2009

EXPLORING IMPACT OF EDUCATIONAL AND ECONOMIC FACTORS ON NATIONAL INTELLECTUAL PRODUCTIVITY USING MACHINE LEARNING METHODS

Canon Fazenbaker
Virginia Commonwealth University

School of Engineering
Virginia Commonwealth University

This is to certify that the thesis prepared by Canon Edward Fazenbaker entitled EXPLORING IMPACT OF EDUCATIONAL AND ECONOMIC FACTORS ON NATIONAL INTELLECTUAL PRODUCTIVITY USING MACHINE LEARNING METHODS has been approved by his or her committee as satisfactory completion of the thesis or dissertation requirement for the degree of Master of Science.

Kayvan Najarian, Ph.D., Committee Chair, Department of Computer Science, School of Engineering

---

Krzysztof Cios, Ph.D., Chair of Department of Computer Science, School of Engineering

---

Peter Aiken, Ph.D, Department of Information Systems, School of Business

---

Dr. F. Douglas Boudinot, Dean of the School of Graduate Studies

[Click here and type the Month, Day and Year this page was signed.]

EXPLORING IMPACT OF EDUCATIONAL AND ECONOMIC FACTORS ON NATIONAL INTELLECTUAL PRODUCTIVITY USING MACHINE LEARNING METHODS

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

by

CANON EDWARD FAZENBAKER
B.S., West Virginia University Institute of Technology, 2005

Director: KAYVAN NAJARIAN
ASSOCIATE PROFESSOR, DEPARTMENT OF COMPUTER SCIENCE

Virginia Commonwealth University
Richmond, Virginia
December, 2009

Acknowledgements

I would like to thank Dr. Soo Yeon Ji and Wenan Chen for their expertise and support with this study. Their research experience and knowledge on machine learning was an immense help and much appreciated.

I would also like to thank Cole Fazenbaker, my brother, for his contributions regarding this topic; his economic background was an important part of this study. Also, I would like to thank my wife, Sara, for her grammatical input into this paper and her patience and support during this study.

Finally, I would like to thank Dr. Kayvan Najarian, Dr. Krzysztof Cios, and Dr. Peter Aiken for their input and knowledge on how to improve this thesis.

Table of Contents

List of Tables

List of Figures

Abstract

By Canon Edward Fazenbaker

A thesis submitted in partial fulfillment of the requirements for the degree of Master of Science at Virginia Commonwealth University.

Virginia Commonwealth University, 2009

Major Director: Kayvan Najarian
Associate Professor, Department of Computer Science

The patent process is representative of a nationwide means for innovations and new ideas to be recognized. The U.S. Patents Office, since its inception in 1790, has issued nearly five million patents. These patents span from the U.S. Patent #1, which was for an improvement "in the making of Pot ash and Pearl ash by a new Apparatus and Process" to today's patents which deal with technologies and mediums that were unimaginable at the Patent Offices' inception. The purpose of this study is to determine what social and economic factors at the federal level have the highest impact on national productivity measured by the number of patents applied for and/or granted each year. Using Machine Learning algorithms and predictive analysis on fifty years worth of data to determine what macroeconomic and educational factors have the most impact on patents.

The first part of this study describes the methods and algorithms used during this research. The second part of this study discusses the results and what those results reveal about the impact of education and economic factors as they relate to national creativity / intellectual productivity. The goal of this study is to determine what factors affect national intellectual productivity in a given year. This data will be useful for governments, both local and federal, when faced with educational and economic issues.

Chapter 1: Introduction

1.1 Overview

A nation's intellectual productivity serves as a contributing factor when considering overall prosperity on a national level. Economic and educational policies set in place by the federal government designed to have an impact on a particular area inevitably influence sometimes unforeseen aspects of other sectors. With a better understanding of what those unforeseen aspects are, a more resourceful federal government will emerge.

1.2 Problem Statement

Macroeconomic fiscal and monetary policies are two types of strategies that the federal government adjusts in order to maintain a stable and prosperous economy. When the Federal Reserve Bank adjusts the Federal Funds rate, which has a direct affect on short term interest rates such as the prime rate, it is clear that the primary concern is with economic growth and inflation [16].

This study gives quantitative evidence that the government needs to closely observe specific factors in macroeconomic planning. In particular, this study lends evidence that one of the government's concerns should be with the influence that their strategic decisions have on national intellectual productivity. This study's intent is not to determine the most predictive method for forecasting national intellectual productivity, but rather lend

evidence to the fact that economic and educational factors both play a part in the nation's overall productivity. Machine learning analysis is used in this study to show these relationships, but recent research by Ben-David and Frank [35] also shows the importance and relevance of "hand crafted" expert systems developed by subject matter experts that have a more detailed understanding of the data itself and the relationships between the individual attributes.

Intellectual productivity is known to be one of the major factors in creating technologies that form industries producing capital; and therefore becoming major sources of prosperity. Wireless and optical communication, biotechnology, and nanotechnology are examples of such intellectual endowers resulting in major industries that shape the U.S. and international economy. These "waves" of technological innovations are important factors to predict, plan, and analyze in order to ensure economic prosperity.

Knowing the role of education on intellectual productivity, an important factor to consider is the government's educational plans. The federal government's role in education is not simple to define. The Department of Education has a mission to promote student achievement and preparation for global competitiveness by fostering educational excellence and ensuring equal access [11], by establishing policies on federal financial aid for education, and distributing as well as monitoring those funds.

Taking into account the framework introduced by Furman [13], national innovative capacity is understood as an economy's potential for producing a stream of commercially relevant innovations. In order for an individual or company to capitalize financially on

those innovations a patent is required. Therefore, while examining national productivity the three main elements of national innovative capacity [14] will be observed as well.

