member from the Board of Senior Thesis Readers

Approved: ____________________________________
Paul Laux, Ph.D.
Director, University Honors Program

ACKNOWLEDGMENTS

I would first like to thank Dr. Falaris, who was willing to take on the role of adviser to this research project. His experience in development economics helped tremendously in steering me towards the proper questions to ask and what data to use.

My gratitude extends to my second reader, Dr. Guillot. This paper turned out to be more mathematical than I had originally intended, but his extensive knowledge of data science and the theory behind the algorithms used greatly aided in my understanding and application of them.

On a more personal note, I must give thanks to my family. Early on in my career as a student I lacked the the motivation to succeed, but my parents realized my potential. All of the time and money they spent towards igniting in me a passion for learning ultimately paid off, as I would not be writing this if it were not for their enduring support. And lastly to my brother, who four years ago was submitting a senior thesis in mathematics and economics at UD (clearly I look to him as a role model). It would be disingenuous not to say that a small part of me chose to write a senior thesis to let the Gross dynasty endure.

TABLE OF CONTENTS

CHAPTERS

APPENDIX

LIST OF TABLES

LIST OF FIGURES

ABSTRACT

This paper attempts to apply clustering techniques from data science to the economic problem of generating a country-level development taxonomy. Development taxonomies currently in use suffer from two key issues. First, the taxonomies are based on very few variables and therefore cannot properly represent something as complex and multifaceted as development. Second, the values used to discriminate groups are chosen arbitrarily. In this work, a univariate analysis is performed using the method of kernel density estimation to empirically generate a single-valued taxonomy which can be directly compared with the income group taxonomy published by the World Bank. Next, a definition of development is derived and a multivariate analysis is performed to create a comprehensive development taxonomy using two forms of k-means clustering. The univariate analysis demonstrates the superiority of a data-driven approach to single-valued taxonomy creation. Conversely, it remains inconclusive as to whether cluster analysis can create a well-defined multivariate development taxonomy.

Chapter 1

INTRODUCTION

A taxonomy is a classification of objects based on a set of attributes such that similar objects are grouped together. Taxonomies have many applications that span a variety of fields. In economics and finance, taxonomies are used to group countries based on sets of development indicators. Credit rating agencies such as Standard & Poor's utilize methodologies to group countries by level of investment risk. The International Monetary Fund uses the results of a taxonomy which groups countries by level of income to create a suite of lending programs for countries classified as "low income" (Mumssen et al., 2012). Similarly, the World Bank implements a classification scheme that groups countries by level of institutional development and capital-market access to determine eligibility to borrow from the International Bank for Reconstruction and Development (Knack et al., 2012). In the field of development economics, the results of country-level taxonomies are used to conduct policy experiments on countries which share particular characteristics. For example, low income countries may be used as the sample with which to study education policy (Filmer and Schady, 2014) or labor regulations (Alatas and Cameron, 2008).

Despite their extensive use in industry and academia, the current taxonomies used to classify countries by level of development are flawed. One of the most widely referenced taxonomies for this purpose is the World Bank Income Groups. The World Bank does not explicitly state the intended purpose of its taxonomy, but policymakers and academics use it as a proxy for country development groups, as evidenced by the examples in the previous paragraph. The World Bank's taxonomy classifies each country as either "low", "lower middle", "upper middle" or "high" income[1]. Two major issues exist in the Bank's taxonomy. First, the income groups have arbitrary cutoffs. The low income threshold is based on an obscure benchmark while the middle and high income thresholds are determined by a decades-old staff report whose authors used an arbitrary number (World Bank, 1989). Second, economic development cannot be explained by a single variable. A country's level of development is determined by a host of measurements which cover all aspects of growth, including education, health, and corruption, to name just a few. Thus, the taxonomy produced by the World Bank provides an unsatisfactory definition of development groups.

Cluster analysis is a tool primarily used in data science to form natural groups in a set of data. Regardless of the clustering method, the goal of cluster analysis is to create groups such that each group is distinct and that every data point falls inside exactly one group. This is accomplished by maximizing "similarity" between points in the same cluster and minimizing "similarity" between points in different clusters, where "similarity" is some measure defined by each clustering method. Cluster analysis solves both problems faced by the World Bank's taxonomy. First, cluster analysis performs optimization to maximize within-group similarity and minimize between-group similarity, providing an objective and substantive framework for choosing group cutoffs. Second, unlike the World Bank's taxonomy, a variety of clustering methods exist which are optimized to perform cluster analysis on high-dimensional data. That is, cluster analysis has the potential to construct an objective, mathematically-backed multivariate country-level development taxonomy.

The rest of the paper will proceed as follows: Section 1.1 will review previous attempts at creating a data-driven development taxonomy. In Chapter 2 kernel density estimation is performed to construct a univariate taxonomy to compare with that

[1] Income is defined as gross national income (GNI) per capita in U.S. dollars calculated using the Bank's Atlas method. For information on the Atlas method, visit https://datahelpdesk.worldbank.org/knowledgebase/articles/378832-the-world-bank-atlas-method-detailed-methodology.

proposed by the World Bank. Chapter 3 contains a survey of clustering methods for high-dimensional data and a multivariate development taxonomy is constructed and tested. Chapter 4 provides a summary of the findings and suggestions for future work.

1.1 Review of previous work

The structure of this paper is based on work by Sumner and Tezanos Vázquez (2012) who use a hierarchical clustering algorithm to create development groups. Their analysis was performed only for countries considered low and middle income as defined by the World Bank. Development is subdivided into four groups: Human development (poverty, inequality, etc.), economic autonomy (structural change, dependence on natural resources, etc.), political freedom (good governance and quality of democracy), and environmental sustainability. Statistical tests run on the cluster centers show significant differences, but no internal validation is performed to determine the quality of the clusters produced. That is, without testing the clustering results there is no knowing how similar countries are within each group.

Zhang and Gao (2015) also use hierarchical clustering methods, but for the purpose of grouping emerging markets based on growth prospects before and after the 2008 financial crisis. In doing so, they create multiple taxonomies based on different correlates of economic growth and compare which countries belong to which clusters across all taxonomies generated. Thus, the model they use is based purely on economic features correlated with growth such as factor endowments and real/financial external linkages.

The application of data science techniques to economic problems arises in the study of macroeconomic indexes, a field that runs tangential to taxonomic analysis. Coccia (2007) proposes the generation of an index that measures country risk and performance across multiple dimensions using principal component analysis. Principal component analysis is a technique to reduce the dimensionality of a dataset down to a specified number of dimensions (i.e., principal components) without disturbing the underlying characteristics of the data. The analysis is run with a single principal component, which corresponds to a single score for each country. Because the score represents country risk, Coccia's indexing method is tailored mainly to international investors.

Chapter 2

UNIVARIATE ANALYSIS

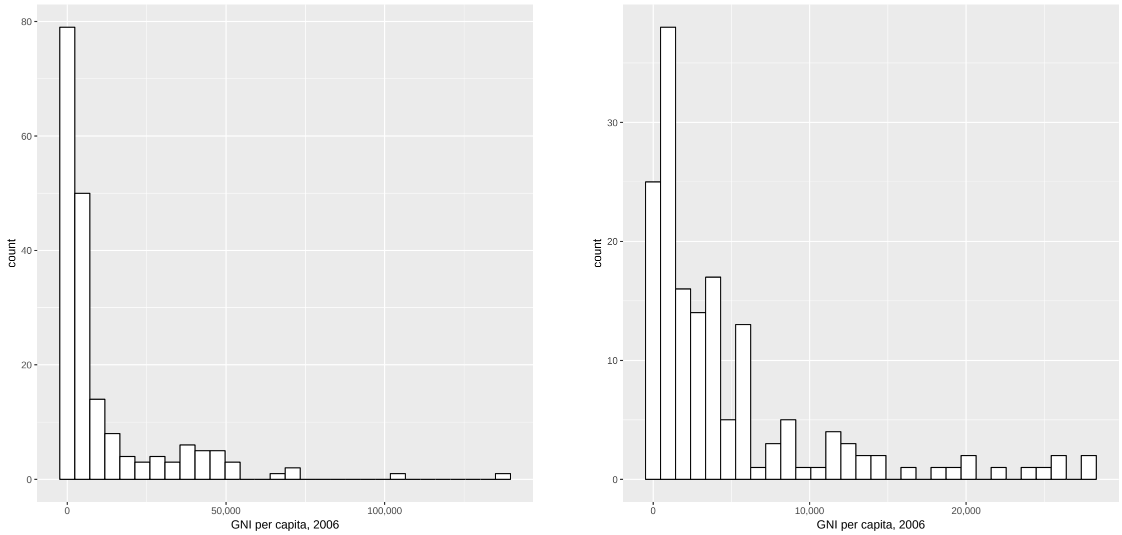
The purpose of this chapter is to establish a baseline for the efficacy of using a data-driven development taxonomy in place of the subjective taxonomies currently in use. Concretely, kernel density estimation will be used to create a taxonomy based solely on gross national income per capita. This way, a direct comparison can be made between the taxonomy produced by cluster analysis and that published by the World Bank.

2.1 Data

As mentioned, the taxonomy is based solely on gross national income per capita (GNI) to allow for direct comparison with the World Bank's income group classification schema, which permits the explicit ranking of taxonomy quality. Thus, the data consists of GNI figures for 189[1] countries from the year 2006[2]. Figure 2.1 shows the distribution of the data. Clearly, the distribution is unimodal and positively skewed, indicating that a disproportionate amount of countries have low and middle standards of living. This will factor into how the group cutoffs are chosen. The fact that the data does not adhere to a normal distribution does not affect the analysis.

[1] Djibouti, Nauru, Somalia, and South Sudan were excluded because no GNI data was recorded for these countries in the chosen year

[2] 2006 was chosen for the purpose of collecting a large sample size during a year of relative stabiliy (i.e., before the recession of 2007-08). The GNI of many countries has increased since 2006, but the purpose of this research is just to test the efficacy of the model.



