DEd in Comparative Education

TITLE: AUTOMATED CORRELATION DISCOVERY OF NATIONAL EDUCATIONAL DATA WITH OTHER NATIONAL CHARACTERISTICS FROM UN, OECD, AND OTHER INTERNATIONAL DATABASES

A RESEARCH PROPOSAL SUBMITTED BY: Jacob J. Walker

51628813@mylife.unisa.ac.za

11/26/2014

ABSTRACT

This is a proposal for data science / data mining meta-research in the field of Comparative Education. The basic methodology is to gather a large number of data sets about the characteristics of countries of the world, and then dependent upon the type of data within these sets, to compare them to each other, finding a maximum correlation coefficient for each pair of data sets. In the end the data sets that are most intriguing will be explored further, and the full results of looking for correlations between national characteristics of all sorts will be released to assist other researchers in the social sciences.

Automated Correlation Discovery of National Educational Data with Other National Characteristics from UN, OECD, and other International Databases

1.1. THE INTRODUCTION/BACKGROUND

All science starts with the act of discovery and then the development of a hypothesis based upon what is discovered. Yet, until recently the process of hypothesis creation was generally considered something that came only from intuition and could not derived through a methodological process (Noé, 1998). Richard Feynman was one such person who believed the human intellect was crucial to the hypothesis creation process, but once joked that there could be a machine set up with a random wheel that could make a succession of guessed hypotheses, and then automatically test them (Feynman, 1964). And while this was originally said in jest, now with the computerization of science, this idea is being put into practice, with the process of knowledge discovery in databases (KDD), which includes data mining, being part of "discovery science" and "data science".

KDD, including automated hypothesis development, has been a direct result of an exponential increase in data often called the "information explosion" in tandem with the computerization of this data and the continual increase in processing power of computers. As an early article about KDD said:

> Across a wide variety of fields, data are being collected and accumulated at a dramatic pace. There is an urgent need for a new generation of computational theories and tools to assist humans in extracting useful information (knowledge) from the rapidly growing volumes of digital data. These theories and tools are the subject of the emerging field of knowledge discovery in databases (KDD). (Fayyad, Piatetsky-Shapiro, & Smyth, 1996)

This trend of having more data available has only increased in recent years, and now businesses are clamouring to take advantage of their "big data" (Barth, Earley, Lawson, & Hall, 2013). Yet the field of education is only starting to catch up with where private industry has already gone. For example the International Educational Data Mining Society was only founded in 2011, and the Journal of Educational Data Mining has only published 7 editions to date. This lack of research using data science methodologies provides an opportunity for the proposed research, which will use data mining methods, to make a new and significant contribution to the field of comparative education.

The research being proposed is extremely broad, yet simple in concept. It is to find potential correlations between characteristics of nations. This will be a form of meta-research as it will consist of taking many sources of data about nations, which will be part of a "Compendium of Countries", and then calculate one or more correlation coefficients between each data set. For instance in some preliminary research that has been conducted, using only World Factbook data (Central Intelligence Agency, 2014), there was found to be a relatively strong linear correlation between the expenditure on education per capita and the expenditure on healthcare per capita (r^2= 0.7956) between countries.

Of course this meta-analysis does not answer why the correlation exists. For example, in the case of the expenditure on education and healthcare, there is a strong possibility that there is little causation between those two variables, and that most of the correlation comes from the fact that as a country becomes wealthier (which can partly be seen in GDP per capita), the residents will invest more in both education and healthcare. (And in fact the linear correlation between expenditure on education per capita and GDP per capita is strong with r^2= 0.7027 and the linear correlation between expenditure on healthcare per capita and GDP per capita is also relatively strong with r^2=0.6679).

This initial research has already produced some interesting results, and this proposed research is likely to discover results that could be very valuable to future researchers.

1.2. LITERATURE REVIEW

Comparative education is an interdisciplinary field, and the proposed research spans several fields, requiring the consideration of relevant academic literature from the field of comparative education, both broad and specific, and also literature from the emerging field of educational data science.

1.2.1. Comparative Education

Traditionally, comparative education has four purposes (Noah, 1985):

1. To describe educational systems, processes, or outcomes
2. To assist in the development of educational institutions and practices
3. To highlight the relationships between education and society
4. To establish generalized statements about education that are valid in more than one country.

The proposed research is in line with each of these purposes, where it primarily seeks to find relationships, in the form of quantitative correlations, between education educational systems, processes, and outcomes within countries compared with the corresponding society of the nations. And, if the correlation has a sufficient correlation coefficient, then it can likely be considered a generalized statement, which is clearly valid in more than one country. Further, the results of this research can be used by other researchers and potentially practitioners to assist in the development of educational institutions and practices.

In the Bray and Thomas Framework for Comparative Education Analysis (Bray, Adamson, & Mason, 2007), one dimension that must be considered when conducting research in comparative education is the geographical locational dimension, in which the framework has seven levels:

- Level 1: World Regions/Continents
- Level 2: Countries
- Level 3: States/Provinces
- Level 4: Districts
- Level 5: Schools
- Level 6: Classrooms
- Level 7: Individuals

There is a long tradition of conducting research at Level 1 or Level 2 of this framework, and the proposed research will continue in this tradition, by comparing countries (Level 2) to each other. This level works well for automated knowledge discovery methodologies, as there is an abundance of data sets about the countries of the world and the various characteristics they have.

For example, in the World Factbook alone, there are nearly 100 quantitative characteristics described for 261 entities (most of which are countries), including data about Geography, People and Society, Government, Economy, Energy, Communications, Transportation, Military, and Transnational Issues (Central Intelligence Agency, 2014).

Country data can also be keyed easily, where every country, recognized by the United Nations, has an ISO 3166-1 Alpha 2 code assigned to it, which helps to ensure that the same entities are being used in each of the comparisons conducted.

While the proposed research will only focus on one level of the geographical locational dimension of the Bray and Thomas framework, on the other 2 dimensions, there will be considerable breadth that the research will cover. For instance, in Aspects of Education and of Society, the proposed research will work to find comparable data and correlations within the categories of curriculum, teaching methods, educational finance, management structure, political change, labour markets, and other aspects of educational systems (whether formal or informal). And within these categories, there may be an additional breakdown of ethnic groups, age groups, religious groups, gender groups, etc.