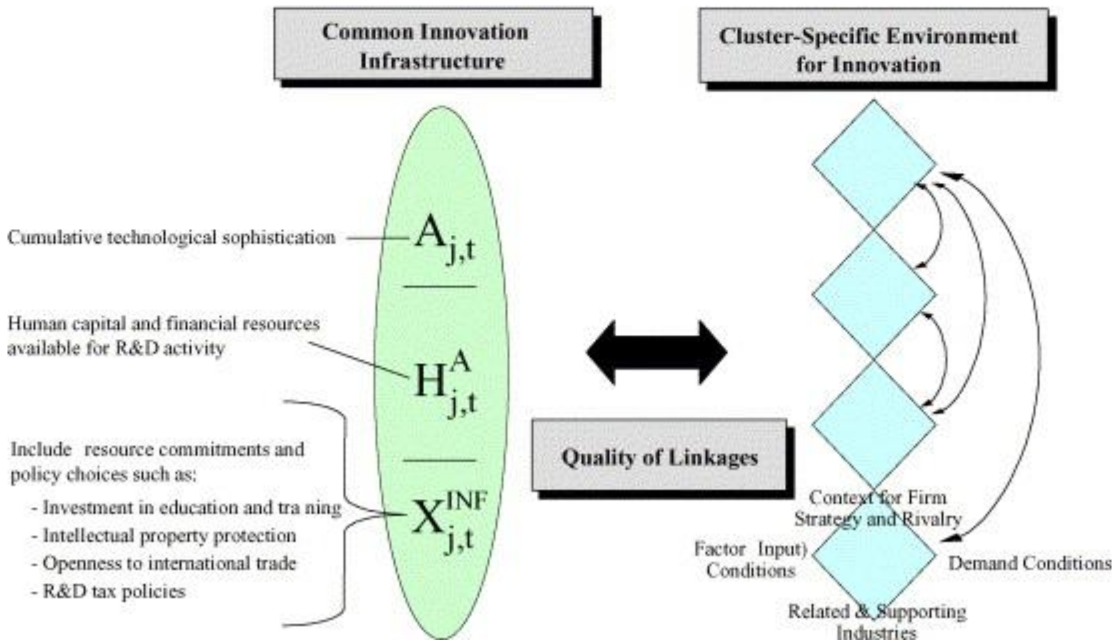

Figure 1.1: National Innovative Capacity (Courtesy of Furman and Hayes [14])

1. The Common Innovation Infrastructure (i.e. Cumulative technological sophistication, Human capital and financial resources available for R&D activity, and resource commitments and policy choices).
2. The Cluster-Specific Environment for Innovation (i.e. the related and supporting industries and the demand conditions.
3. The quality of linkages between the infrastructure and the environment.

Machine learning methods have been used in many different industries to analyze a wide array of issues [37] [40]. While there have been attempts to intuitively predict classes of industries that are more likely to impact the future economy, there has been little work done on quantitative analysis of factors that most identify and impact the innovative capacity / potential of the nation. This study will apply advanced machine learning methods to analyze different attributes as potential factors impacting intellectual productivity and identify the most significant attributes among this list as described in Specific Aims.

1.3 Specific Aim

The main objective of this project is to quantitatively analyze various macroeconomic measures and identify the ones that can most effectively help maintain a stable level of intellectual productivity, which in turn facilitates a more stable and prosperous economy. Specifically, this study uses public education enrollment statistics [21], as well as private school enrollment to determine if there is a significant relation between private and public school enrollment and national intellectual productivity.

Starting with a data set of both economic and educational data (see Appendix A for full list of dataset) this study determines the most predictive attributes that relate to national intellectual productivity. Macroeconomic and educational data were chosen because of the federal government's impact on policies and funding. Whether that impact is direct, such is the case with the interest rate, or the impact is indirect as with the

unemployment rate, it is clear that the actions of the federal government have an effect on those attributes.

Machine learning methods including M5 Rules [2] [3] [4], Decision Table [5], and Conjunctive Rule will be used for analyzing the data. The use of the rule-based system will allow human users to understand the reasoning behind the extracted knowledge.

1.3.1 Patent Issuance as a Measure of National Productivity

Patent issuance measures one particular type of output of national productivity – intellectual productivity. A patent grants the right to exclude others from making, using, offering for sale, or selling the invention throughout the United States or importing the invention into the United States [15].

Even though patent issuance is not the only measure of intellectual productivity, due to the legal structure that protects the rights for the intellectual property, it is logical that patent issuance would be the most significant measure to assess intellectual productivity. It is understandable that not all patents are pursued as a commercial product and not all commercial products formed out of a patent are truly innovative; however, assuming that the ratio of the patents that contribute to intellectual productively remain relatively constant, the total number of patents issued is a reliable measure to assess national intellectual productivity.

1.3.2 National Intellectual Productivity

This study considers National Innovative Capacity [14], as an independent entity and uses its overall schema as a black-box type of concept for the National Intellectual Productivity model. As described in the model (see Figure 4), the federal government produces the environment, or inputs to the model, then based on these inputs and the National Innovative Capacity black-box National Intellectual Productivity is captured. This model is laid out in Figure 1.2.