(a) All countries

(b) GNI less than $30,000

Figure 2.1: Distribution of country GNI per capita

2.2 Theory

This section is devoted to creating a univariate taxonomy. Thus, popular clustering methods such as k-means which do not take advantage of certain properties of the real numbers (e.g., order) will be put aside until Chapter 3. Instead, kernel density estimation will be used. Kernel density estimation (KDE) is a non-parametric technique independently discovered by Rosenblatt (1956) and Parzen (1962) used to estimate a probability density function (PDF) for discrete, finite sets of data. Clustering algorithms aim to create dense groupings. Thus, KDE is performed to serve the goal of finding the local minima of the estimated PDF, which act as group cutoffs.

The estimated density function follows the form

$$\hat{f}(y) = \frac{1}{nh} \sum_{i=1}^{n} K\left(\frac{y - x_i}{h}\right),$$

where K is the kernel function (i.e., the distribution), n denotes the sample size, h represents the width of the bins, and x_i is the i^{th} sample observation. Regardless of what K is chosen, the underlying idea behind KDE is to define each observed point x_i as the mean of a distribution with standard deviation equal to the width h of the histogram bin which contains x_i and take the sum of these individual distributions to derive the density estimate. Intuitively, intervals that contain more sample points add more "weight" to the overall distribution estimate compared to more sparse intervals.

2.2.1 Bandwidth selection

Calculating the KDE requires two parameters: the choice of kernel and the bandwidth. For the purposes of this analysis the Gaussian kernel function with weight $K(y) = \frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}x^2}$ is used for its smoothness and continuity. The choice of bandwidth is critical, as it determines the smoothness of $\hat{f}$. Too narrow a bandwidth and $\hat{f}$ will be under-smoothed, producing a jagged PDF with dozens of local minima. A bandwidth that is too wide will lead to over-smoothing, masking important features in the data. The cross validation maximum likelihood method is used to find the optimal bandwidth.

The maximum likelihood problem is set up as follows. Let $X_1, X_2, ..., X_n$ be a set of i.i.d. random variables whereby the form of their distribution is dependent on the bandwidth h. The goal of maximum likelihood estimation is to maximize the likelihood of drawing the sampled data by maximizing the joint probability density function $\hat{f}_h(X_1, X_2, ...X_n)$. In doing so, an estimate of h, call it h_{MLCV}, will be derived. The random variables are assumed to be independent. Thus, $\hat{f}_h(X_1, X_2, ..., X_n) = \prod_{i=1}^{n} \hat{f}_h(X_i)$. Maximizing this product with respect to h leads to the trivial solution $h_{MLCV} = 0$. This is not feasible because a bin cannot have a width of zero. Thus, cross-validation must be used to find a non-trivial optimal bandwidth.

Cross-validation is a data validation technique to assist in generalizing the result of a statistical procedure to any independent set of data. In the context of selecting a suitable bandwidth, cross-validation provides an estimate of h that is optimal not just for the given sample, but for any sample provided. Specifically, leave-one-out (n-fold) cross-validation is used. The original sample is subdivided into n subsamples. That is, each observation is excluded from exactly one subsample such that each subsample contains $n-1$ observations. Maximum-likelihood estimation is performed on each subsample, and the results from each procedure are averaged. This process runs through every potential value of h in a defined interval. The value of h that maximizes the average maximum-likelihoods is the optimal bandwidth. Mathematically, the process is described as follows. We begin by replacing $\hat{f}_h(X_i)$ with

$$\hat{f}_{\{h,i\}}(X_i) = \frac{1}{(n-1)h} \sum_{\substack{j=1 \\ j \neq i}}^{n} K\left(\frac{X_j - X_i}{h}\right),$$

where n is the sample size, h is the estimated bandwidth, and K is the kernel, here assumed to be Gaussian. For simplicity in taking derivatives, the maximum log-likelihood is used. We wish to find $h_{MLCV} = \text{argmax}_{h>0}\, MLCV(h)$, where

$$MLCV(h) = \frac{1}{n} \sum_{i=1}^{n} log\left[\sum_{j \neq i} K\left(\frac{X_j - X_i}{h}\right)\right] - log\left[(n-1)\, h\right].$$

The method was optimized over a predefined interval of possible values of h. The results demonstrate the bandwidth asymptotically approaching a value between 577 and 580. Taking the mean of the solutions produced in the last several iterations gives an optimal bandwidth of 574.16. The same procedure was conducted with other bandwidth optimization algorithms and similar results were produced. Figure 2.1 shows a distribution of country GNI per capita skewed heavily to the right. Thus, KDE was performed for countries with GNI per capita less than US$30,000 in order to analyze local minima at a more granular level.

2.3 Results

Figure 2.2 shows the estimated PDF produced by kernel density estimation using the optimal bandwidth obtained by maximum likelihood cross validation. The graph shows four significant dips in the range of $0-$10,000 but becomes relatively flat thereafter, including for the set of higher values of GNI that are not represented. Thus, the first 4 minima will act as the income group cutoffs and 5 income groups are defined in the new taxonomy. This is in contrast to the World Bank taxonomy which defines only 4 income groups. The two taxonomies are described in Tables 2.1 & 2.2 and a list of countries with their respective KDE-defined income groups is found in Table A.1. Visually, it is clear that the World Bank cut-offs are not located at local minima, indicating a sub-optimal clustering. One way to empirically compare

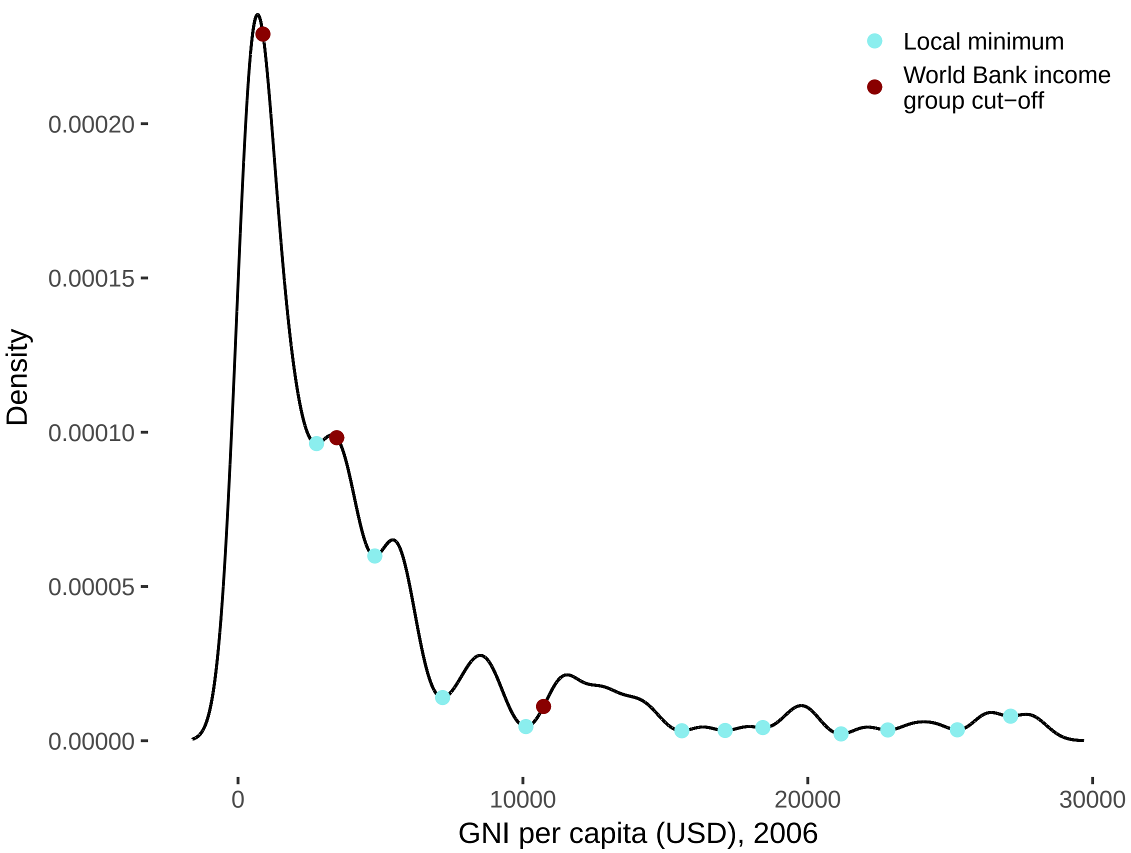


Figure 2.2: Estimated probability density function produced by KDE

the relative accuracy of each clustering scheme is by way of the silhouette index, an internal validation measure proposed by Rousseeuw (1987). Let $\mathbf{x}_i$ be the vector of characteristics for country i belonging to some cluster A. Also, let C be another cluster. Then define

$$a(i) := \frac{1}{n_A} \sum_{j \in A} ||\mathbf{x}_i - \mathbf{x}_j||_2$$

$$b(i) := \min_{C \neq A} \left(\frac{1}{n_C} \sum_{j \in C} ||\mathbf{x}_i - \mathbf{x}_j||_2 \right),$$

where n_K is the number of countries in cluster K. Then the silhouette index for country i is $s(i) := \frac{b(i)-a(i)}{max(a(i),b(i))}$. Furthermore, the mean cluster silhouette index is $s_K := \frac{1}{n_K}\sum_{i\in K} s(k)$ and the global silhouette index is $s := \frac{1}{|\mathbb{K}|}\sum_{K\in\mathbb{K}} s_K$, where $\mathbb{K}$ is the set of all clusters formed. The silhouette index can range from -1 to 1. When $s = 1$, the within-cluster distance is minimized and the "between" cluster distance is maximized, indicating a well-clustered set of data. The opposite is true for $s = -1$. In the case of the analysis just performed, the KDE-produced taxonomy obtained an s of 0.54 while the World Bank's taxonomy scored a 0.38. A silhouette index of 0.54 signifies an adequate clustering while an index of 0.38 indicates a weak clustering (Kaufman and Rousseeuw, 1990). Thus, according to the silhouette index, the taxonomy produced by KDE is superior to that created by the World Bank.