1.2.2. Educational Data Science and Data Mining

The field of data science is relatively new, with methodologies such as data mining were being developed in the 1990's (Frawley, Piatetsky-Shapiro, & Matheus, 1992), and academics started to recognize data science as an independent discipline in the early 2000's (Cleveland, 2001). And, as has been briefly discussed, using these techniques in educational research has only occurred more recently, with the International Educational Data Mining Society being founded in 2011, and the suggestion that educational data science may be becoming its own field has been formalized only this year (Piety, Hickey, & Bishop, 2014).

Much of early data science activities came from institutional research, where educational institutions adopted a wide variety of student information systems and used data from these and related systems to attempt to understand their student behaviour. More recently data science activities have been able to centre more directly on the learner, such as using predictive analytics to personalize computer based instruction, or conducting data mining and analytic research on data produced from video games, cognitive tutors, and learning management systems (LMSs), including massive open online courses (MOOCs).

Broader system-wide data science methods have been used within the context of individual nations. For example, the United States has encouraged much of this form of research by having laws and policies such as "No Child Left Behind" which has required schools to collect data in standardized manners, with the creation of state longitudinal data systems (SLDS). Much of this research has been comparative in nature, but not international, such that it would focus on comparing individual institutions within a state or between states within the United States.

The use of data science specific methodologies with international data has been rarer. The research that does exist, has occurred primarily with internet-based environments, such as MOOCs and social network sites, which have users that span many countries. For example, in the paper "The MOOC phenomenon: who takes massive open online courses and why?" there was a look at which nations students who use MOOCs came from (Christensen et al., 2013). An example of looking at using data mining of social media within a multinational context can be found in "Identification of user patterns in social networks by data mining techniques: Facebook case" (Bozkır, Mazman, & Sezer, 2010). It is likely that there is additional research that has been conducted on social media, search engine, and other international website data which relates directly to education across nations, but that this has not been publicly released, as it would have been conducted for primarily financial gain.

1.2.3. Sources of Data

As noted, there is an immense amount of data that has already been produced regarding the educational and other characteristics of countries. And while there has been some compiled sources, such as The World Factbook, these are relatively small. A larger effort to collect national characteristics is occurring with The Open Compendium of Countries (CompendiumOfCountries.org), a sister project to this proposed doctoral research. This compendium is attempting to have as much existing data about countries of the world from as many sources as possible, gathered into one single database. These data sets will then be used for this thesis, as well as for other research.

The compendium will include as many data sources as it can, even if there is some duplication of data. For instance, the U.N. and the U.S. each have their own estimates for the population of nations. But because these might use slightly different methodologies, the researcher has decided to include both in the initial analysis, and sort out which data sources have the most validity after doing the analysis.

The United Nations will be the primary source of data within The Compendium of Countries. The United Nations Statistics Division recently launched the UNdata website (http://data.un.org) which contains 34 databases and 60 million records about various characteristics of countries. One of these databases is the UIS Data Centre, from the UNESCO Institute for Statistics. According to the UNdata website:

> The UNESCO Institute for Statistics (UIS) provides UNdata with a subset of the more than 1,000 indicators which may be found in the UIS Data Centre. The UIS Data Centre contains indicators and raw data on education, literacy, science, culture and communication. The UIS collects these data from more than 200 countries and international organizations.
>
> The UIS is the primary data source of education, literacy and science data for leading publications and databases, such as: Education For All Global Monitoring Report; World Development Indicators; Human Development Report; State of the World's Children and many others (UNESCO Institute for Statistics, 2014).

Another major source of International data comes from the Organisation for Economic Co-operation and Development (OECD), which not only tracks information about its member states, but also often includes data on non-member states that choose to participate in their activities, such as the Programme for International Student Assessment (PISA). Some of this data is freely available via the OECD.StatExtracts website (http://stats.oecd.org) and the OECD Data Portal (http://data.oecd.org). The researcher will also attempt to use sources that are not as open to the public, as long as this data can still be published openly.

Individual countries also keep track of data about others, with the United States being a prime example, with its World Factbook. The U.S. also tracks other information, such as the U.S. Department of State's Trafficking in Persons (TIP) report. Data sets that can be found from other U.S. agencies will also be included, with www.data.gov being a source of over 131,546 datasets, although most of them are only U.S. focused.

Other data will be drawn from additional Global Performance Indicators (GPIs), which rank, rate, or categorize countries in many different issue areas. Examples include, Transparency International's Corruption Perceptions Index (www.transparency.org/cpi2013/results) and Reporters without Borders' Press Freedom Index (http://rsf.org/index2014). The reliability of some of these indices has been called into question ("How to lie with indices," 2014), and this will be discussed in the Issues of Reliability/Validity section of this proposal.

1.2.4. Some Previously Discovered Educational Correlations

While no known study to date, has compared as many variables as the proposed research will, there have clearly been a tremendous number of studies that have looked at individual variables between nations to attempt to find correlations. While is impossible to cite (or even find) all of these, this proposal will look at some of the relevant educational correlations that have been explored with scores on international achievement tests being a common GPI to compare to other variables.

The three most commonly used achievement tests for this type of analysis are the PISA (Program for International Student Assessment), TIMSS (Trends in Mathematics and Science Study), and PIRLS (Progress in International Reading Study). This type of analysis has generally been done to affect public policy decisions, such that it generally attempts to show one of three things:

- Certain variables, such as spending per pupil, do or do not have an effect on test scores.
- Test scores do or do not predict other outcomes, such as income, or GDP, etc.
- Test scores are or are not valid performance indicators because of how they do or do not correlate with other variables.