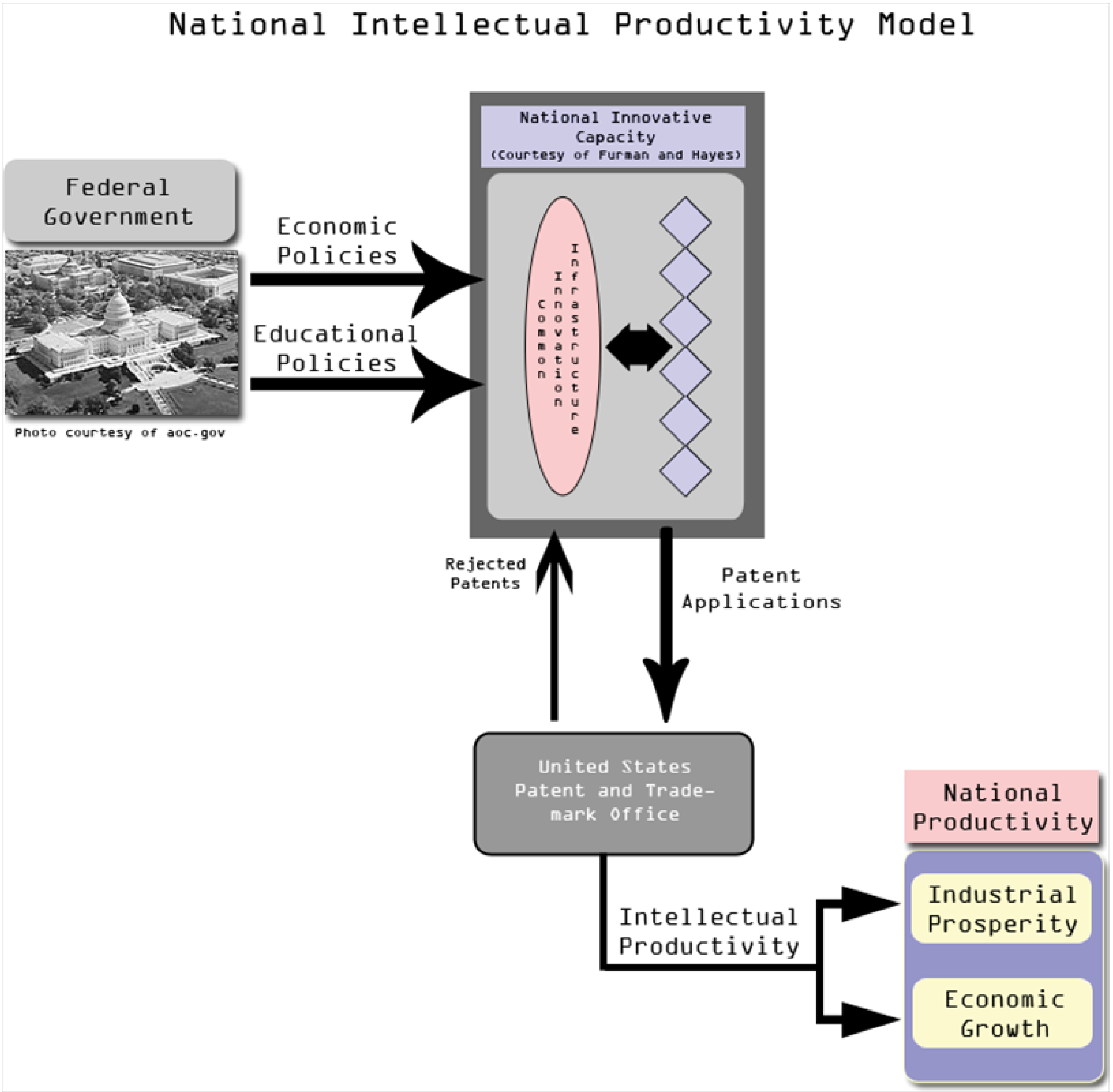

Figure 1.2: National Intellectual Productivity Model

1.3.3 Patent Considerations

When measuring national intellectual productivity using patents, it is vital to not look at the statistics in a vacuum. There are other factors that play a part in the number of patents applied for each year. Among these factors are the fees associated with filing a patent. Previous studies, although more focused on the European Patent Office (EPO), have shown that a 10% increase in filing fees would lead to a reduction of about 5% in the filing of patents [17]. Although, patent fees are unlikely to have an effect on patents claiming technological breakthroughs, it is safe to assume that inventions with less potential financial gain would be affected by this variable.

Another variable to keep in mind is that the patent data excludes reissues. If a patent has been reissued, which sometimes broadens the scope to include previously neglected aspects of the invention, then that patent is only represented once in the statistics, and in turn also not represented in this study’s data and corresponding research.

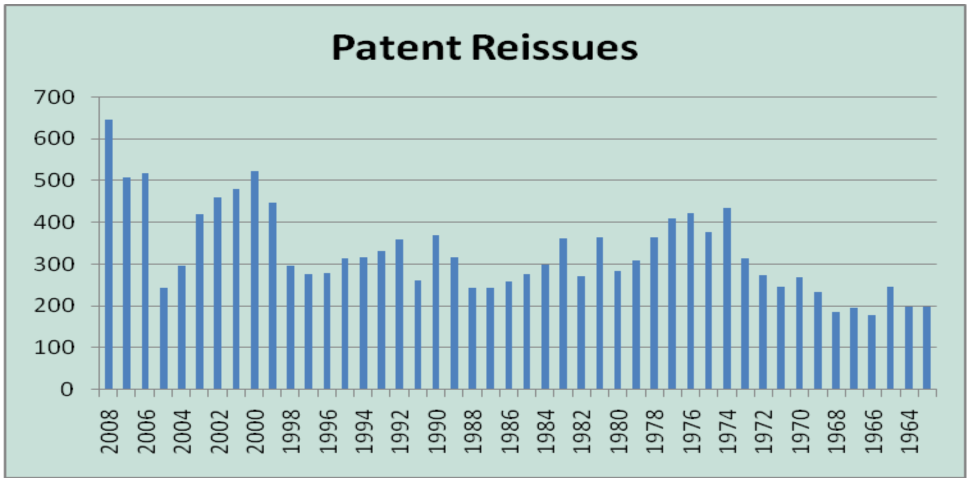

Figure 1.3: Patent Reissues

For completeness purposes and using the U.S Patent and Trademark Office data [18], this study has calculated the mean number of patent reissues from 1963 to 2008 as 314.88 with a standard deviation of 102.91. These data are shown in Figure 1.3.

1.4 Summary

In Chapter 1, a brief introduction to the ideas of the project is given. First, the problem statement and specific aims are provided; the main objective of the study is to apply machine learning methods to identify factors that impact national intellectual productivity. Knowing that intellectual productivity is a major factor in ensuring a stable and prosperous economy, it is important to find the factors that help maintain a high level of intellectual productivity. In this study, it is hypothesized that educational policies and plans are among the most important factors that affect intellectual productivity; this

hypothesis is tested using the historical data representing the number of patents applied for and issued in the United States.

Chapter 2: Methods

2.1 Overview

Three classification algorithms; M5 Rules [2] [3] [4], Decision Table [5], and Conjunctive Rule , are applied to classify the created data sets. By restricting specific attributes from the data set, and then comparing the results of each run, the most relevant data becomes evident. The extraneous data that is removed from the data set allows for more accurate numeric projections [26]. These algorithms were implemented using the WEKA toolkit [1].

In addition, the ReliefFAttributeEval [7] [8] [9] algorithm for selection of most relevant attributes, which is implemented in WEKA, was used to investigate the most predictive attributes of the data set. In order to determine if a combination of economic and educational data would produce a more highly accurate forecast of national intellectual productivity some pre-processing, in the form of attribute selection was done using the entire data set as a whole.

2.1.1 Test Options

Each classifier was run using three different test options: 10 fold cross-validation, 49 fold cross-validation (49 fold was used because it is the maximum allowed by the dataset), and 66% percentage split. Cross-validation [6], defines and generates a number of

folds, n, that randomly reorders and splits the data set into equally sized folds. In each test, a single fold among the *n* folds is used for testing while the remaining *n*-1folds are used for training the classifier. The results are then collected and averaged over all tests. Percentage split uses a certain percentage, *m*, of the data to use for training, and the remaining data, 100 – *m*, to perform testing.

2.2 Classifiers

The classifiers used in this study, i.e. M5, Conjunctive Rule, and Decision Table, are models for prediction and classification. This study uses each of these classifiers to predict the number of patents applied for and granted using various data sets, and compare the results.

Next the three classifiers are very briefly introduced.

2.2.1 M5Rules

M5 Generates a decision list for regression problems using separate-and-conquer [29]. Each iteration of the algorithm builds a model tree using M5 and makes the "best" leaf into a rule. The M5Rules algorithm was chosen as one of the methods used for prediction based on the results of previous studies using model trees for classification [19] which concludes that versions of the M5 algorithm outperformed a state-of-the-art decision tree learner on problems with numeric attributes. As such, this algorithm was used in this paper as one of the three algorithms to be compared with other algorithms known to work well with numeric attributes.