Table 2.1: World Bank Income Groups, 2006

Low Income	Lower Middle Income	Upper Middle Income	High Income
≤$875	$876-$3,465	$3,466-$10,725	>$10,725

Table 2.2: Income Groups using Kernel Density Estimation

Group 1	Group 2	Group 3	Group 4	Group 5
≤$2,756	$2,756-$4,800	$4,801-$7,180	$7,181-$10,105	>$10,105

Use of the empirical taxonomic model has implications in determining which countries are eligible to receive concessional loans. The IMF's Poverty Reduction and Growth Trust program (PRGT) provides a suite of concessional lending benefits for countries that fall below the operational cutoff established by the International Development Association, which maintains a ratio of approximately 1.52:1 with the low-income cutoff established by the World Bank (International Development Association, 2001). In 2006, the operational cutoff was $1,675. By applying the ratio to the empirical model, the modified operational cutoff is $4,189. Under the taxonomy proposed, the number of countries qualified to receive concessional lending under PRGT would expand to include those such as Colombia and Ukraine which are not eligible under the paradigm founded on the World Bank taxonomy.

Chapter 3

MULTIVARIATE ANALYSIS

Now that a data-driven approach has been shown to produce a more accurate univariate taxonomy than those used today, the analysis will be extended to higher dimensions. Assessing a country's level of development involves observing a multitude of variables that span all sectors of society. Thus, the results of a multivariate clustering analysis will prove particularly useful in creating a taxonomy to measure economic development.

3.1 Data

Historically, the West has decided what constitutes development (Vázquez and Sumner, 2012). Thus, variables such as societal openness and system of government have typically been included in its definition (United Nation General Assembly, 1948). In recent years these ideas have been questioned, but analyzing the shift in global norms goes beyond the scope of this paper. Distinguished development economist Amartya Sen defines development as "a process of expanding the real freedoms that people enjoy," which include not just the freedom of economic opportunity but also more human freedoms such as the freedom to live a healthy life and the freedom to learn (Sen, 1999). Most development economists have adopted Sen's definition as the leading interpretation of development. Thus, the variables in the multivariate analysis were chosen to fit within this development paradigm. Broadly, they fall under the following categories: investment environment, economic stability, public health, research and development, education, and human rights and corruption. Unless otherwise specified, the data come from the World Bank Development Indicators database (2000–2017).

The data span 17 years, but an effort was made to choose values for the year nearest to 2006. Provided below are descriptions of the categories used to define development.

- *Investment environment* includes several indicators that represent a country's efforts to promote business and investment. These include the cost, the number of procedures, and the time required to start a business.

- *Economic stability* refers to the strength of a country's economy. For the purposes of this paper, economic strength will be based on three measures: Standard of living, dependence on natural resources, and inequality. Standard of living will be measured using GNI per capita as in the univariate analysis. A country's dependence on natural resources is a proxy for the level of diversification in its economy. A more diversified economy can rebound quicker after a negative shock to any given sector. Thus, this will be measured by a country's total natural resources rent as a percent of GDP, where rent is the excess revenue after accounting for the costs of resource extraction (OECD). Lastly, economic inequality is measured by the poverty headcount ratio as a percent of the population, defined as the ratio of people whose income fall below their respective national poverty line.

- *Public health* describes the health of a country's population with a special focus on areas in which the state plays a role. Most of the variables chosen in this area pertain to early childhood health because of its correlation with future economic productivity (Schultz, 2010). These variables include percent of children immunized against diphtheria and measles and percent of low birth weight babies (UNICEF). Other indicators include life expectancy, and adult and infant mortality rates.

- *Research and development environment* refers to the scientific output of a country. It serves mainly as a proxy for assessing the quantity and quality of the proportion of the population who have completed tertiary education and beyond. Two variables are used to describe this environment: Number of scientific documents published per million people and the country *h*-index. The *h*-index is an author-level index that quantifies each author's impact based on a combination of their published articles and number of citations. Thus, the country *h*-index is just an aggregate of the *h*-indexes of every author in a given country. The data come from the SCImago Journal & Country Rank database (SCImago, 2016).

- *Education* is widely known to be a correlate of development, ranging from its positive relationship with economic growth and public health to its negative relationship with crime (Hanushek, 2007; Lochner and Moretti, 2004; Silles, 2009). The only variable available for a sufficient sample of countries is average years of schooling. The data come from the Barro-Lee Educational Attainment Database (Barro and Lee, 2013).

	Measure	Variable
Investment Environment	Cost to start a business (% of income per capita)	$cost\_biz$
	Procedures required to start a business (days)	$proc\_req$
	Time required to start a business (days)	tim_req
Economic Stability	GNI per capita, Atlas method (current US$)	gni
	Total natural resources rents (% of GDP)	nat_res_rent
	Poverty headcount ratio at national poverty lines (% of population)	$povHCR$
Public Health	Immunization, DPT (% children ages 12-23 mo)	imz_dpt
	Immunization, measles (% children ages 12-23 mo)	imz_msls
	Life expectancy at birth, total (years)	lf_expec
	Low birth weight babies (% of births)	low_bw
	Mortality rate, infant (per 1,000 live births)	$mort\_rt\_inf$
	Mortality rate, adult (per 1,000 adults)	$mort\_rt\_ad$
Research and Development	Documents published (per 1 million people)	num_doc
	h-index	$hidx$
Education	Average years of schooling attended (years)	avg_yr_sch
Human Rights and Corruption	Physical integrity rights (0-8)	$physint$
	Freedom of foreign movement (0-2)	$formov$
	Freedom of domestic movement (0-2)	$dommov$
	Independence of the judiciary (0-2)	$injud$

Table 3.1: Breakdown of development indicators and their corresponding variable

- *Human rights and corruption* are important factors in identifying development, as they are tied to a country's economic growth (Mauro, 1995; Mo, 2001). For example, a corrupt judicial system may favor domestic businesses over their international counterparts in a land ownership dispute. This would make for a hostile environment for foreign direct investment, potentially leading to slower economic growth. The CIRI Human Rights Data Project created a set of indexes which measure human rights and corruption (Cingranelli et al., 2014). Among these include measures of government respect for human rights, freedom of foreign and domestic movement, and independence of the judiciary, where a higher score indicates greater government respect for the given variable. Further explanation of each index can be found on CIRI's website[1].

Table 3.1 provides a summary of the indicators which will be used in the multivariate analysis along with their corresponding variable name. From this point onward the indicators will be referenced by their respective variable name. The sample for the multivariate analysis consists of 94 countries located across every continent. Table 3.2 provides a summary of the correlation matrix of the development indicators (Table A.3

[1] http://www.humanrightsdata.com/p/data-documentation.html

Table 3.2: Distribution of correlation coefficients ρ

Min	1st Quartile	Median	Mean	3rd Quartile	Max
-0.95	-0.42	-0.18	-0.01	0.39	0.91

provides the full correlation matrix). The majority of the correlation coefficients lie between -0.41 and 0.41, meaning most pairs of indicators are not too highly correlated, an indication that each variable plays a unique role in differentiating the countries sampled[2].

3.2 *k*-Means

k-means clustering is one of the most widely used algorithms for clustering high-dimensional data due to its simple and intuitive nature. At its core, the goal of *k*-means is to group similar points and separate dissimilar points, where "similarity" is defined by the user. The algorithm is rudimentary, so *k*-means is most effective when the data form clearly separate groups, all with similar size and density (MacKay, 2005). Thus, the multivariate analysis will begin first with basic *k*-means and later be performed with more advanced *k*-means techniques in an effort to create a more accurate multi-dimensional development taxonomy.

3.2.1 Theory

Given an n-dimensional dataset of countries and a number of clusters k, let $\mathbf{x}_i$ be the vector that characterizes country i. The algorithm begins by randomly assigning k countries as the cluster centroids. That is, each country chosen as a centroid is the center of the p^{th} cluster C_p. The centroid (mean) of each cluster $\mathbf{m}_p$ is recorded. The algorithm loops through the following two steps until each country remains in the same cluster after two consecutive iterations:

[2] No pair of indicators has an absolute correlation coefficient greater than 0.90 except for life expectancy and adult mortality rate ($\rho = -0.95$). Also, the summary table excludes diagonal entries of the correlation matrix, which explains why the max ρ is not 1.0

1. *Assignment.* Assign each $\mathbf{x}_i$ to the nearest cluster C_{p^*}. That is, for a given $\mathbf{x}_i$, choose $p^* = \operatorname{argmin}_{p=1,2,\ldots,k} (\|\mathbf{x}_i - \mathbf{m}_p\|_2)$. Thus, $\mathbf{x}_i$ will be assigned to C_{p^*}. Formally, the ${p^*}^{th}$ cluster C_{p^*} is defined as follows:

$$C_{p^*} := \{i | \|\mathbf{x}_i - \mathbf{m}_{p^*}\|_2 \leq \|\mathbf{x}_i - \mathbf{m}_p\|_2 \forall p = 1, 2, ..., k\}.$$

2. *Update.* Update the means of each cluster centroid to reflect the new composition of the clusters. Mathematically, reevaluate each $\mathbf{m}_p$ using

$$\mathbf{m}_p = \frac{\sum_{\mathbf{x}_j \in C_p} \mathbf{x}_j}{|C_p|}, \quad \forall p = 1, 2, ..., k.$$

Thus, in the case of basic k-means, "similarity" is a measured by Euclidean distance and the goal is to minimize the sum of squared deviations of each country from its respective cluster centroid. The user must decide what number of clusters k to use in running k-means. With the stated goal, it is reasonable to choose the k which has the most impact in decreasing the within sum of squared deviations. Figure 3.1 displays the within sum of squared deviations of points to their respective clusters for multiple runs of k-means. This provides a visual aid for choosing k. From the plot,

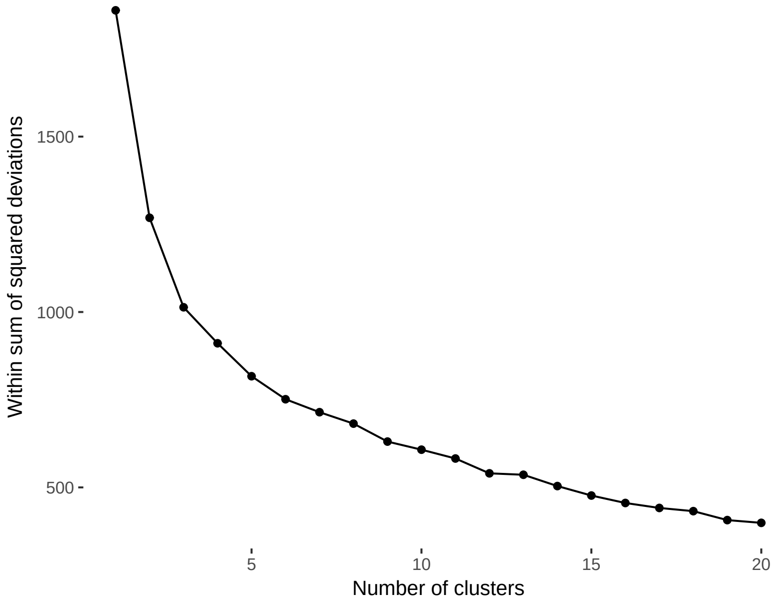


Figure 3.1: Scree plot for k-means

there appears to be a large drop in within sum of squares from $k = 3$ to $k = 4$ and a marginal drop from $k = 4$ to $k = 5$. Thus, $k = 4$ appears to be a reasonable choice for the number of clusters.