The OECD publishes a series of monthly short reports titled PISA in Focus, of which approximately 1/3 of the reports have an analysis of correlations between country characteristics, often involving PISA scores and other characteristics that may affect those scores. One counterintuitive analysis found a negative correlation between hours per week students study science and their average science scores (OECD, 2011). Another counterintuitive correlation analysis suggested that countries with greater perseverance levels of their students did not automatically correlate to having better math scores (OECD, 2014a). Of results that are more intuitive, another analysis found a positive correlation between cumulative expenditures on education in a country and average reading performance, suggesting that for countries that don't invest much, that these can gain a lot with more investment, but that there is a "law of diminishing returns" such that countries that already spend a lot, only see modest increases in scores, if any at all (OECD, 2012). Also, there is a minor positive correlation between competition of schools within a country and their math performance, but with a coefficient of determination of only 0.011 (OECD, 2014b). And there is a fairly clear negative correlation between how much students miss school and how well they perform on the PISA math test (OECD, 2014c).

Correlations that have been explored regarding how improving on the PISA improves a country's economic development, with research conducted by the OECD suggesting that an increase of 100 on the PISA scale (one standard deviation) "would yield an annual growth rate that is 1.74 percentage points higher" (Hanushek & Woessmann, 2010). A meta-analysis of 144 human development indicators (HDIs) compared to the mean of PISA scores for different countries showed those that did and did not have a linear correlation looking using Pearson coefficients. The results were partially mixed, but many HDIs were shown to have a correlation (Baykal, 2014).

But other research has called into question that there is an economic development correlation, by suggesting that there was little to no statistically significant correlation between PISA scores and the G20 Ranks of Creativity, Innovation, and Entrepreneurship using a Spearman Rho test (Tienken & Mullen, 2014). Further, some evidence suggests that there is a high correlation between a country's PISA score and how many of that country's negotiated items were included on the test (Tienken, 2013), which calls into question some of the validity of the PISA (see also Issues of Reliability/Validity).

1.3. THE PROBLEM STATEMENT

It is common for research in comparative education to compare a relatively small set of variables and characteristics between countries, but as of yet, there has not been systemic research conducted to find a large number of potential correlations. By conducting an automated systematic comparison of variables between nations from a collection of machine readable data sets, unexpected correlations may be found, and finding these previously undiscovered correlations can lead to new questions and associated research, which may be able to solve problems in manners not even conceived of.

An additional problem that exists, is that most educational researchers do not tools to conduct such knowledge discovery and data mining activities. The researcher will publicly and freely share the tools used in this research, in the form of Excel spreadsheets, and Visual Basic for Applications (VBA) libraries, such that they can freely be used by other researchers. These tools will also be included as part of the Libre Excel Tools for Data Science that the researcher has been developing, thus giving more "effect for the effort".

Further, while this project will focus on comparing educational characteristics of nations to other characteristics, there is the collateral benefit that the research will also automatically be comparing many non-educational characteristics to each other, and as such there will likely be discoveries that can benefit other fields of social science.

1.4. THE RESEARCH QUESTIONS

Based upon the problem at hand, of there not being a previous systematic search for correlations of characteristics nations, the broad research question can be stated as:

> What are potential correlations that exist between formal and informal educational characteristics of nations, and other characteristics of nations?

Two more specific research question regarding methods is:

> How can Microsoft Excel be used to conduct automated correlation discovery over a large number of data sets, each of which contain data about a significant number of nations?
>
> Can Microsoft Excel also efficiently conduct various forms of correlation discovery, including linear, non-linear, and also correlation discovery for nominal and ordinal data?

These specific research questions are important, because any research using data science methods rely heavily upon effective use of appropriate software. While data science is often conducted with the Python programming language or with statistical packages such as R, Excel is a tool also used by many data scientists, as well as researchers in many other fields. If Excel can effectively be used to perform this type of automated correlation discovery over data sets about nations, the same techniques can be used by other researchers in other fields to conduct knowledge discovery.

1.5. THE AIMS AND OBJECTIVES OF THE STUDY

1.5.1. Research Aim

The aim of this research is to expand the field of comparative education by discovering new potential correlations with educational factors of nations and other characteristics of these nations, including geographic, economic, social, governmental, technological, and potentially even military quantitative characteristics.

1.5.2. Research Objectives

The research tasks can be modelled as a series of 8 stages: data sourcing, data acquisition, data preparation, data transformation, model building, model evaluation, model visualization, and model use, as shown in Figure 1. It should be noted, that given the iterative and parallel nature of data mining, these tasks will not be completed in the exact order shown. Further, the data sourcing, acquisition, preparation, and some transformation will be part of the sister project, the Open Compendium of Countries.

- Develop a research proposal that defines the research project (this is complete)
- **Data Sourcing**: Determine which data sets will be used to create a compendium of countries
- **Data Acquisition**: Collect all data sets in individualized spreadsheets
- **Data Preparation**: Ensure the disparate data in the individualized spreadsheets are standardized, including using the same keying system for countries.
- **Data Transformation**: Integrate individual spreadsheets together, and denormalized the data to prepare of analysis.
- **Model Building**: Use existing Excel formulas to be used to discover potential correlations with nominal and ordinal data, and linear correlations for interval and ratio scaled data, and develop user defined functions (UDFs) or macros in Visual Basic for Applications (VBA) to determine potential non-linear correlations for interval and ratio scaled data.
- **Model Evaluation**: Run each of the models over the data sets, and determine which has the greatest correlation coefficients.
- **Model Visualization**: Graph the interesting discoveries to better understand them.
- **Model Use**: Conduct some deeper level analysis on interesting discoveries, including hypothesis testing and surface level analysis about potential causalities.
- Present the results of the research in the Thesis.