2.2.2 ConjunctiveRules

Conjunctive rule is a two-stage algorithm [31] that first produces a set of classification rules and then prunes and orders those rules during the execution using Reduced Error Pruning [32]. Conjunctive rule implements a single conjunctive rule learner that can predict numeric values [6]. The rules created by Conjuctive rule, as other rule learners in general, can sometimes create complicated and long rules. Although research exists as to the validity and usability of the more complicated rules [30], this study is only interested in the overall predictive performance of these rules.

2.2.3 Decision Table

Decision table builds and executes a simple Decision Table Majority (DTM) [5] with two components consisting of a schema and a body. Decision table [5], in some instances, outperforms state-of-the-art classifiers such as C4.5. DTM uses the wrapper model [33] [34] to identify optimal attributes during the execution of the classifier. Best-first search, the wrapper model algorithm used in this study, works in conjunction with the classifier to identify the optimal features of the data set.

2.3 Attribute Selection

Attribute selection [10] is used to further refine the data that provides the most predictive qualities and reduces the number of dimensions describing data [38]. Attribute selection, sometimes referred to as feature selection, is the process in which a subset of a

given data set is selected based on its connection to the desired input variable. Feature selection is an essential step when the goal is to produce high accuracy classifications [39]. The attribute selection machine learning method, ReliefFAttributeEval, is used in this study to investigate specific attributes to determine which are the most predictive.

This method is further described next.

2.3.1 ReliefFAttributeEval

The RELIEF approach [24] [28] describes two fundamental approaches to attribute selection as:

(1) A filter that works independently of the classifier and

(2) A wrapper approach that selects attributes to optimize classification using the algorithm.

For the M5 and Conjuctive rule executions this study applies the former - an independent filter approach which selects the optimal set of attributes independently of the classifier algorithms used. Recent research aimed at optimizing ReliefF [36], referred to as Supervised Model Construction (FSSMC), is designed to reduce processing time while maintaining accuracy. The data set used in this study does not call for the use of this new implementation since processing time in our instance is a matter of seconds.

2.4 Data Set Formation

When using machine learning methods to performing statistical analysis such as regression, it is preferable to create the data set in such a way that takes advantage of the attribute with the highest frequency of measurement. The patent data provided (see Figure 2.1) by the United States Patent and Trademark Office [23] being yearly, lead to the data with a more frequent measurements such as the mortgage rate and savings rate to be normalized by taking the yearly maximum, minimum, median, and mean values.

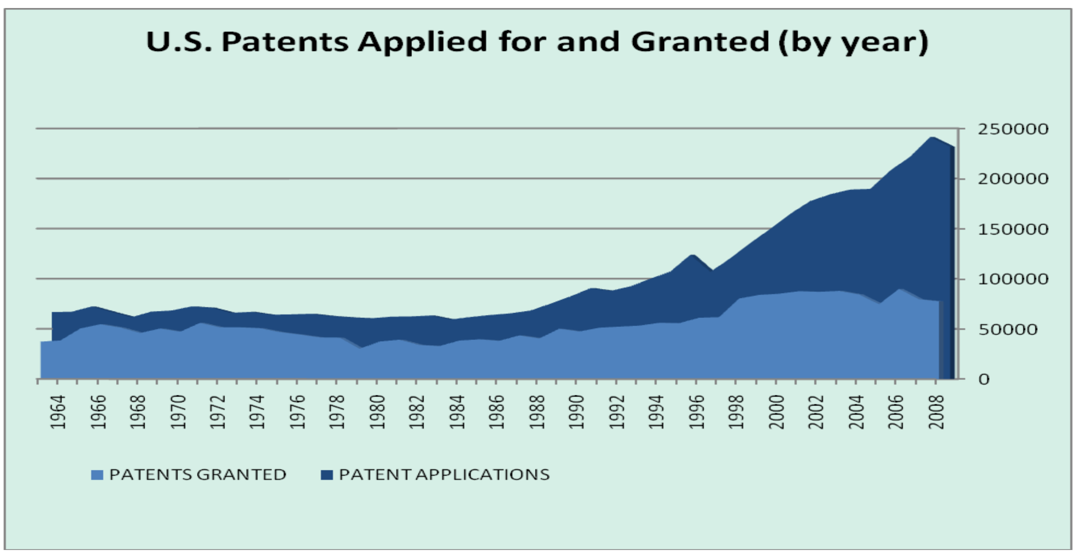

Figure 2.1: U.S. Patents Applied for and Granted (see Appendix D for relating data)

2.4.1 Economic Data

The attributes that make up the economic data set are by and large made up of macroeconomic factors. Unemployment rate, mortgage rate, savings rate, and gross domestic product (GDP) represent this study's macroeconomic attributes.

2.4.2 Educational Data

Educational data was obtained from the National Center for Educational statistics which is a part of the U.S. Department of Education [22]. The attributes that make up the educational data set represent a broad range of enrollment statistics. Enrollment statistics, both private and public, are broken out by elementary, secondary schools preschool through eighth grade, grades nine through twelve, and post secondary degrees.

Chapter 3 – Results and Discussion

3.1 Overview

This section is dedicated to presentation of the results as well as the discussion of the obtained results. Three different methods are used for the analyses and their results are compared with each other.

3.2 Analysis Conditions

As discussed in the previous chapters, three classifier algorithms; M5 Rules, Conjunctive Rule, and Decision Table, are used in this study to enumerate national intellectual productivity. Each run of the classifier is used to compute the relative absolute error of the projected patent attribute to the actual patent data (Table 3.1). Each classifier was run using three different test options: 10 fold cross-validation, 49 fold cross-validation, and 66% percentage split.

Each classification algorithm applied this study's standard economic or educational data set (see Appendix B and C for more details on these datasets). The U.S. population [20] was then added into each data set and the classification tasks were run again. This was done to quantify the effect that the raw population has on national intellectual productivity.

3.3 Evaluation of Economic Data

The attributes that make up the economic data set are primarily made up of macroeconomic factors. Unemployment rate, mortgage rate, savings rate, and gross domestic product (GDP) represent this study's macroeconomic attributes. Using the relative absolute error as the indication of predictive capability, Table 3.1 details the performance of the economic data set when used to project the number of patent applications filed for a given year.