3.2.2 Results

Figure 3.2 provides a summary of the clusters formed using basic k-means. A full description of the taxonomy including which countries belong to what clusters can be found in Table A.4. The heat map gives a ranking of the normalized mean values of each cluster, where a cluster is represented by the cells whose labels correspond to the cluster number. Because the means are normalized, they have no economic interpretation. Instead, they can be used to compare and rank clusters across each dimension by order of magnitude. Analysis of variance (ANOVA) is performed to determine whether the cluster centroids (i.e., the means of each cluster) are significantly different. This test serves as a weak proxy in determining the degree of separation between clusters, which

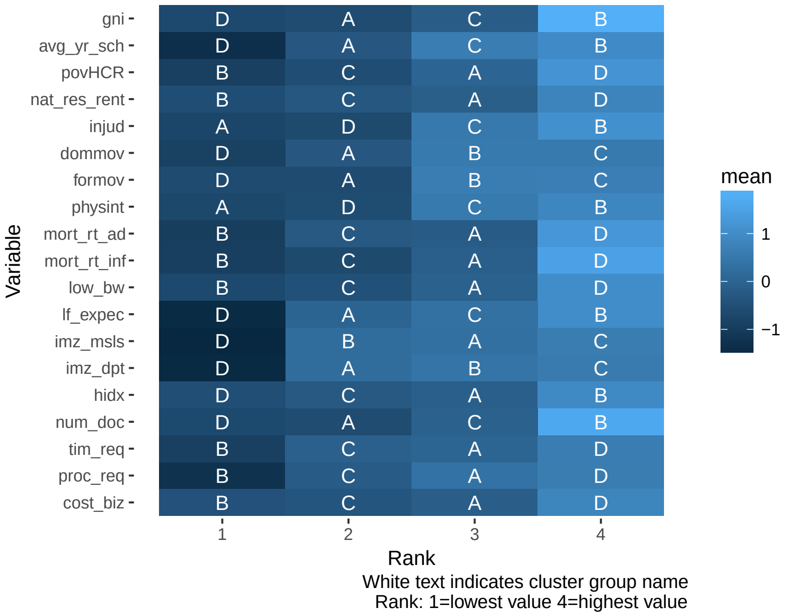


Figure 3.2: Description of the taxonomy produced by basic k-means

varName	df	sumSq	meanSq	fVal	pVal
cost_biz	3.00	21.38	7.13	8.96	0.00
proc_req	3.00	41.22	13.74	23.88	0.00
tim_req	3.00	22.27	7.42	9.45	0.00
num_doc	3.00	69.31	23.10	87.79	0.00
hidx	3.00	23.19	7.73	9.97	0.00
imz_dpt	3.00	59.37	19.79	52.96	0.00
imz_msls	3.00	64.65	21.55	68.42	0.00
lf_expec	3.00	68.13	22.71	82.16	0.00
low_bw	3.00	35.78	11.93	18.76	0.00
mort_rt_inf	3.00	72.79	24.26	108.05	0.00
mort_rt_ad	3.00	51.92	17.31	37.92	0.00
physint	3.00	41.86	13.95	24.56	0.00
formov	3.00	37.98	12.66	20.71	0.00
dommov	3.00	31.09	10.36	15.06	0.00
injud	3.00	54.49	18.16	42.45	0.00
nat_res_rent	3.00	21.99	7.33	9.29	0.00
povHCR	3.00	50.85	16.95	36.20	0.00
avg_yr_sch	3.00	67.02	22.34	77.39	0.00
gni	3.00	80.32	26.77	189.99	0.00

Table 3.3: Summary of ANOVA results for basic k-means

in turn provides information about the quality of the taxonomy produced. Table 3.3 displays the results of the ANOVA test.

Every variable received a p-value (pVal) of less than 0.01, indicating that for each dimension there exists at least one pair of clusters that have significantly different means. This does not mean that means of each cluster are significantly different across every dimension, but this is not necessary in a high-dimensional taxonomy, for two distinct clusters need not differ in value across every variable. The f-value (fVal) measures the effect each variable has in determining the dissimilarity of cluster centroids. The higher the f-value, the greater its contribution in defining distinct centroids. By this metric, it appears measures of health and education/R&D have the greatest effect in distinguishing centroids, with the exception of GNI contributes the most from any single variable. Variables classified under human rights & corruption and investment environment have low f-values, signaling their weak role in establishing distinct clusters.

3.3 Drawbacks to basic k-means

As mentioned in the previous section, ANOVA is a weak test to determine the quality of a taxonomy because it only measures distinctness of cluster centroids. The cluster means may be well separated, but that provides no information regarding the the density of the clusters. If the clusters are not dense, the boundaries of each cluster may be in close proximity to the others, which brings into question the uniqueness of the groups produced. Fortunately, there exist a multitude of internal validation measures specifically created for cluster analysis to determine the quality of the results of a particular clustering method, such as the silhouette index introduced in Chapter 2. The silhouette index for the taxonomy produced by basic k-means is 0.19. According to the rating system introduced in Chapter 2, a value of 0.19 indicates the clustering cannot even be considered a taxonomy given its non-existent structure. As alluded to previously, the low silhouette index may be due to the sparsity of the data. That is, the countries may be too dispersed in order to form well-defined clusters. Density-based spatial clustering of applications with noise and a high-dimensional outlier detection schema provide evidence that supports this hypothesis. The following subsections provide brief descriptions of each method.

3.3.1 Density testing

Density-based spatial clustering of applications with noise (DBSCAN) is an alternative to k-means based clustering methods. k-means partitions a dataset into k clusters to then reshape them through an iterative process based on measuring a point's distance from the mean. In contrast, DBSCAN uses a density-based approach to form groups from the data (Ester et al., 1996). Given a user-defined ϵ radius and minimum number of neighbors $minPts$, DBSCAN classifies the data in three ways. A data point p is a:

- *core point* if there exists a set of n points $\{q_i\}_{i=1}^{n}$ all within ϵ of p and $n \geq minPts$. It is also said that q is directly density reachable to p.
- *border point* if there exists a chain of points $p_1, p_2, ..., p_n = p$ such that p_i and p_{i+1} are directly density reachable.
- *outlier* if it does not satisfy any of the above requirements.

Thus in the context of development taxonomies, DBSCAN creates a graph of countries, where groups are formed from core countries and expanded to include countries associated with them.

DBSCAN was performed on the data using $\epsilon = 1.995$ and $minPts = 3$. The result is a clustering that classifies 68% of the sampled countries as random noise, leaving less than a fifth of the sample available for clustering. Such a large quantity of data assigned as noise with such relaxed parameters leaves reason to believe that the data are highly scattered and not conducive to clustering.

3.3.2 Outlier detection

The results of DBSCAN are confirmed by running a high-dimensional outlier detection schema proposed by Krigel et al. (2009). The algorithm randomly chooses subspaces of the dataset and determines by how much each observation deviates from its k-nearest neighbors, where k is defined by the user. The result is a vector of outlier scores for each observation. How far an observation's score deviates from zero determines the degree to which it is an outlier. In this paper an outlier will be considered an observation with score 0.08 or above[3]. The detection algorithm was performed on the data, and the distribution of the outlier scores is shown in Table 3.4.

Table 3.4: Distribution of outlier scores

Min	1^{st} Quartile	Median	Mean	3^{rd} Quartile	Max
0.02	0.06	0.11	0.14	0.17	0.71

The large interquartile range gives evidence of a wide distribution of outlier scores so the median will be used as a measure of center. The median is 0.11, which is higher than the threshold of 0.8. This also means at least half of the observations are considered outliers, which supports the results produced by DBSCAN and provides

[3] There does not exist an explicit cutoff which determines whether an observation qualifies as an outlier, but the authors provide an example in which the observation they constructed as an outlier received a score of 0.08.

further evidence that the distribution of countries is scattered and unfit for basic k-means.

3.4 Robust Sparse k-Means

One way to remedy the issue of sparse data is through an algorithm that assigns weights to each feature called robust sparse k-means (RSKC). The weighted features adjust the distances between observations such that those that are weighted more heavily force the clustering algorithm to discriminate each cluster by those features more than those that are weighted less. This method solves the issues of sparse data and over-representation of certain features. The sparse k-means clustering procedure was introduced by Witten and Tibshirani (2010). Robustness was later added by Kondo et al. (2016).

3.4.1 Theory

Let n be the number of countries and $d_{i,i'}$ the Euclidean distance between $\mathbf{x}_i$ and $\mathbf{x}_{i'}$, the vectorized versions of countries i and $i^{'}$, respectively. Then the total sum of squares is equal to $tss := \frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{i,i'}^2$.

Let $\mathcal{C} := \{C_1, C_2, ..., C_K\}$ be a set of K clusters. Then the total within sum of squares is $wss := \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i'\in C_k} d_{i,i'}^2$, where n_k denotes the number of countries in C_k. It follows that the between sum of squares is $bss := tss - wss$. The additive property of bss allows it to be broken down to the component level. Let p denote the number of variables used to represent a country. Then bss can be redefined as

$$\sum_{j=1}^{p}\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{(i,i')_j}^2 - \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i'\in C_k} d_{(i,i')_j}^2\right),$$

where $d_{(i,i')_j}$ refers to the Euclidean distance between the j^{th} components of $\mathbf{x}_i$ and $\mathbf{x}_{i'}$. Let $\mathbf{w}$ be the vector of weights assigned to each variable. Then the optimization problem solved by sparse k-means is as follows:

$$\max_{C_1,C_2,...,C_k;\mathbf{w}}\sum_{j=1}^{p} w_j\left(\frac{1}{n}\sum_{i=1}^{n}\sum_{i'=1}^{n} d_{(i,i')_j}^2 - \sum_{k=1}^{K}\frac{1}{n_k}\sum_{i,i'\in C_k} d_{(i,i')_j}^2\right)$$

$$\|\mathbf{w}\|_2 \leq 1, \|\mathbf{w}\|_1 \leq l, w_j \geq 0,$$

where $l > 1$ determines the degree of sparsity of the solution.