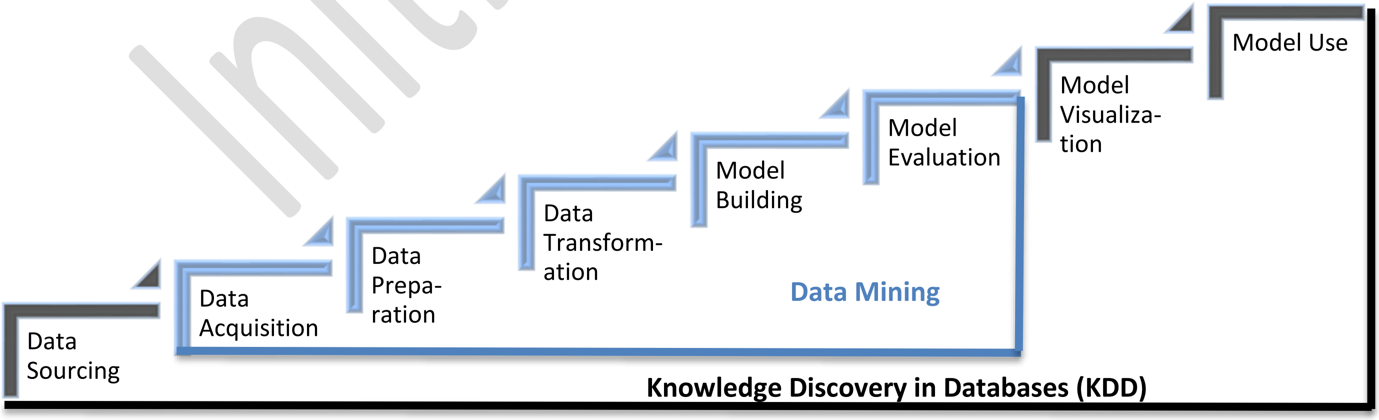

Figure 1: Adapted Generalized Knowledge Discovery in Databases (KDD) Process (Miner, Nisbet, & Elder, 2009, p. 17)

1.6. RESEARCH DESIGN AND METHODOLOGY

This research will be purely quantitative in nature, and focus on regression between 2 sets of variables at a time, using a Knowledge Discovery in Databases (KDD) process.

1.6.1. Data Sourcing

As will be discussed under Issues of Reliability/Validity, this research will attempt to broadly gather datasets. This "everything and the kitchen sink" approach has dramatic implications to interpreting the results of the proposed research, but does not hinder the methodology used within it. This form of data sourcing fits mostly within the Inductive Database theoretical framework for data mining (Mannila, 2000), in which there is an attempt to include all the data available plus all the questions that could be asked about the patterns in the data (Miner et al., 2009, p. 19).

But as there is more data in the world than could be obtained for any research, there will be limits on what will be included and inherent bias in the selection of datasets. The first bias will be purposeful, which is to concentrate on gathering datasets that most directly relate to the field of education, as this is a research proposal for a doctorate in Comparative Education. Thus additional effort will be made to get educational characteristics and indicators of nations, even if additional data transformations are needed to analyse this data. The next bias is more ethical and legal in nature, which is that the proposed research will only be able to use open data, such that not only access can be obtained to it, but that it can also be freely released as part of the results. This issue will be discussed more in the Ethical Considerations portion of this proposal. The last major bias is pragmatic in nature, data that is already machine readable, especially in a flat file or relational format (CSV, XLS, SQL, MDB, etc.), will take less time to include in the research, and thus will likely be represented more than data that is hierarchical or semantic in nature (XML, RDF, etc.), or than data that is contained purely within human readable documents (PDF, DOC, HTML, etc.).

1.6.2. Data Acquisition and Pre-Processing

While traditional data mining often relies upon the selection of data from internal databases, the research being proposed is integrating data from many publicly available databases and data sets. Data that is already in an Excel or CSV format will need a minimal amount of pre-processing. For data that is in a relational database format, such as MySQL or Microsoft Access, individual tables can be exported to sheets and tables within those sheets in Microsoft Excel.

For data that is in PDF format, the Adobe PDF conversion service will be used from within Adobe Reader to convert the data into either a Word or Excel format. For Word documents, tables can generally be copied and pasted easily from the document to Excel. For data that is only web based (in HTML), if the data is contained within a single page, then copy and paste techniques can be used. For data that needs to be gathered over multiple pages, then web scraping techniques may be implemented, using either Excel's built-in web queries, or using VBA combined with the Selenium web testing tool, or possibly writing a web scraper in Python. Since extracting data from these forms of documents is more labour intensive, the researcher may employ the use of research assistants to help with converting this type of data into an Excel format.

Data stored in XML (including RDF/XML files), pose a challenge that XML is inherently a hierarchical language, while data in Excel can at best act somewhat like a relational database. While there are several conversion methods between XML and Relational Models, each has trade-offs (Lee, Mani, & Chu, 2003). To do any necessary conversions, a combination of tools may be employed, including Excel's XML feature or potentially writing an extraction script in Python, or manually converting.

1.6.3. Data Preparation

Given that the data being acquired is coming from many disparate databases, the data needs to be prepared in a manner that it can ultimately all fit into one database. Separate spreadsheets will be used to collect data in a standardized format, and in the end these will be concatenated together.

1.6.3.1. Keying Data

While the data is being copied to Microsoft Excel, it needs to have each country from the source data connected to a unique key for that country. This is important because different source data sets may list a country in different ways. As way of example, the following names all refer to the same country: Republic of Korea; Korea, Republic of; South Korea; Korea, South; and, Korea

An alias lookup table will be use to convert country names to a single ISO-3166 Alpha 2 code as a key, which in the above example would be "KR".

1.6.3.2. Standardizing and Partially Normalizing Data

All data will be copied into a single table with a single data model. Given that the original data may have many different data models, it will not be possible for the data to have full database normalization, further Excel does not lend itself to full database normalization. Thus the following structure attempts to not lose any original data, but also be analysed for the purposes of this research:

Field Name / Column Heading	Description
Source Name	This is the name of the database or source of the data
URI	In most cases this will be the URL of the source database.
Country Name	This will be the ISO English Short Name of the Country
Code	This will generally be the ISO 3166-1 Alpha 2 Code. In some cases for entities that are in the original database, that do not have an ISO-3166-1 Alpha 2 Code
Characteristic of Nation	This will have the "major" field name from the source database. This field can be viewed as a superset of the "Aspects of Education and of Society" dimension in the Bray and Thomas Framework.
Demographic Subgroup	This is a subfield of the characteristic. It may be NULL (blank) if the source data does not have subgroup breakdowns. It may also be a compound field name if more than one subgroup is broken down in the source data (thus being partially de-normalized). This field corresponds to the "Nonlocational Demographic Groups" of the Bray and Thomas Framework.
Year	This is the point in time when the data was collected. If the original data has a finer time-grain than a year, then a fractional year will be specified. For instance July 1, 2010 would be specified as 2010.5
Value	This is the quantitative data. In the case of nominal data, this will be 0 or 1, as a binary/Boolean value for a statement such as "Has Compulsory Education".
Units	This is the type of units the quantitative data represents, such as U.S. dollars, students, teachers, dollars per capita, etc. In some cases units will not exist.
Type of Scale	This is the type of scale and granularity that the data has, which will be used to determine what correlation coefficient is used. Scales will be categorized as: Ratio, Interval, Ordinal, or Nominal (Binary).
Notes	Any additional information that was specified in the original data source that does not fit in one of the other fields, will be entered into this field.