Table 3.1: Economic Results – Patent Applications
(see Appendix B for input data attributes)

Classification Algorithm and Test Options	Relative Absolute Error (Data set without U.S. Population)	Relative Absolute Error (Data set with U.S. Population included)
M5Rules (cross validation 49 folds)	47.86%	47.86%
M5Rules (cross validation 10 folds)	*56.68%*	*56.68%*
M5Rules (percentage split 66%)	33.10%	33.10%
ConjunctiveRule (cross validation 49 folds)	36.99%	36.99%
ConjunctiveRule (cross validation 10 folds)	42.94%	42.94%
ConjunctiveRule (percentage split 66%)	29.04%	29.04%
DecisionTable (cross validation 49 folds)	18.39%	18.81%
DecisionTable (cross validation 10 folds)	**18.29%**	**18.56%**
DecisionTable (percentage split 66%)	18.90%	21.37%

The results of Table 3.1 show that the Decision Table classifier is the most predictive when computing national productivity measured by patent applications. Each Decision Table run outperformed all of the other executions of Conjunctive Rule and M5 Rules. A difference of 38.39% is evident between the least predictive run of M5 and the

most predictive run of the decision table, which witnesses to the superiority of the performance of the decision table in this modeling task.

Table 3.2 shows the performance of economic data set when used to predict the number of granted patents for a given year.

Table 3.2: Economic Results – Granted Patent
(see Appendix B for input data attributes)

Classification Algorithm and Test Options	Relative Absolute Error (Data set without U.S. Population)	Relative Absolute Error (Data set with U.S. Population included)
M5Rules (cross validation 49 folds)	47.13%	47.13%
M5Rules (cross validation 10 folds)	*55.01%*	*55.01%*
M5Rules (percentage split 66%)	46.07%	46.07%
ConjunctiveRule (cross validation 49 folds)	48.93%	48.93%
ConjunctiveRule (cross validation 10 folds)	43.40%	43.40%
ConjunctiveRule (percentage split 66%)	43.45%	43.45%
DecisionTable (cross validation 49 folds)	29.14%	32.11%
DecisionTable (cross validation 10 folds)	28.02%	**28.02%**
DecisionTable (percentage split 66%)	**22.64%**	31.98%

As shown by Table 3.2, again Decision Table yields the most predictive results when used to project the number of patents granted. Table 3.2 also shows that the economic data set is more accurate (by 4.35%) when projecting the number of patents applied for than the number granted.

3.4 Discussion of Results: Economic Data

Table 3.3 compares the results achieved by the classifiers for the economic data set. As indicated before, Decision Table is the most predictive resource for projecting national intellectual productivity (for both patent applications and granted patents), while the M5Rules algorithm is the least predictive when using economic data to predict national intellectual productivity.

Table 3.3: Economic Results Summary
(see Appendix B for input data attributes)

Results Summary	Patent Applications Data w/o Population	Patent Applications Data with Population	Patent Granted Data w/o Population	Patent Granted Data With Population
Mean Relative Absolute Error	30.98%	31.54%	38.89%	40.86%
RAE Standard Deviation	*13.86%*	*13.46%*	*11.05%*	*9.06%*
Least Predictive Value	*56.68%*	*56.68%*	*55.01%*	*55.01%*
Least Predictive Algorithm	M5Rules (cross validation 10 folds)	M5Rules (cross validation 10 folds)	M5Rules (cross validation 10 folds)	M5Rules (cross validation 10 folds)
Most Predictive Value	**18.29%**	**18.56%**	**22.64%**	**28.02%**
Most Predictive Algorithm	DecisionTable (cross validation 10 folds)	DecisionTable (cross validation 10 folds)	DecisionTable (percentage split 66%)	DecisionTable (cross validation 10 folds)

As it can be seen in Table 3.3:

1. Given the set of economic inputs introduced in this study, Decision Table is capable of predicting the number of patents with relatively high accuracy.

2. Addition of population as an input does not help with the accuracy of the prediction, showing that the information in population cannot be very informative once the other input factors are processed by the Decision Table.

3.5 Evaluation of Educational Data

The attributes that make up the educational data set represent a broad range of enrollment statistics. Enrollment statistics, both private and public, are broken out by elementary, secondary schools preschool through eighth grade, grades nine through twelve, and post secondary degrees.

As with the economic data set, the relative absolute error is used as the indication of predictive capability. Table 3.4 details the performance of the educational data set when used to calculate the number of patent applications filed for a given year.

Table 3.4: Educational Results – Patent Applications
(see Appendix H for DecisionTable actual results and Appendix C for input data attributes)

Classification Algorithm and Test Options	Relative Absolute Error (Data set without U.S. Population)	Relative Absolute Error (Data set with U.S. Population included)
M5Rules (cross validation 49 folds)	45.97%	45.97%
M5Rules (cross validation 10 folds)	*52.36%*	*52.36%*
M5Rules (percentage split 66%)	41.06%	41.06%
ConjunctiveRule (cross validation 49 folds)	36.99%	36.99%
ConjunctiveRule (cross validation 10 folds)	50.70%	50.70%
ConjunctiveRule (percentage split 66%)	29.04%	29.04%
DecisionTable (cross validation 49 folds)	30.10%	18.69%
DecisionTable (cross validation 10 folds)	29.49%	**16.34%**
DecisionTable (percentage split 66%)	**11.91%**	21.37%

Table 3.4 reveals that the Decision Table classifier is the most predictive when computing national productivity measured by patent applications. Although not all Decision Table runs outperform other classifiers (Conjunctive Rule with a percent split of 66% was more accurate than Decision Table with 49 folds cross validation), the relative absolute error achieved with the 66% split run resulted in a 6.38% improvement over the most predictive economic data set execution. A difference of 40.45% was shown between the least predictive run of M5 and the most predictive run of the Decision Table.

Table 3.5 shows the performance of the educational data set when used to predict the number of granted patents for a given year.

Table 3.5: Educational Results – Granted Patent
(see Appendix C for input data attributes)

Classification Algorithm and Test Options	Relative Absolute Error (Data set without U.S. Population)	Relative Absolute Error (Data set with U.S. Population included)
M5Rules (cross validation 49 folds)	51.88%	51.88%
M5Rules (cross validation 10 folds)	*59.87%*	*59.87%*
M5Rules (percentage split 66%)	45.52%	45.52%
ConjunctiveRule (cross validation 49 folds)	48.93%	48.93%
ConjunctiveRule (cross validation 10 folds)	48.12%	48.12%
ConjunctiveRule (percentage split 66%)	43.45%	43.45%
DecisionTable (cross validation 49 folds)	30.57%	32.30%
DecisionTable (cross validation 10 folds)	31.39%	35.53%
DecisionTable (percentage split 66%)	**27.33%**	**31.98%**

As shown by Table 3.5, Decision Table also yields the most predictive results when used to predict the number of patents granted (this was also the case with the economic

data set). The educational data set is more accurate, by 15.42%, when predicting the number of patents applied for than the number granted.

3.6 Discussion of Results: Educational Data

Table 3.6 further compares the educational data executions. The Decision Table classifier is the most predictive for both patent applications and granted patents. While the M5Rules algorithm is the least predictive when using economic data to predict national productivity.