3.4.2 Results

The outlier detection test was performed on the weighted dataset and the distribution of scores is presented in Table 3.5. The distribution shows a marked shift left in outlier scores, as most fall between 0.01 and 0.04, which is well within the range of points that do not qualify as outliers. In fact, the maximum outlier score for the weighted data is equivalent to the mean score of the unweighted dataset. Thus, RSKC was successful in reducing noise and forming a denser distribution of countries. The silhouette index is computed for the taxonomy generated by RSKC and is shown to be 0.28 compared to the 0.19 produced by basic k-means, indicating a significant improvement in the quality of the taxonomy produced. However, it should be noted that a score of 0.28 corresponds to a weak clustering. This is an improvement over basic k-means in that the taxonomy is substantial, but RSKC still fails to produce clearly distinct clusters. Thus, RSKC not only reduces the noise in the dataset, but it also produces a more accurate taxonomy than basic k-means.

Table 3.5: Distribution of outlier scores for the weighted data

Min	1$^{\text{st}}$ Quartile	Median	Mean	3$^{\text{rd}}$ Quartile	Max
0.00	0.01	0.02	0.03	0.04	0.08

The optimal weight vector $\mathbf{w}^*$ contains the weights placed on each variable. Table 3.6 provides a comprehensive list of each variable with its respective weight. Weights signify the relative importance of the variables in discriminating clusters. That is, the greater the weight placed on a variable, the more influence it has on the final partition produced by RSKC. The five most weighted variables are *gni*, *num_doc*, *avg_yr_sch*, *mort_rt_inf*, and *imz_msls*, which cover R&D environment, public health, education, and standard of living. The five least-weighted features are *cost_biz*, *nat_res_rent*,

Variable	Weight	Variable	Weight
cost_biz	0.03	mort_rt_ad	0.19
proc_req	0.21	physint	0.24
tim_req	0.10	formov	0.15
num_doc	0.36	dommov	0.10
hidx	0.07	injud	0.28
imz_dpt	0.18	nat_res_rent	0.05
imz_msls	0.25	povHCR	0.19
lf_expec	0.28	avg_yr_sch	0.32
low_bw	0.22	gni	0.40
mort_rt_inf	0.29		

Table 3.6: Optimal weight assignments produced by RSKC

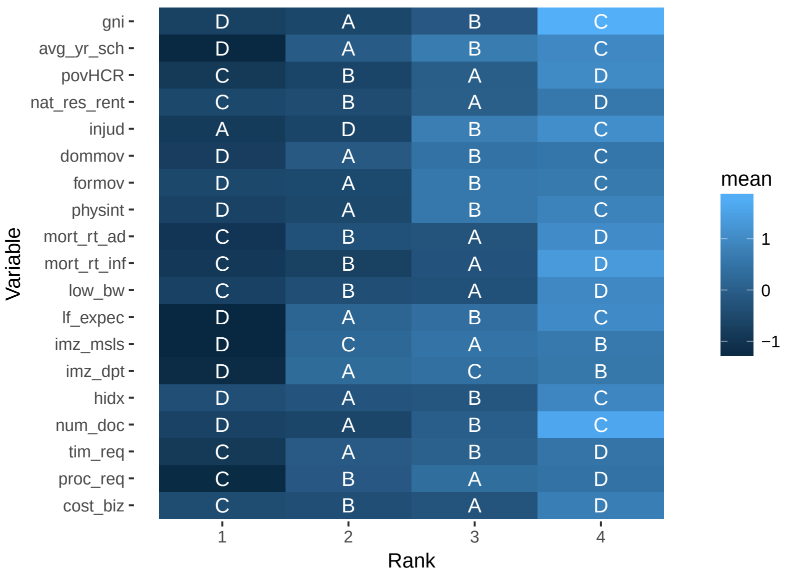
hidx, *tim_req*, and *dommov*, which cover human rights, economic stability, and investment environment.

A full description of the development taxonomy produced by RSKC including which countries belong to what clusters can be found in Table A.6. Figure 3.3 provides a summary of the taxonomy. The rank is determined by comparing the cluster means based on the unweighted normalized data. Descriptions of each cluster are as follows:

- Cluster D characterizes countries considered to be **underdeveloped**. The majority of these countries have hostile business environments, almost non-existent research and development communities, and very high poverty levels. Their heavy reliance on natural resources may feed in to their high levels of corruption and historically poor human rights records. In addition, their populations are generally poorly educated and unhealthy. The countries in Cluster D can be characterized best by their research and development environments, standards of living, and levels of education.

- Cluster A characterizes countries considered to be **developing**. Typically, a country in this group hosts a poor business environment, unproductive but influential research community[4], and medium-high levels of poverty. Their economies are adequately diversified, and their corruption and human rights records are mediocre. In addition, their populations are somewhat educated and healthy. The countries in Cluster A can be characterized best by their research environments, standards of living, investment environment, and levels of public health.

[4] On average, few papers were published in these countries, but those that were received a disproportionate number of citations



White text indicates cluster group name
Rank: 1=lowest value 4=highest value

Figure 3.3: Description of the taxonomy produced by RSKC

- Cluster B characterizes countries considered to be **fairly developed**. These countries host a welcoming business environment, productive but only adequately influential research environment, and low levels of poverty. Their economies are diversified and their human rights and corruptions records fairly clean. Their populations are fairly educated and healthy. The countries in Cluster B can be characterized best by their levels of public health and business environments, and human rights records.

- Cluster C characterizes countries considered to be **fully developed**. These countries host very friendly business environments, have productive and influential research communities, and claim very low rates of poverty. Their economies are highly diversified and their human rights and corruption records are very clean. Their populations are well-educated and boast excellent health. The countries in Cluster C can be characterized best by their human rights records, investment environments, and levels of public health.

The term "best" used in the last sentence of each cluster description is defined as the set of variables which show the smallest within-cluster standard deviation using the normalized unweighted data (for within-cluster summary statistics, see Table A.7). Low within-cluster standard deviation for a given variable in a cluster indicates that most countries in that cluster share similar values for that variable. Thus, the cluster as a whole can be best characterized by those variables that produce the smallest within-cluster standard deviations. Conversely, there exist variables whose values do not properly characterize an entire cluster. That is, there exist countries within each cluster that do not align exactly with their respective cluster profiles. For example, the majority of countries in Cluster D have values for *nat_res_rent* that range between 4.18% and 15.8% of GDP, except for Iraq which towers these figures with 61.61% of GDP. The same is true for the human rights and corruption indexes. The fact that their values are defined on small discrete intervals make it such that clusters which on average have clean human rights records (i.e., Cluster C) include countries with mediocre scores in $injud$ and $formov$.

Chapter 4

CONCLUSION

This paper investigated the question of whether an empirical approach to producing a development taxonomy could rival the techniques currently in practice. Kernel density estimation was implemented to generate a univariate taxonomy which could be directly compared to that published by the World Bank. Internal validity measurements confirmed the superiority of this data-driven taxonomy. With the base case established, the analysis extended to higher dimensions in an attempt to create a development taxonomy. Development was defined by a set of variables which fall under six broad categories typically associated with growth. Two k-means clustering methods were used to create a multivariate development taxonomy. The first was basic k-means, which after performing an internal validation measure showed the taxonomy it produced to be unusable. Several density tests were performed on the underlying data which showed high levels of noise. This issue was resolved using a k-means algorithm that accounts for outliers and noise called RSKC. Never before used in the creation of a development taxonomy, RSKC produced a taxonomy strong enough for use in economic analysis, but still fairly weak in defining distinct and cohesive development groups. This latter point brings into question the efficacy of a data-driven approach to produce a development taxonomy.

Further work can be done to find development indicators for countries whose values more naturally form well separated clusters. Once that data is available, implementation would not be difficult given the flexibility of the clustering algorithms presented. The efficacy of the model proposed can also be tested by performing comparisons of multiple development taxonomies spanning several decades to observe changes in the composition of the groups. Countries that ascend to clusters which represent more developed countries can be analyzed to determine if the model correctly identified upward growth.

References

Alatas, V., and Cameron, L. A. (2008, jan). The Impact of Minimum Wages on Employment in a Low-Income Country: A Quasi-Natural Experiment in Indonesia. *ILR Review*, *61*(2), 201–223. Retrieved from http://journals.sagepub.com/doi/10.1177/001979390806100204 doi: 10.1177/001979390806100204

Barro, R. J., and Lee, J. W. (2013). A new data set of educational attainment in the world, 1950-2010. *Journal of Development Economics*, *104*, 184–198. doi: 10.1016/j.jdeveco.2012.10.001

Cingranelli, D. L., Richards, D. L., and Clay, K. C. (2014). *The CIRI Human Rights Dataset.* Retrieved from http://www.humanrightsdata.com

Coccia, M. (2007). A New Taxonomy of Country Performance and Risk Based on Economic and Technological Indicators. *Journal of Applied Economics*, *10*(1), 29–42.

Ester, M., Kriegel, H. P., Sander, J., and Xu, X. (1996). A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise. *Proceedings of the 2nd International Conference on Knowledge Discovery and Data Mining*, 226–231. Retrieved from https://www.aaai.org/Papers/KDD/1996/KDD96-037.pdf doi: 10.1.1.71.1980

Filmer, D., and Schady, N. (2014, jul). The Medium-Term Effects of Scholarships in a Low-Income Country. *Journal of Human Resources*, *49*(3), 663–694. Retrieved from http://jhr.uwpress.org/lookup/doi/10.3368/jhr.49.3.663 doi: 10.3368/jhr.49.3.663

Hanushek, E. a. (2007). The Role of Education Quality in Economic Growth The Role of School Improvement in Economic Development. *Humanities*, *46*, 607–677. Retrieved from http://papers.ssrn.com/sol3/papers.cfm?abstract _id=960379 doi: 10.2139/ssrn.960379

International Development Association. (2001). *IDA Eligibility, Terms and Graduation Policies* (Tech. Rep. No. January). World Bank.