1.6.4. Data Transformation

Once all of the individual spreadsheets have been created from the different data sources, these will be combined together and then have the data transformed in a manner that will allow for the analysis to occur.

1.6.4.1. Flattening (De-normalizing) Data

After the data has been entered into the partially normalized format, a Pivot Table will be used to "flatten" (de-normalize) the data, so it is more readable, and so that it will be set up in a manner that can have the automated correlation discovery occur. This flattened format will have a row for each country (based upon its code), and then will have a column for each individual characteristic that will be ultimately part of the automated correlation discovery. Each column will include:

- Source Name
- Characteristic of Nation
- Demographic Subgroup
- Year

The source name is included, because it is possible to have the same type of data from more than one source. The year will be included because education is a long term process, in which a correlation may not be seen for several years after something has occurred.

1.6.4.2. Derived Attributes

There may often be times when the attributes (variables) that are in the initial data source do not work as well for correlation purposes as variables that can be derived from these. For instance, it may be better to search for correlations with educational expenditure per capita than educational expenditures as a portion of GDP. Thus there is value in having derived attributes. While this could occur in the data preparation stage, it is often more efficient to do this after flattening the data, by using the Calculated Items feature of Excel PivotTables.

1.6.5. Correlation Model Building

There are several statistical methods that can be used to determine correlation coefficient between two variables. Which method is best used for a circumstance is dependent upon the type of scales the variables are part of (Kufs, 2011, p. 267). In general, there are 4 major types of scales: nominal scales, ordinal scales, interval scales, and ratio scales (Stevens, 1946). In the case where at least one set of data is not on a ratio or interval scale, attempts to define the correlation by a Cartesian equation through traditional regression techniques do not have much meaning, but statistical methods still exist to derive a correlation coefficient. When both sets of variables are using interval or ratio scales, they can generally be treated as being continuous (even if they are discrete, such as population). In the case of variables that are both either using an interval or ratio scale, than there are one or more equations that the data might correlate to (including non-linear equations), in which case there is a need to test multiple types of equations for correlations.

Based upon the Type of Scale of each variable, the following correlation coefficients will be used:

Scale of 1st Variable	Scale of 2nd Variable	Correlation Coefficient Used
Binary (Nominal)	Binary (Nominal)	Phi and Tetrachoric
Binary (Nominal)	Ordinal	Rank biserial
Binary (Nominal)	Interval	Biserial
Binary (Nominal)	Ratio	Point-biserial
Ordinal	Binary (Nominal)	Rank biserial
Ordinal	Ordinal	Polychoric
Ordinal	Interval	Spearman Rho
Ordinal	Ratio	Spearman Rho
Interval	Binary (Nominal)	Biserial
Interval	Ordinal	Spearman's Rho
Interval	Interval	Pearson Coefficient and equivalents
Interval	Ratio	Pearson Coefficient and equivalents
Ratio	Binary (Nominal)	Point-biserial
Ratio	Ordinal	Spearman Rho
Ratio	Interval	Pearson Coefficient and equivalents
Ratio	Ratio	Pearson Coefficient and equivalents

See Kufs, 2011, p. 267

1.6.5.1. Non-Parametric Models for Nominal and Ordinal Data

Data that has one or more variables that are either nominal (as measured by a binary value of either 0 or 1 for "is or isn't") or ordinal (rankings), these cannot have traditional equations constructed from them that have real meaning. But they can have an appropriate correlation coefficient determined which is generally nonparametric in nature.

When comparing two sets of nominal binary data, the Phi and Tetrachoric correlation coefficients both are commonly used. Which one is most appropriate is dependent upon the underlying distribution of the data sets, where the tetrachoric correlation coefficient is the linear correlation of an underlying bivariate normal distribution, the phi-coefficient is the linear correlation of an underlying bivariate discrete distribution (Ekström, 2011). Given that this underlying distribution is not easily discernible in the research being proposed, the researcher has decided to calculate both coefficients and use the greater of the two to determine whether the results show potential.

For ordinal data, such as when there may only be the ranking of countries released as a GPI, it is common to use the Spearman Rho correlation coefficient or the Kendall Tau. But the Spearman Rho is considered to be more comparable with a Pearson Coefficient (Nagpaul, 1999), and thus will be used. But, it is important to recognize that the Spearman Rho makes the assumption that there is a monotonic relationship between the variables (Lund Research Ltd, 2013), thus if an underlying relationship exists where there is an "optimal" value such as a correlation that could be approximated by a downward parabola or a normal distribution, the Spearman Rho would produce a Type 2 error.

The other non-parametric correlation coefficients likely have some similar issues with only finding monotonic relationships, but given the nature of the scale of the data this probably cannot be easily fixed, although the researcher plans to see if also looking for parametric models would be appropriate.

1.6.5.2. Linear Models

When both sets of data are on an interval or ratio scale, then it is possible to see if a "line of best fit" can be generated using linear regression. A Pearson product-moment correlation coefficient r can be determined as a measure of how well the line fits the data. While it is also common to use the coefficient of determination r^2, this would not be as comparable to the other correlation coefficients that will be used in this research.

1.6.5.3. Non-Linear Models

The Pearson correlation coefficient can only be determined for linear models. But non-linear models that may fit the data such as logarithmic, power, exponential models and even non-monotonic models such as some polynomial models and normal distribution models, can still have a correlation of determination r^2 calculated. By taking the square root of this, an equivalent of a Pearson product-moment correlation coefficient can be calculated.