Table 3.6: Educational Results Summary
(see Appendix C for input data attributes)

Results Summary	Patent Applications Data w/o Population	Patent Applications Data with Population	Patent Granted Data w/o Population	Patent Granted Data With Population
Mean Relative Absolute Error	33.76%	32.00%	41.68%	43.26%
RAE Standard Deviation	*12.82%*	*13.89%*	*10.99%*	*9.42%*
Least Predictive Value	*52.36%*	*52.36%*	*59.87%*	*59.87%*
Least Predictive Algorithm	M5Rules (cross validation 10 folds)	M5Rules (cross validation 10 folds)	M5Rules (cross validation 10 folds)	M5Rules (cross validation 10 folds)
Most Predictive Value	**11.91%**	**16.34%**	**27.33%**	**31.98%**
Most Predictive Algorithm	DecisionTable (percentage split 66%)	DecisionTable (cross validation 10 folds)	DecisionTable (percentage split 66%)	DecisionTable (percentage split 66%)

As it can be seen in Table 3.6:

1. Given the set of educational inputs introduced in this study, Decision Table is capable of predicting the number of patents with relatively high accuracy.
2. Addition of population as an input does not help with the accuracy of the prediction, showing that the information in population cannot be very informative once the other input factors are processed by the Decision Table.
3. An interesting observation is the error of the educational factors in predicting the intellectual productivity which is less than that of economic factors. This supports the idea that the educational factors may be at least as important (if not more important than) the economic factors when identifying the future productivity of the nation.

3.7 Evaluation of Combined Attribute data

The combined data set includes both economic and educational attributes (see Appendix E for complete listing of attributes that make up the data set). The ReliefFAttributeEval attribute selection algorithm was used to aid in the creation of the combined attributes data sets (see Appendix G for complete run for patent applications). Attribute selection was used to create a data set using both patents applied for and granted patents as the attribute evaluator. The top ten ranked attributes for each run make up the combined data sets.

Table 8 details the performance of the combined data set when used to calculate the number of patent applications and the number of patents granted. See Appendix E for the combined attributes for patents applied for and granted patents.

Table 3.7: Combined Data Results – Applied for and Granted Patents
(see Appendix E for input data attributes)

Classification Algorithm and Test Options	Relative Absolute Error (Patents Applied for)	Relative Absolute Error (Granted Patents)
M5Rules (cross validation 49 folds)	49.66%	50.86%
M5Rules (cross validation 10 folds)	*54.74%*	*59.87%*
M5Rules (percentage split 66%)	32.92%	45.52%
ConjunctiveRule (cross validation 49 folds)	36.99%	48.93%
ConjunctiveRule (cross validation 10 folds)	42.94%	48.12%
ConjunctiveRule (percentage split 66%)	29.04%	43.45%
DecisionTable (cross validation 49 folds)	16.54%	30.57%
DecisionTable (cross validation 10 folds)	16.13%	31.39%
DecisionTable (percentage split 66%)	**11.66%**	**27.33%**

3.8 Combined Attribute Summary

Table 3.8 presents the results of combined using attributes data set that incorporates the information in both economical and educational data to predict intellectual productivity.

Table 3.8: Combined Data Results Summary
(see Appendix E for input data attributes)

Results Summary	Patent Applications	Patents Granted
Mean Relative Absolute Error	28.61%	41.58%
RAE Standard Deviation	15.36%	10.89%
Least Predictive Value	*54.74%*	*59.87%*
Least Predictive Algorithm	M5Rules (cross validation 10 folds)	M5Rules (cross validation 10 folds)
Most Predictive Value	**11.66%**	**27.33%**
Most Predictive Algorithm	DecisionTable (percentage split 66%)	DecisionTable (percentage split 66%)

3.9 Overall Performance When Using Economic, Educational, and Combined Datasets

Table 3.9 shows the overall performance of each data set when used to predict intellectual productivity. The table reveals that the combined attributes data set produced the smallest relative absolute error.

Table 3.9: Complete Analysis Results Summary

Most Predictive Data Set	Algorithm	Most Predictive RAE
Patent Applications		
Economic Data Set (w/o Population)	DecisionTable (cross validation 10 folds)	18.29%
Educational Data Set (w/o Population)	DecisionTable (percentage split 66%)	11.91%
Combined Attributes Data Set	DecisionTable (percentage split 66%)	11.66%
Granted Patents		
Economic Data Set (w/o Population)	DecisionTable (percentage split 66%)	22.64%
Educational Data Set (w/o Population)	DecisionTable (percentage split 66%)	27.33%
Combined Attributes Data Set	DecisionTable (percentage split 66%)	27.33%

As Figure 3.1 shows the relative absolute error when predicting the number of patents applied for is much smaller than the number of granted patents. The main factor for this discrepancy is hypothesized to be because of the time that elapses between when a patent is first filed for and it is granted. This hypothesis could be further investigated by modifying the underlying data sets to include a two to three year mean of previous years' data and then performing the analysis described in this study again.

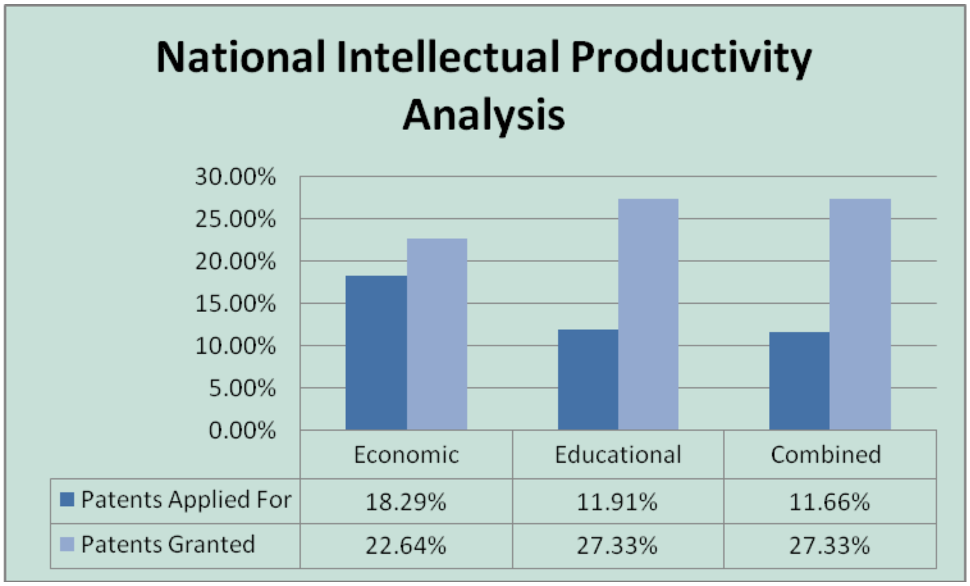

Figure 3.1: Complete Analysis Results Chart

From the results presented above, it is clear that the DecisionTable classifier outperforms the other two methods for forecasting national intellectual productivity in almost all cases.