Kaufman, L., and Rousseeuw, P. J. (1990). *Finding Groups in Data: An Introduction to Cluster Analysis (Wiley Series in Probability and Statistics).* Retrieved from http://books.google.com/books?hl=en&lr=&id=YeFQHiikNo0C&oi=fnd&pg=PR11&dq=Finding+Groups+in+Data+-+An+introduction+to+Cluster+Analysis&ots=5zp9F4PGxF&sig=SeUYzccb34LjgB8 doi: 10.1007/s13398-014-0173-7.2

Knack, S., Rogers, F. H., and Heckelman, J. C. (2012). Crossing the threshold: A positive analysis of IBRD graduation policy. *Review of International Organizations*, *7*(2), 145–176. doi: 10.1007/s11558-011-9136-3

Kondo, Y., Salibian-Barrera, M., and Zamar, R. (2016). **RSKC**: An *R* Package for a Robust and Sparse K-Means Clustering Algorithm. *Journal of Statistical Software*, *72*(5). Retrieved from http://www.jstatsoft.org/v72/i05/ doi: 10.18637/jss.v072.i05

Kriege, H. P., Kröger, P., Schubert, E., and Zimek, A. (2009). Outlier detection in axis-parallel subspaces of high dimensional data. In *Lecture notes in computer science (including subseries lecture notes in artificial intelligence and lecture notes in bioinformatics)* (Vol. 5476 LNAI, pp. 831–838). doi: 10.1007/978-3-642-01307-2_86

Lochner, L., and Moretti, E. (2004). The effect of education on crime: Evidence from prison inmates, arrests, and self-reports. *American Economic Review*, *94*(1), 155–189. doi: 10.1257/000282804322970751

MacKay, D. J. C. (2005). *Information Theory, Inference, and Learning Algorithms David J.C. MacKay* (Vol. 100). Retrieved from http://www.cambridge.org/0521642981 doi: 10.1198/jasa.2005.s54

Mauro, P. (1995). Corruption and Growth. *The Quarterly Journal of Economics*, *110*(3), 681–712. Retrieved from https://academic.oup.com/qje/article

-lookup/doi/10.2307/2946696 doi: 10.2307/2946696

Mo, P. H. (2001). Corruption and Economic Growth. *Journal of Comparative Economics*, *29*(1), 66–79. doi: 10.1006/jcec.2000.1703

Mumssen, C., Fabrizio, S., Lane, C., Mukhopadhyay, B., Bal Gündüz, Y., Bersch, J., . . . Yang, F. (2012). *Review of Facilities for Low-Income Countries* (Tech. Rep.). International Monetary Fund. Retrieved from http://www.imf.org/external/np/pp/eng/2012/072612.pdf

OECD. (n.d.). Poverty rate. Retrieved from /content/indicator/0fe1315d-en doi: http://dx.doi.org/10.1787/0fe1315d-en

Parzen, E. (1962). On Estimation of a Probability Density Function and Mode. *The Annals of Mathematical Statistics*, *33*(3), 1065–1076. Retrieved from http://projecteuclid.org/euclid.aoms/1177704472 doi: 10.1214/aoms/1177704472

Rosenblatt, M. (1956). Remarks on Some Nonparametric Estimates of a Density Function. *The Annals of Mathematical Statistics*, *27*(3), 832–837. Retrieved from http://projecteuclid.org/euclid.aoms/1177728190 doi: 10.1214/aoms/1177728190

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, *20*(C), 53–65. doi: 10.1016/0377-0427(87)90125-7

Schultz, T. P. (2010). Population and health policies. In D. Rodrik and M. R. Rosenzweig (Eds.), *Handbook of development economics* (pp. 4785–4881). Oxford: North-Holland.

SCImago. (2016). SJR - SCImago Journal & Country Rank. *SJR - SCImago Journal & Country Rank*, 1–4. Retrieved from http://www.scimagojr.com/journalsearch.php?q=21100228129&tip=sid&clean=0 doi: 10.3145/epi.2008.nov.12

Sen, A. (1999). *Development as freedom.* New York: Alfred Knopf.

Silles, M. A. (2009). The causal effect of education on health: Evidence from the

United Kingdom. *Economics of Education Review*, *28*(1), 122–128. doi: 10.1016/j.econedurev.2008.02.003

The World Bank. (2000–2017). *World Development Indicators (2000-2010).* Retrieved from http://databank.worldbank.org/data/reports.aspx?source=2&type=metadata&series=SI.POV.GINI#advancedDownloadOptions

UNICEF. (2000–2014). *Low Birthweight database.* Retrieved from https://data.unicef.org/topic/nutrition/low-birthweight/

United Nation General Assembly. (1948). *Universal Declaration of Human Rights.* Retrieved from http://www.un.org/en/documents/udhr/ doi: 10.1080/13642989808406748

Vázquez, S. T., and Sumner, A. (2012). Beyond low and middle income countries: What if there were five clusters of developing countries? *IDS Working Papers*, *2012*(404), 1-40. Retrieved from https://onlinelibrary.wiley.com/doi/abs/10.1111/j.2040-0209.2012.00404.x doi: 10.1111/j.2040-0209.2012.00404.x

Witten, D. M., and Tibshirani, R. (2010). A framework for feature selection. *American Statistician*, *105*(490), 713–726. Retrieved from http://www.pubmedcentral.nih.gov/articlerender.fcgi?artid=2930825&tool=pmcentrez&rendertype=abstract doi: 10.1198/jasa.2010.tm09415.A

World Bank. (1989, jan). *Per capita income : estimating internationally comparable numbers* (Tech. Rep.). The World Bank. Retrieved from http://documents.worldbank.org/curated/en/496091468180250433/Per-capita-income-estimating-internationally-comparable-numbers

Zhang, Z., and Gao, Y. (2015). *Emerging Market Heterogeneity: Insights from Cluster and Taxonomy Analysis* (Vol. 15-155).

Appendix

APPENDIX

Table A.1: Composition of univariate taxonomy produced by KDE

Group 1	Group 1 (cont.)	Group 2	Group 3	Group 4	Group 5
Liberia	Yemen, Rep.	Peru	Romania	Chile	Croatia
Burundi	Senegal	El Salvador	Costa Rica	Turkey	Hungary
Ethiopia	Pakistan	Swaziland	Panama	Equatorial Guinea	Estonia
Congo, Dem. Rep.	Mauritania	Albania	Botswana	Mexico	Oman
Niger	Nigeria	Algeria	Uruguay	Poland	St. Kitts and Nevis
Afghanistan	Sao Tome and Principe	Thailand	Argentina	Libya	Antigua and Barbuda
Madagascar	Cote d'Ivoire	Ecuador	St. Vincent and the Grenadines	Lithuania	Seychelles
Eritrea	Solomon Islands	Macedonia, FYR	Dominica	Latvia	Slovak Republic
Guinea	Cameroon	Tunisia	Gabon	Palau	Trinidad and Tobago
Myanmar	Moldova	Iran, Islamic Rep.	Lebanon		Czech Republic
Malawi	Bolivia	Colombia	South Africa		Saudi Arabia
Sierra Leone	Mongolia	Bosnia and Herzegovina	Russian Federation		Barbados
Rwanda	Congo, Rep.	Dominican Republic	Malaysia		Malta
Nepal	Lesotho	Belarus	Mauritius		Bahrain
Uganda	Nicaragua	Marshall Islands	Venezuela, RB		Portugal
Central African Republic	Timor-Leste	Fiji	St. Lucia		Korea, Rep.
Mozambique	Egypt, Arab Rep.	Tuvalu	Grenada		Slovenia
Tajikistan	Bhutan	Belize			Israel
Togo	Guyana	Kazakhstan			Bahamas, The
Gambia, The	Honduras	Namibia			Greece
Zimbabwe	Indonesia	Serbia			Cyprus
Burkina Faso	Sri Lanka	Suriname			New Zealand
Guinea-Bissau	Paraguay	Jamaica			Brunei Darussalam
Tanzania	Angola	Maldives			Spain
Haiti	Syrian Arab Republic	Bulgaria			Singapore
Mali	Philippines	Montenegro			Australia
Kyrgyz Republic	Kiribati	Cuba			Italy
Lao PDR	Georgia	Brazil			Canada
Cambodia	Azerbaijan				France
Chad	Turkmenistan				Germany
Bangladesh	Ukraine				United Arab Emirates
Kenya	Vanuatu				Japan
Ghana	Armenia				Belgium
Uzbekistan	Iraq				Austria
Benin	China				Kuwait
Comoros	Guatemala				Finland
Papua New Guinea	Cabo Verde				Andorra
Zambia	Morocco				San Marino
Sudan	Kosovo				United Kingdom
Vietnam	Jordan				Netherlands
India	Micronesia, Fed. Sts.				United States
	Samoa				Ireland
	Tonga				Sweden
					Qatar
					Denmark
					Iceland
					Switzerland
					Norway
					Luxembourg
					Liechtenstein
					Monaco