Given that Excel's primary method of calculating r^2 values for nonlinear data is a process of simple linearization, this does not produce accurate results. An improved method is to use the solver feature to find curves of fit using least squares and then calculate the r^2 from those curves' equations (Kemmer & Keller, 2010; Middleton, 2010). Since solver cannot be done in a simple formula, either a user defined function (UDF) or macro will be developed to calculate the best curves of fit.

1.6.6. Model Evaluation

Each appropriate model will be calculated between the data sets. In the case of interval and ratio scaled data, there will be many common non-linear models tested to find the correlation coefficient of each. When multiple correlation coefficients are calculated, whichever one has the highest absolute value of the correlation coefficients, will be used as the model that is considered the best for that comparison between data sets.

1.6.7. Model Visualization

For correlations that seem particularly strong, especially when they appear to be counterintuitive, these will be deemed to be "interesting" and the results will be graphed in a scatterplot (or other appropriate visualization for non-parametric correlation coefficients) with the curve of best fit overlaid. The lack of finding a correlation between variables may also be explored via data visualization when they also have counterintuitive results.

1.6.8. Model Use

In the select cases that will be explored more in depth, hypothesis testing may be done, if sufficient data is available. This testing may either be done through a t-test or an ANOVA F-test. The scope of this particular research will not generally go beyond this point, as it will be assumed that other researchers will take the data generated by what is found and look at it more deeply and critically. Or the researcher may do some of this himself in his post-doctorate work.

1.7. ISSUES OF RELIABILITY/VALIDITY

The methodology discussed is statistically reliable and valid for what it intends to accomplish, which is to show potential correlations between current data sets. But it is critical to recognize the limitations of the methodology, especially because the primary value of this research is to be able to use its results for future research.

1.7.1. Correlation Does Not Imply Causation

As has been noted earlier in this proposal, correlation does not imply causation. While this research should hopefully uncover potential causations that might be able to have better causation hypothesis testing done in other research, none of its results will be able to actually prove causation.

1.7.2. Potential Problems with "Garbage In, Garbage Out" (GIGO)

As a form of meta-analysis this research relies upon the accuracy of the underlying data sets, and if the source data are not accurate, then clearly the results will not be accurate. In computer science, this is often called GIGO (Garbage In, Garbage Out). Researchers have been calling into question the objectivity and validity of many Global Performance Indicators (Kelley & Simmons, 2014), and this issue needs to be considered as part of the proposed research. And even for GPIs that are considered more rigorous such as international academic testing, there has been research calling into doubt their methodologies and whether it is even possible to compare "apples to apples" between multiple education systems (Sjøberg, 2012; Tienken, 2013).

Thus, there are two school of thoughts about how to choose data for the proposed research. The researcher could try to evaluate the accuracy and validity of the source data before using it (as is traditional), or the researcher could include every piece of data that can be obtained. Originally, the researcher was considering filtering the original data before using it; but instead the second more inclusive method is being planned. The reason is that searching for correlations may in fact help ferret out underlying data that is not reliable, for example a strong correlation between two indicators, one of which is a GPI, could indicate that the GPI could have been created as a composite using that other variable. In which case, one could then consider whether it was an appropriate composite or not. Further, other forms of statistical analysis conducted on the data set produced by this research could potentially do an even better job of specifically looking for suspicious GPIs. Given this potential benefit and that about as much time would be used to attempt to filter data is it would be to include it, the more inclusive method data sourcing will be used.

1.7.3. Potential Issues from National Indices using Composite Data

As has just been discussed, many organizations that create GPIs may have their rankings use some secondary research as part of the calculation in scoring a nation. In these cases, the indices will inherently have a correlation with any data set that was used as part of the scoring. There can be an argument made that this is not a direct issue of validity, but it is clear no new insight was discovered.

1.7.4. Potential Type I and Type II Errors

In any statistical analysis there is always the potential for Type I and Type II errors (which are also sometimes called "false positives" and "false negatives"), and this research is no different, where in some cases there will be a greater correlation found than truly exists (Type I Error) and in other cases the true correlation may be greater than what the research found (Type II Error). While both types of errors may occur, there are more sources of potential Type I errors (false positives) than there are for Type II errors (false negatives).

1.7.4.1. Potential Type I Errors from Random Correlations

Given the large number of potential correlations being calculated, and the relatively low number of data points that each calculated correlation is composed of (250 or less), there is a probability for one of the calculations to discover a spurious correlation. These probabilities will be determined for each comparison of data sets, and included in the results of the research.

1.7.4.2. Potential Type I Errors with Ecological Correlations

Because this research will find “ecological correlations”, because the data sets are all either averages or otherwise aggregated, there will be a tendency that the results will overstate the strength or associations between the data sets. (Freedman, 1999)

1.7.4.3. Potential Type I Errors for Higher Order Polynomial Relationships

If the research was to include higher order polynomial models using linear algebra methods, these might be able to “perfectly” fit the data sets, but would not be valid. As such, the only polynomial model that will be tested with data will be parabolic, as this can often approximate models that have an optimal value.

1.7.4.4. Potential Type II Errors: A Valid Correlation May Exist that Wasn’t Tested

As has been discussed, for data that is not at an interval or ratio scale, generally only monotonic correlations can be determined. And even for interval and ratio scaled data, there is a limited number of models that this research will test for correlation with. There is always the potential that the data could be better described in a manner not considered by the research. This is even the case when a high coefficient is found for a particular model; another model still could be more accurate.

1.8. ETHICAL CONSIDERATIONS

As a form of meta-analysis, most of the traditional ethical research principles are not as applicable, as the research will not be directly getting data from individuals, and all of the data that will be used in the research is publicly available. And while this particular research won’t have these issues, it should be noted that data science conducted on social media data should have considerations about confidentiality and non-disclosure; voluntary informed consent; voluntary participation; the right to withdraw; openness and justice to research participants; and commitment to causing no harm.