Although this study reveals that the population alone is not an indication of intellectual productivity, it is understood that the effect of population is apparent through enrollment statistics.

3.10 Summary

This chapter presented the predictive capabilities of both economic and educational factors in estimating the intellectual productivity. A combination of educational and economic factors was also used as a basis for testing –with the majority of the attributes being educational. The results indicates that the Decision Table provides the most suitable

model, among the three models tested, in predicting the intellectual productivity from an economic, educational, and combined input data. The results also indicated that the educational factors can better predict the intellectual productivity, and as such, may be at least as important as the economic factors in identifying the nation's intellectual productivity. Using combined dataset provides slightly better results than just using educational data.

Chapter 4: Conclusions

The main conclusions of the study can be summarized as follows:

- Economic and educational policies were shown to have a tangible relationship with national intellectual productivity.
- Machine learning methods are shown to have the capability of predicting the intellectual productivity with accuracies close of 90%. Such models can allow the government to better control macroeconomic factors and allocate budget and resources towards educational projects in order to optimize intellectual productivity of the nation.
- Education was shown to have the higher impact on intellectual productivity. Educational attributes made up nine out of ten attributes for granted patents combined data set. Of all the attributes, both economic and educational, post-secondary enrollment was shown to have the highest impact on intellectual productivity.
- Of the top ten attributes ranked using attribute selection for patents applied for, six of the ten are from the educational data set.
- This finding demonstrates the value of higher education as it relates to national productivity.

References

[1] I.H. Witten and E. Frank: Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations. Morgan Kaufmann, 1999.

[2] Geoffrey Holmes, Mark Hall, Eibe Frank: Generating Rule Sets from Model Trees. In: Twelfth Australian Joint Conference on Artificial Intelligence, 1-12, 1999.

[3] Ross J. Quinlan: Learning with Continuous Classes. In: 5th Australian Joint Conference on Artificial Intelligence, Singapore, 343-348, 1992.

[4] Y. Wang, I. H. Witten: Induction of model trees for predicting continuous classes. In: Poster papers of the 9th European Conference on Machine Learning, 1997.

[5] Ron Kohavi: The Power of Decision Tables. In: 8th European Conference on Machine Learning, 174-189, 1995.

[6] WEKA Documentation
http://weka.wikispaces.com/Primer
http://weka.sourceforge.net/doc/weka/classifiers/rules/ConjunctiveRule.html

[7] Kenji Kira, Larry A. Rendell: A Practical Approach to Feature Selection. In: Ninth International Workshop on Machine Learning, 249-256, 1992.

[8] Igor Kononenko: Estimating Attributes: Analysis and Extensions of RELIEF. In: European Conference on Machine Learning, 171-182, 1994.

[9] Marko Robnik-Sikonja, Igor Kononenko: An adaptation of Relief for attribute estimation in regression. In: Fourteenth International Conference on Machine Learning, 296-304, 1997.

[10] Igor Kononenko, Se June Hone: Attribute selection for modeling. In: Future Generation Computer Systems, Volume 13, Issues 2-3, 1997
doi: 10.1016/S0167-739X(97)81974-7.

[11] U.S. Department of Education
http://www.ed.gov/about/landing.jhtml

[12] Porter, M.E, Stern, S.: Measuring the 'Ideas' Production Function: Evidence from International Patent Output, In: NBER Working Paper 7891, 2000.

[13] J.L. Furman, M.E. Porter and S. Stern: The determinants of national innovative capacity. In: Research Policy volume 31 pages 899 – 933, 2002.

[14] Jeffrey L. Furman, Richard Hayes: Catching up or standing still? : National innovative productivity among 'follower' countries, 1978-1999. In: Research Policy, Volume 33, Issue 9, pages 1253-1431, November 2004.

[15] U.S. Consolidated Laws. Appendix L Patent Laws, United States Code Title 35 – Patents, page 154.

[16] James C. Cooper: The Fed will be in no rush to raise rates. In: Business Week, 00077135, 6/1/2009, Issue 4133.

[17] G. Rassenfosse, van Pottelsberghe: A first look at the price elasticity of patents. In: Oxford Review of Economic Policy, Volume: 23, Issue: 4, pages: 588 – 604, 2007.

[18] U.S. PATENT AND TRADEMARK OFFICE, Electronic Information Products Division / Patent Technology Monitoring Team, us_stat_08f.xls, June 2009.

[19] Eibe Frank, Yong Wang, Stuart Inglis, Geoffrey Holmes, Ian H. Witten: Using Model Trees for Classification. In: Machine Learning Journal, ISSN: 0885-6125, Volume 32, Number 1, July 1998.

[20] U.S. Government, Energy Information Administration
http://www.eia.doe.gov/emeu/aer/txt/ptb1601.html

[21] United States Census Bureau
http://www.census.gov/population/www/socdemo/school/past-schen.html

[22] U.S. Department of Education Institute of Educational Sciences, National Center for Educational statistics
http://nces.ed.gov/programs/digest/d08/tables/dt08_003.asp

[23] U.S. Patent and Trademark Office, Electronic Information Products Division Patent Technology Monitoring Team (PTMT), U.S. Patent Statistics Chart Calendar Years 1963 – 2008,
http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm
http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.pdf

[24] K. Kira and L. Rendell: A practical approach to feature selection, In: Proceedings of the International. Conference on Machine Learning, ICML-92, pages 49-256, 1992.

[25] M.A. Hall and G. Holmes: Benchmarking attribute selection techniques for discrete class data Mining, In: IEEE Transactions on Knowledge and Data Engineering 15, pages 1437-1447, 2003.

[26] P. Langley and W. Iba: Average-case analysis of a newarest neighbor algorithm. In: Proceedings of the thirteenth international joint conference on artificial intelligence, pages 889-894, 1993.

[27] Salvador Garcia, Alberto Fernandez, and Francisco Herrera: Enhancing the effectiveness and interpretability of decision tree and rule induction classifiers with evolutionary training set selection over imbalanced problems. In: Applied soft computing [1568-4946] Volume 9, Issue 4, pages 1304-1314, 2009.

[28] Victoria J. Hodge, Simon O'Keefe, and Jim Austin: A binary neural decision table classifier. In: Neurocomputing [0925-2312] Volume 69, Issues 16-18, pages 1850 – 1859, October 2006.

[29] J. Furnkranz: Separate-and-conquer rule learning. In: Artificial Intelligence Review 13, pages 3 – 54, 1999.

[30] Ulrich Rűckert and Luc De Raedt: An experimental evaluation of simplicity in rule learning. In: Artificial Intelligence Review 172, pages 19 – 20, 2008.