Figure A.1: Map of countries sampled for multivariate analysis

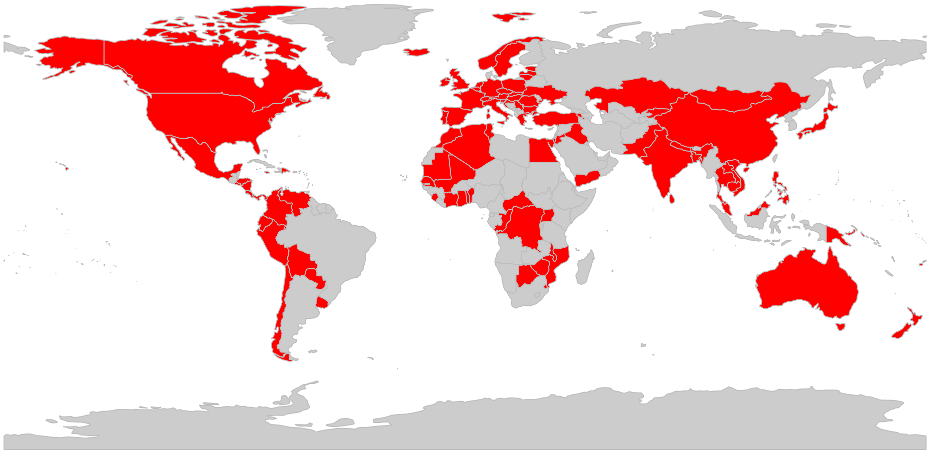


Table A.2: World Bank Income Groups, 2006

Low Income	Lower Middle Income	Upper Middle Income	Upper Income
Liberia	Sao Tome and Principe	Dominican Republic	Croatia
Burundi	Cote d'Ivoire	Belarus	Hungary
Ethiopia	Solomon Islands	Marshall Islands	Estonia
Congo, Dem. Rep.	Cameroon	Fiji	Oman
Niger	Moldova	Tuvalu	St. Kitts and Nevis
Afghanistan	Bolivia	Belize	Antigua and Barbuda
Madagascar	Mongolia	Kazakhstan	Seychelles
Eritrea	Congo, Rep.	Namibia	Slovak Republic
Guinea	Lesotho	Serbia	Trinidad and Tobago
Myanmar	Nicaragua	Suriname	Czech Republic
Malawi	Timor-Leste	Jamaica	Saudi Arabia
Sierra Leone	Egypt, Arab Rep.	Maldives	Barbados
Rwanda	Bhutan	Bulgaria	Malta
Nepal	Guyana	Montenegro	Bahrain
Uganda	Honduras	Cuba	Portugal
Central African Republic	Indonesia	Brazil	Korea, Rep.
Mozambique	Sri Lanka	Romania	Slovenia
Tajikistan	Paraguay	Costa Rica	Israel
Togo	Angola	Panama	Bahamas, The
Gambia, The	Syrian Arab Republic	Botswana	Greece
Zimbabwe	Philippines	Uruguay	Cyprus
Burkina Faso	Kiribati	Argentina	New Zealand
Guinea-Bissau	Georgia	St. Vincent and the Grenadines	Brunei Darussalam
Tanzania	Azerbaijan	Dominica	Spain
Haiti	Turkmenistan	Gabon	Singapore
Mali	Ukraine	Lebanon	Australia
Kyrgyz Republic	Vanuatu	South Africa	Italy
Lao PDR	Armenia	Russian Federation	Canada
Cambodia	Iraq	Malaysia	France
Chad	China	Mauritius	Germany
Bangladesh	Guatemala	Venezuela, RB	United Arab Emirates
Kenya	Cabo Verde	St. Lucia	Japan
Ghana	Morocco	Grenada	Belgium
Uzbekistan	Kosovo	Chile	Austria
Benin	Jordan	Turkey	Kuwait
Comoros	Micronesia, Fed. Sts.	Equatorial Guinea	Finland
Papua New Guinea	Samoa	Mexico	Andorra
Zambia	Tonga	Poland	San Marino
Sudan	Peru	Libya	United Kingdom
Vietnam	El Salvador	Lithuania	Netherlands
India	Swaziland	Latvia	United States
Yemen, Rep.	Albania	Palau	Ireland
Senegal	Algeria		Sweden
Pakistan	Thailand		Qatar
Mauritania	Ecuador		Denmark
Nigeria	Macedonia, FYR		Iceland
	Tunisia		Switzerland
	Iran, Islamic Rep.		Norway
	Colombia		Luxembourg
	Bosnia and Herzegovina		Liechtenstein
			Monaco

Table A.3: Correlation matrix of development indicators

	cost_biz	proc_req	tim_req	num_doc	hidx	imz_dpt	imz_msls	lf_expec	low_bw	mort_rt_inf	mort_rt_ad	physint	formov	dommov	injud	nat_res_rent	povHCR	avg_yr_sch	gni
cost_biz	1.00	0.19	0.27	-0.24	-0.16	-0.48	-0.48	-0.54	0.17	0.64	0.44	-0.19	-0.29	-0.37	-0.22	0.17	0.52	-0.46	-0.25
proc_req	0.19	1.00	0.59	-0.56	-0.31	-0.36	-0.29	-0.44	0.35	0.42	0.36	-0.48	-0.42	-0.26	-0.57	0.23	0.42	-0.47	-0.57
tim_req	0.27	0.59	1.00	-0.39	-0.35	-0.42	-0.40	-0.42	0.26	0.39	0.39	-0.24	-0.24	-0.29	-0.30	0.25	0.47	-0.41	-0.42
num_doc	-0.24	-0.56	-0.39	1.00	0.39	0.31	0.25	0.59	-0.38	-0.53	-0.50	0.52	0.38	0.36	0.63	-0.31	-0.53	0.59	0.91
hidx	-0.16	-0.31	-0.35	0.39	1.00	0.19	0.14	0.31	-0.16	-0.28	-0.28	0.08	0.13	0.09	0.33	-0.18	-0.30	0.35	0.44
imz_dpt	-0.48	-0.36	-0.42	0.31	0.19	1.00	0.89	0.64	-0.46	-0.75	-0.53	0.35	0.36	0.46	0.38	-0.49	-0.53	0.60	0.32
imz_msls	-0.48	-0.29	-0.40	0.25	0.14	0.89	1.00	0.63	-0.56	-0.75	-0.49	0.30	0.32	0.48	0.35	-0.46	-0.54	0.63	0.25
lf_expec	-0.54	-0.44	-0.42	0.59	0.31	0.64	0.63	1.00	-0.45	-0.93	-0.95	0.38	0.34	0.52	0.48	-0.33	-0.78	0.74	0.62
low_bw	0.17	0.35	0.26	-0.38	-0.16	-0.46	-0.56	-0.45	1.00	0.56	0.30	-0.46	-0.28	-0.24	-0.34	0.25	0.37	-0.59	-0.38
mort_rt_inf	0.64	0.42	0.39	-0.53	-0.28	-0.75	-0.75	-0.93	0.56	1.00	0.79	-0.42	-0.37	-0.54	-0.51	0.33	0.77	-0.83	-0.54
mort_rt_ad	0.44	0.36	0.39	-0.50	-0.28	-0.53	-0.49	-0.95	0.30	0.79	1.00	-0.28	-0.25	-0.44	-0.35	0.27	0.73	-0.58	-0.52
physint	-0.19	-0.48	-0.24	0.52	0.08	0.35	0.30	0.38	-0.46	-0.42	-0.28	1.00	0.66	0.43	0.51	-0.26	-0.35	0.50	0.53
formov	-0.29	-0.42	-0.24	0.38	0.13	0.36	0.32	0.34	-0.28	-0.37	-0.25	0.66	1.00	0.46	0.51	-0.34	-0.37	0.41	0.40
dommov	-0.37	-0.26	-0.29	0.36	0.09	0.46	0.48	0.52	-0.24	-0.54	-0.44	0.43	0.46	1.00	0.46	-0.27	-0.46	0.41	0.37
injud	-0.22	-0.57	-0.30	0.63	0.33	0.38	0.35	0.48	-0.34	-0.51	-0.35	0.51	0.51	0.46	1.00	-0.25	-0.53	0.57	0.63
nat_res_rent	0.17	0.23	0.25	-0.31	-0.18	-0.49	-0.46	-0.33	0.25	0.33	0.27	-0.26	-0.34	-0.27	-0.25	1.00	0.28	-0.33	-0.29
povHCR	0.52	0.42	0.47	-0.53	-0.30	-0.53	-0.54	-0.78	0.37	0.77	0.73	-0.35	-0.37	-0.46	-0.53	0.28	1.00	-0.72	-0.53
avg_yr_sch	-0.46	-0.47	-0.41	0.59	0.35	0.60	0.63	0.74	-0.59	-0.83	-0.58	0.50	0.41	0.41	0.57	-0.33	-0.72	1.00	0.60
gni	-0.25	-0.57	-0.42	0.91	0.44	0.32	0.25	0.62	-0.38	-0.54	-0.52	0.53	0.40	0.37	0.63	-0.29	-0.53	0.60	1.00

Table A.4: Composition of development taxonomy produced by basic k-means

Cluster A	Cluster B	Cluster C	Cluster D
Algeria	Australia	Botswana	Benin
Armenia	Austria	Bulgaria	Central African Republic
Bangladesh	Belgium	Chile	Congo, Dem. Rep.
Bolivia	Canada	Costa Rica	Congo, Rep.
Cambodia	France	Croatia	Cote d'Ivoire
China	Germany	Czech Republic	India
Colombia	Iceland	Estonia	Iraq
Dominican Republic	Ireland	Fiji	Lao PDR
Ecuador	Israel	Greece	Malawi
Egypt, Arab Rep.	Italy	Hungary	Mali
El Salvador	Japan	Jamaica	Mauritania
Gambia, The	Netherlands	Latvia	Mozambique
Ghana	New Zealand	Lithuania	Pakistan
Honduras	Norway	Mexico	Papua New Guinea
Jordan	Sweden	Mongolia	Senegal
Kazakhstan	Switzerland	Panama	Sierra Leone
Kyrgyz Republic	United Kingdom	Peru	Swaziland
Malaysia	United States	Poland	Togo
Morocco		Portugal	Uganda
Nepal		Romania	Yemen, Rep.
Nicaragua		Slovak Republic	Zimbabwe
Paraguay		Slovenia	
Philippines		Spain	
Serbia		Ukraine	
Sri Lanka		Uruguay	
Thailand			
Tunisia			
Turkey			
Venezuela, RB			
Vietnam			