But this is to not say that the research doesn’t have ethical considerations. An accepted principle of ethics in the scientific community is that one does not use data that was previously gathered in an unethical manner. Thus, if there is known evidence that an original data set was gathered in a non-ethical manner, it will not be included in this research. For example, there has been some publicly released research conducted by social media companies that have not received consent from the participants (Sullivan, 2014), there has also been research conducted by anonymous hackers that have used malware installed on computers to gather data. Data from research that has these or other forms of ethical issues will not be used.

Transparency is also a critical ethical consideration. The University of South Africa has taken a tremendous step forward by working to ensure all theses can be released publicly, and it is the intention of the researcher, to ensure this thesis will be released under an open libre license. In this same vain, only data sets that can be freely and completely released with the thesis will be included. Given that in in the United States the laws have been interpreted that data sets are not generally covered by copyright law, this should legally allow most data sets obtained by the researcher to be released. But, other countries may have different copyright laws, and the researcher is not in the financial position to take on a lawsuit. Thus any data sets where it seems that the original researcher may attempt to claim a form of intellectual property and not want the data released, then this data will simply not be used, and this lack of scientific transparency on the part of the original researcher will be noted.

1.9. PRELIMINARY CHAPTER OUTLINE

The following are the preliminary chapters that will be in the thesis:

Chapter 1: Introduction and overview
Chapter 2: Literature review of data science & data mining methodologies
Chapter 3: Literature review of known correlations between nations
Chapter 4: Research design and data collection
Chapter 5: Results and discussion of select correlations
Chapter 6: Conclusions, recommendations and limitations of the study

1.10. TIME FRAME

It is estimated that this research should take no more than one year's time. It is important that the research does not extend much beyond this, as it is often preferred to have the correlations between country characteristics being about current data, and so if the research is extended beyond this, then data would often need to be re-gathered, which would be a form of "chasing your tail".

The researcher feels confident that the project can be complete within a year's time, given that much of the data is already in a machine readable form, and that by working with the sister project of The Open Compendium of Countries for data sourcing, data can be gathered by other individuals as well the researcher. The time for the development of the analysis tools is estimated to be about 6 months of part-time work, and the actual processing time required for the model building and evaluation is unknown, but if it appears to be too great, cloud computing or distributed computing could be employed to expedite the process.

The following are the different components of the research, although these do not necessarily act as distinct phases, since much of the research can be done in an iterative and parallel manner:

1. **Data Sourcing, Acquisition, and Preparation**: This work will be conducted with The Open Compendium of Countries. It will be ongoing, with a system employed that as data is acquired and prepared from original data sources, these can be fairly easily and quickly incorporated into the research.
2. **Develop the Tools for Data Transformation, Model Building, and Evaluation**: It is estimated that the full process of developing these tools will take about 6-months of part-time work, although it may take less time dependent upon the other events occurring in the researcher's life. Given that this development work can be done in tandem with the sourcing of data, delays in this portion of the research should not delay the ultimate results.
3. **Conduct Data Mining, using Developed Tools:** As the data sets get larger, the processing required for this portion of the research will grow in a parabolic manner (not exponentially), if this produces significant time delays, then strategies such as using cloud or distributed computing may be utilized, or if necessary Excel could be abandoned as the primary tool, and another tools such as Weka, or a custom programmed R or Python program, could be used.
4. **Explore "Interesting" Results:** As discussed, data visualization and potentially hypothesis testing will be conducted on results that seem interesting. This can often be done as these results are found, and thus done in tandem with other steps.
5. **Write the Thesis:** Portions of the thesis such as the literature review can be written in tandem with the other steps, but the full thesis cannot be complete until the other steps are done.

1.11. REFERENCES

Note: This proposal and the resulting thesis uses/will use the American Psychological Association's 6th Edition citation style. This style has been chosen as it internationally recognized as being appropriate for research involving the social sciences.

Barth, P., Earley, S., Lawson, L., & Hall, S. (2013). *Turning Big Data Into Useful Information*. eWeek. Retrieved from http://www.eweek.com/research/editorial/big-data-analytics?assettype=pdf

Baykal, A. (2014). Human Development Indicators and PISA Scores Correlated. In *IICE 2014 Proceedings*. Dublin, Ireland. Retrieved from https://www.academia.edu/7031378/Human_Development_Indicators_and_PISA_Scores_Correlated

Bozkır, A. S., Mazman, S. G., & Sezer, E. A. (2010). Identification of user patterns in social networks by data mining techniques: Facebook case. In *Technological Convergence and Social Networks in Information Management* (pp. 145–153). Springer. Retrieved from http://link.springer.com/chapter/10.1007/978-3-642-16032-5_13

Bray, M., Adamson, B., & Mason, M. (2007). *Comparative Education Research: Approaches and Methods*. シュプリンガー・ジャパン株式会社. Retrieved from http://public.eblib.com/EBLPublic/PublicView.do?ptiID=371572

Central Intelligence Agency. (2014). *The World Factbook parsed into XML*. Washington, DC: Sourceforge. Retrieved from http://jmatchparser.sourceforge.net/factbook/

Christensen, G., Steinmetz, A., Alcorn, B., Bennett, A., Woods, D., & Emanuel, E. J. (2013). The MOOC phenomenon: who takes massive open online courses and why? *University of Pennsylvania, Nd Web*, *6*. Retrieved from http://m4ed4dev.linhost1.jbsinternational.com/sites/default/files/the_mooc_phenomenon.pdf

Cleveland, W. S. (2001). Data science: an action plan for expanding the technical areas of the field of statistics. *International Statistical Review*, *69*(1), 21–26.

Ekström, J. (2011). The Phi-coefficient, the Tetrachoric Correlation Coefficient, and the Pearson-Yule Debate. *Department of Statistics, UCLA*. Retrieved from https://escholarship.org/uc/item/7qp4604r.pdf

Fayyad, U., Piatetsky-Shapiro, G., & Smyth, P. (1996). From data mining to knowledge discovery in databases. *AI Magazine*, *17*(3), 37.