Table A.5: Summary statistics for taxonomy produced by basic *k*-means

Cluster	Variable	Mean	St. Dev.	Min	Max	Cluster	Variable	Mean	St. Dev.	Min	Max
A	cost_biz	56.76	67.28	5.10	292.10	A	mort_rt_ad	168.49	50.49	93.11	285.08
B	cost_biz	3.94	4.70	0.20	20.00	B	mort_rt_ad	74.92	12.21	63.20	111.32
C	cost_biz	13.68	9.83	2.80	44.20	C	mort_rt_ad	160.72	83.57	73.22	468.51
D	cost_biz	229.63	331.70	15.80	1314.60	D	mort_rt_ad	347.64	138.76	179.46	653.92
A	proc_req	10.63	2.91	6.00	17.00	A	physint	3.17	1.76	0.00	6.00
B	proc_req	5.33	2.11	2.00	9.00	B	physint	6.78	1.56	2.00	8.00
C	proc_req	8.80	2.29	5.00	15.00	C	physint	6.04	1.62	1.00	8.00
D	proc_req	11.43	2.27	6.00	17.00	D	physint	3.43	2.04	0.00	6.00
A	tim_req	40.15	27.78	7.00	143.00	A	formov	1.10	0.55	0.00	2.00
B	tim_req	13.22	7.66	3.00	27.00	B	formov	1.89	0.32	1.00	2.00
C	tim_req	37.14	24.89	7.50	105.00	C	formov	1.92	0.28	1.00	2.00
D	tim_req	57.05	32.81	22.00	133.00	D	formov	1.10	0.70	0.00	2.00
A	num_doc	70.40	85.13	4.91	333.46	A	dommov	1.47	0.68	0.00	2.00
B	num_doc	1952.59	798.67	32.79	3763.26	B	dommov	2.00	0.00	2.00	2.00
C	num_doc	523.89	476.51	23.47	1698.17	C	dommov	2.00	0.00	2.00	2.00
D	num_doc	16.12	11.35	0.37	40.17	D	dommov	1.10	0.77	0.00	2.00
A	hidx	239.90	242.57	24.00	1059.00	A	injud	0.33	0.55	0.00	2.00
B	hidx	561.78	492.75	64.00	1965.00	B	injud	1.94	0.24	1.00	2.00
C	hidx	193.52	181.01	18.00	839.00	C	injud	1.44	0.65	0.00	2.00
D	hidx	110.14	71.90	16.00	251.00	D	injud	0.43	0.68	0.00	2.00
A	imz_dpt	92.13	6.40	71.00	99.00	A	nat_res_rent	6.84	7.97	0.09	26.84
B	imz_dpt	94.39	3.91	83.00	99.00	B	nat_res_rent	1.60	3.08	0.00	11.72
C	imz_dpt	96.08	2.83	88.00	99.00	C	nat_res_rent	4.34	8.05	0.05	35.08
D	imz_dpt	72.71	11.58	51.00	99.00	D	nat_res_rent	18.49	19.33	2.08	61.61
A	imz_msls	92.17	6.09	78.00	99.00	A	povHCR	26.53	13.35	2.03	62.19
B	imz_msls	90.94	5.26	80.00	97.00	B	povHCR	11.07	3.52	5.80	18.10
C	imz_msls	95.48	3.02	87.00	99.00	C	povHCR	16.61	7.50	5.60	34.60
D	imz_msls	70.14	10.20	48.00	93.00	D	povHCR	45.11	15.36	18.90	72.30
A	lf_expec	70.30	4.38	58.33	74.42	A	avg_yr_sch	7.27	2.08	3.25	11.60
B	lf_expec	80.24	1.09	77.69	82.32	B	avg_yr_sch	11.09	0.95	9.15	12.86
C	lf_expec	73.47	5.54	53.59	80.82	C	avg_yr_sch	10.03	1.39	7.02	12.73
D	lf_expec	55.75	7.22	44.55	68.26	D	avg_yr_sch	4.10	1.55	1.28	7.47
A	low_bw	10.30	4.89	2.22	22.00	A	gni	2403.67	1827.34	340.00	7820.00
B	low_bw	6.56	1.38	4.00	10.00	B	gni	43104.44	11740.16	22060.00	69980.00
C	low_bw	7.83	2.42	5.00	13.70	C	gni	9520.40	7007.81	1120.00	27970.00
D	low_bw	16.94	7.93	9.00	35.00	D	gni	772.86	657.63	230.00	3050.00
A	mort_rt_inf	26.99	13.18	6.80	55.90						
B	mort_rt_inf	4.12	1.05	2.30	6.70						
C	mort_rt_inf	11.70	9.45	3.30	42.90						
D	mort_rt_inf	71.99	20.95	32.60	124.40						

Table A.6: Composition of development taxonomy produced by RSKC

Cluster A	Cluster B	Cluster C	Cluster D
Algeria	Botswana	Australia	Bangladesh
Armenia	Bulgaria	Austria	Benin
Bolivia	Chile	Belgium	Cambodia
China	Costa Rica	Canada	Central African Republic
Colombia	Croatia	France	Congo, Dem. Rep.
Dominican Republic	Czech Republic	Germany	Congo, Rep.
Ecuador	Estonia	Iceland	Cote d'Ivoire
Egypt, Arab Rep.	Fiji	Ireland	Gambia, The
El Salvador	Greece	Israel	Ghana
Honduras	Hungary	Italy	India
Jordan	Jamaica	Japan	Iraq
Kazakhstan	Latvia	Netherlands	Lao PDR
Kyrgyz Republic	Lithuania	New Zealand	Malawi
Malaysia	Peru	Norway	Mali
Mexico	Poland	Sweden	Mauritania
Mongolia	Portugal	Switzerland	Mozambique
Morocco	Romania	United Kingdom	Nepal
Nicaragua	Slovak Republic	United States	Pakistan
Panama	Slovenia		Papua New Guinea
Paraguay	Spain		Senegal
Philippines	Sri Lanka		Sierra Leone
Serbia	Uruguay		Swaziland
Thailand			Togo
Tunisia			Uganda
Turkey			Yemen, Rep.
Ukraine			Zimbabwe
Venezuela, RB			
Vietnam			

Table A.7: Summary statistics for taxonomy produced by RSKC

Cluster	Variable	Mean	St. Dev	Min	Max	Cluster	Variable	Mean	St. Dev	Min	Max
A	cost_biz	35.99	37.69	5.10	145.50	A	mort_rt_ad	162.62	48.74	93.11	270.99
B	cost_biz	15.03	11.99	2.80	44.20	B	mort_rt_ad	154.52	83.38	73.22	468.51
C	cost_biz	3.94	4.70	0.20	20.00	C	mort_rt_ad	74.92	12.21	63.20	111.32
D	cost_biz	212.64	302.30	15.80	1314.60	D	mort_rt_ad	323.86	135.52	150.56	653.92
A	proc_req	10.57	3.11	6.00	17.00	A	physint	3.54	1.88	0.00	7.00
B	proc_req	9.00	2.31	5.00	15.00	B	physint	6.14	1.52	2.00	8.00
C	proc_req	5.33	2.11	2.00	9.00	C	physint	6.78	1.56	2.00	8.00
D	proc_req	10.96	2.34	6.00	17.00	D	physint	3.23	2.03	0.00	6.00
A	tim_req	36.70	28.10	7.00	143.00	A	formov	1.18	0.61	0.00	2.00
B	tim_req	40.95	24.78	7.50	105.00	B	formov	1.86	0.35	1.00	2.00
C	tim_req	13.22	7.66	3.00	27.00	C	formov	1.89	0.32	1.00	2.00
D	tim_req	53.94	31.93	17.00	133.00	D	formov	1.15	0.67	0.00	2.00
A	num_doc	82.66	86.17	4.91	333.46	A	dommov	1.57	0.69	0.00	2.00
B	num_doc	579.59	482.09	23.47	1698.17	B	dommov	1.95	0.21	1.00	2.00
C	num_doc	1952.59	798.67	32.79	3763.26	C	dommov	2.00	0.00	2.00	2.00
D	num_doc	18.55	15.57	0.37	75.12	D	dommov	1.15	0.73	0.00	2.00
A	hidx	198.25	186.59	24.00	871.00	A	injud	0.25	0.44	0.00	1.00
B	hidx	212.18	192.82	18.00	839.00	B	injud	1.64	0.49	1.00	2.00
C	hidx	561.78	492.75	64.00	1965.00	C	injud	1.94	0.24	1.00	2.00
D	hidx	158.81	204.56	16.00	1059.00	D	injud	0.46	0.65	0.00	2.00
A	imz_dpt	93.11	6.17	71.00	99.00	A	nat_res_rent	8.43	9.58	0.09	35.08
B	imz_dpt	96.23	2.33	89.00	99.00	B	nat_res_rent	2.76	5.10	0.05	21.44
C	imz_dpt	94.39	3.91	83.00	99.00	C	nat_res_rent	1.60	3.08	0.00	11.72
D	imz_dpt	75.73	12.40	51.00	99.00	D	nat_res_rent	15.58	18.36	1.08	61.61
A	imz_msls	94.07	4.84	81.00	99.00	A	povHCR	25.39	13.17	2.03	62.19
B	imz_msls	95.36	3.17	87.00	99.00	B	povHCR	15.63	7.20	5.60	34.60
C	imz_msls	90.94	5.26	80.00	97.00	C	povHCR	11.07	3.52	5.80	18.10
D	imz_msls	72.81	10.81	48.00	93.00	D	povHCR	42.45	15.42	18.90	72.30
A	lf_expec	71.59	3.06	64.11	76.14	A	avg_yr_sch	7.92	1.73	4.34	11.60
B	lf_expec	73.89	5.48	53.59	80.82	B	avg_yr_sch	10.17	1.38	7.02	12.73
C	lf_expec	80.24	1.09	77.69	82.32	C	avg_yr_sch	11.09	0.95	9.15	12.86
D	lf_expec	57.18	7.31	44.55	68.41	D	avg_yr_sch	4.21	1.54	1.28	7.47
A	low_bw	8.68	3.78	2.22	21.00	A	gni	3032.50	2030.41	500.00	8370.00
B	low_bw	8.48	3.08	5.00	17.00	B	gni	10126.36	7200.40	1370.00	27970.00
C	low_bw	6.56	1.38	4.00	10.00	C	gni	43104.44	11740.16	22060.00	69980.00
D	low_bw	16.57	7.53	9.00	35.00	D	gni	717.69	600.95	230.00	3050.00
A	mort_rt_inf	22.28	8.01	6.80	44.10						
B	mort_rt_inf	10.25	8.88	3.30	42.90						
C	mort_rt_inf	4.12	1.05	2.30	6.70						
D	mort_rt_inf	67.87	20.70	32.60	124.40						