Feynman, R. (1964). *Messsenger Lectures, The Character of Physical Law -7 -Seeking New Laws*. Retrieved from https://www.youtube.com/watch?v=j3mhkYbznBk&feature=youtube_gdata_player

Frawley, W. J., Piatetsky-Shapiro, G., & Matheus, C. J. (1992). Knowledge discovery in databases: An overview. *AI Magazine*, *13*(3), 57.

Freedman, D. A. (1999). Ecological inference and the ecological fallacy. *International Encyclopedia of the Social & Behavioral Sciences*, *6*, 4027–4030.

Hanushek, E. A., & Woessmann, L. (2010). *The High Cost of Low Educational Performance: The Long-Run Economic Impact of Improving PISA Outcomes*. Paris: Organization for Economic Cooperation and Development. Retrieved from http://www.oecd.org/pisa/44417824.pdf

How to lie with indices. (2014, November 8). *The Economist*. Retrieved from http://www.economist.com/news/leaders/21631025-learn-ruses-international-country-rankings-how-lie-indices

Kelley, J. G., & Simmons, B. A. (2014). The Power of Performance Indicators: Rankings, Ratings and Reactivity in International Relations. Presented at the Annual Meeting of the American Political Science Association, Rochester, NY: Social Science Research Network. Retrieved from http://papers.ssrn.com/abstract=2451319

Kemmer, G., & Keller, S. (2010). Nonlinear least-squares data fitting in Excel spreadsheets. *Nature Protocols*, *5*(2), 267–281. doi:10.1038/nprot.2009.182

Kufs, C. (2011). *Stats with Cats: The Domesticated Guide to Statistics, Models, Graphs, and Other Breeds of Data Analysis*. Wheatmark, Inc.

Lee, D., Mani, M., & Chu, W. W. (2003). Schema Conversion Methods between XML and Relational Models. In *Knowledge Transformation for the Semantic Web*. IOS Press. Retrieved from http://www.cobase.cs.ucla.edu/tech-docs/dongwon/book03.pdf

Lund Research Ltd. (2013). Spearman's Rank-Order Correlation - A guide to when to use it, what it does and what the assumptions are. Retrieved November 4, 2014, from https://statistics.laerd.com/statistical-guides/spearmans-rank-order-correlation-statistical-guide.php

Mannila, H. (2000). Theoretical frameworks for data mining. *ACM SIGKDD Explorations Newsletter*, *1*(2), 30–32.

Middleton, M. R. (2010). Better Exponential Curve Fitting Using Excel. Presented at the 41st Decision Sciences Institute Annual Meeting 2010, San Diego, CA: Decision Sciences Institute. Retrieved from http://www.mikemiddleton.com/Excel-Exponential-Curve-Fit-2010.pdf

Miner, G., Nisbet, R., & Elder, J. (2009). *Handbook of Statistical Analysis and Data Mining Applications*. Academic Press. Retrieved from http://books.google.com/books?id=U5np34a5fmQC

Nagpaul, P. S. (1999). Non-parametric Measures of Bivariate Relationships. In *Guide to Advanced Data Analysis using IDAMS Software*. New Delhi (India): Division of Information and Informatics UNESCO. Retrieved from http://www.unesco.org/webworld/idams/advguide/Chapt4_2.htm

Noah, H. J. (1985). Comparative Education. In T. Husén & T. N. Postlethwaite (Eds.), *The International encyclopedia of education: research and studies* (pp. 869–872). New York, NY: Pergamon Press.

Noé, K. (1998). Philosophical aspects of scientific discovery: A historical survey. In *Discovey Science* (pp. 1–11). Springer. Retrieved from http://link.springer.com/chapter/10.1007/3-540-49292-5_1

OECD. (2011). *Does investing in after-school classes pay off?* (No. N°3). Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd.org/portugal/47573005.pdf

OECD. (2012). *Does money buy strong performance in PISA?* (No. N°13). Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisainfocus/49685503.pdf

OECD. (2014a). *Do students have the drive to succeed?* (No. N°37). Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisainfocus/pisa-in-focus-n37-%28eng%29-final.pdf

OECD. (2014b). *When is competition between schools beneficial?* (No. N°42). Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisainfocus/PISA-in-Focus-N42-%28eng%29-FINAL.pdf

OECD. (2014c). *Who are the school truants?* (No. N°35). Organisation for Economic Co-operation and Development. Retrieved from http://www.oecd.org/pisa/pisaproducts/pisainfocus/PISA-in-Focus-n35-%28eng%29-FINAL.pdf

Piety, P. J., Hickey, D. T., & Bishop, M. J. (2014). Educational data sciences: framing emergent practices for analytics of learning, organizations, and systems. In *Proceedings of the Fourth International Conference on Learning Analytics And Knowledge* (pp. 193–202). Indianapolis, Indiana: ACM Press. doi:10.1145/2567574.2567582

Sjøberg, S. (2012). PISA: Politics, fundamental problems and intriguing results. *La Revue*, *14*. Retrieved from http://www.uhr.no/documents/6b_Sjoberg_PISA_English_La_Revue_no_20.03..pdf

Stevens, S. S. (1946). On the theory of scales of measurement. *Science*, *103*(2684), 677–680.

Sullivan, G. (2014, July 1). Cornell ethics board did not pre-approve Facebook mood manipulation study. *The Washington Post*. Retrieved from http://www.washingtonpost.com/news/morning-mix/wp/2014/07/01/facebooks-emotional-manipulation-study-was-even-worse-than-you-thought/

Tienken, C. H. (2013). Conclusions from PISA and TIMSS Testing. *Kappa Delta Pi Record*, *49*(2), 56–58. doi:10.1080/00228958.2013.786588

Tienken, C. H., & Mullen, C. A. (2014). The Curious Case of International Student Assessment: Rankings and Realities in the Innovation Economy. In *Building Cultural Community Through Global Educational Leadership* (pp. 146–164). Retrieved from http://christienken.com/wp-content/uploads/2013/01/Tienken_Mullen_PageProofs_MAy21_CLEAN-1.pdf

UNESCO Institute for Statistics. (2014, June 30). UNdata | explorer | UIS Data Centre. Retrieved October 12, 2014, from http://data.un.org/Explorer.aspx?d=UNESCO