[31] Frans Coenen and Paul Leng: The effect of threshold values on association rule based classification accuracy. In: Data and Knowledge Engineering, Volume 60, Issue 2, pages 345 – 360, February 2007.

[32] J. Fűrnkranz and F. Widme: Incremental reduced error pruning. In: Proceedings of the 11th International Conference on Machine Learning, Morgan Kaufman, pages 70 – 77, 1994.

[33] G. John, R. Kohavi, and K Pfleger: Irrelevant features and the subset selection problem. In: Machine Learning Proceedings of the Eleventh International Conference, Morgan Kaufmann, pages 121 – 129, 1994.

[34] Ron Kohavi and George H. John: Wrappers for feature subset selection. In: Artificial Intelligence [0004-3702] Volume 97, Issues 1 – 2, pages 273 – 324, 1997. doi:10.1016/S0004-3702(97)00043-X

[35] Arie Ben-David and Eibe Frank: Accuracy of machine larning models versus "hand crafted" expert systems – A credit scoring case study. In: Expert Systems with Applications, Volume 36, Issue 3, Part1, Pages 5264 – 5271, April 2009.

[36] Yue Huang, Paul Jo. McCullagh, and Norman D. Black: An optimization of ReliefF for classification in large datasets. In: Data and Knowledge Engineering, Volume 68, Issue 11, pages 1348 – 1356, November 2009.

[37] Pat Langley and Herbert A. Simon: Applications of machine learning and rule induction. In: Communications of the ACM, Volume 38, Issue 11, pages 54 – 64, ISN: 0001-0782, 1995.

[38] Huawen Liu, Jigui Sun, Lei Liu, and Huijie Zhang: Feature selection with dynamic mutual information. In: Pattern Recognition, Volume 42, Issue 7, pages 1330 – 1339, July 2009.

[39] J.W. Han and M. Kamber: Data Mining: Concepts and Techniques. Morgan Kaufmann Publishers, San Francisco, 2001.

[40] S.S. Panda, A.K. Singh, D. Chakraborty, and S.K. Pal: Drill wear monitoring using back propagation neural network. In: Journal of Materials Processing Technology, Volume 172, pages 283 – 290, 2006.

APPENDIX A – Dataset Attributes

NUM_PATENTS_APPLICATIONS
NUM_PATENTS_GRANTED
UNEMPLOYMENT_RATE
PUBLIC_SCHOOL_ENROLLMENT
PRIVATE_SCHOOL_ENROLLMENT
GDP_Q1
GDP_Q2
GDP_Q3
GDP_Q4
MORTGAGE_RATE_MAX
MORTGAGE_RATE_MIN
MORTGAGE_RATE_MEDIAN
MORTGAGE_RATE_MEAN
SAVINGS_RATE_MAX
SAVINGS_RATE_MIN
SAVINGS_RATE_MEDIAN
SAVINGS_RATE_MEAN
COLLEGE_ENROLLMENT_NUMBER_census
EDU_TOTAL_ENROLLMENT_ALL_LEVELS
EDU_ELEMENTARY_AND_SECONDARY_TOTAL
EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_TOTAL
EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_PRESCHOOL_THROUGH_8
EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_GRADES_9_TO_12
EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_TOTAL
EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_PRESCHOOL_TO_8
EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_GRADES_9_TO_12
EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_TOTAL
EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PUBLIC
EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PRIVATE
US_POPULATION

APPENDIX B – Economic Data Attributes

NUM_PATENTS_APPLICATIONS
NUM_PATENTS_GRANTED
UNEMPLOYMENT_RATE
GDP_Q1
GDP_Q2
GDP_Q3
GDP_Q4
MORTGAGE_RATE_MAX
MORTGAGE_RATE_MIN
MORTGAGE_RATE_MEDIAN
MORTGAGE_RATE_MEAN
SAVINGS_RATE_MAX
SAVINGS_RATE_MIN
SAVINGS_RATE_MEDIAN
SAVINGS_RATE_MEAN

APPENDIX C – Educational Data Attributes

NUM_PATENTS_APPLICATIONS
NUM_PATENTS_GRANTED
PUBLIC_SCHOOL_ENROLLMENT
PRIVATE_SCHOOL_ENROLLMENT
COLLEGE_ENROLLMENT_NUMBER_census
EDU_TOTAL_ENROLLMENT_ALL_LEVELS
EDU_ELEMENTARY_AND_SECONDARY_TOTAL
EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_TOTAL
EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_PRESCHOOL_THROUGH_8
EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_GRADES_9_TO_12
EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_TOTAL
EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_PRESCHOOL_TO_8
EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_GRADES_9_TO_12
EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_TOTAL
EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PUBLIC
EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PRIVATE

APPENDIX D – NUMBER OF PATENTS APPLIED FOR AND GRANTED BY THE USPTO

YEAR	PATENT APPLICATIONS	PATENTS GRANTED
2008	231588	77501
2007	241347	79526
2006	221784	89823
2005	207867	74637
2004	189536	84270
2003	188941	87893
2002	184245	86971
2001	177511	87600
2000	164795	85068
1999	149825	83905
1998	135483	80289
1997	120445	61708
1996	106892	61104
1995	123958	55739
1994	107233	56066
1993	99955	53231
1992	92425	52253
1991	87955	51177
1990	90643	47391
1989	82370	50184
1988	75192	40498
1987	68315	43519
1986	65487	38126
1985	63874	39556
1984	61841	38373
1983	59390	32868
1982	63316	33890
1981	62404	39218
1980	62098	37350
1979	60535	30074

1978	61441	41250
1977	62863	41488
1976	65050	44280
1975	64445	46712
1974	64093	50646
1973	66935	51501
1972	65943	51519
1971	71089	55975
1970	72343	47073
1969	68243	50394
1968	67180	45781
1967	61651	51274
1966	66855	54634
1965	72317	50331
1964	67013	38410
1963	66715	37174

APPENDIX E – Combined Attributes for Patents Applied for and Granted Patents

Patents Applied For

0.0885	27	EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PUBLIC
0.0669	26	EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_TOTAL
0.0669	24	EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_PRESCHOOL_TO_8
0.0608	17	COLLEGE_ENROLLMENT_NUMBER_census
0.0593	13	SAVINGS_RATE_MAX
0.0588	22	EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_GRADES_9_TO_12
0.0527	5	GDP_Q1
0.0516	6	GDP_Q2
0.0506	7	GDP_Q3
0.0504	28	EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PRIVATE

Granted Patents

0.06922	26	EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_TOTAL
0.06922	24	EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_PRESCHOOL_TO_8
0.05677	27	EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PUBLIC
0.05212	21	EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_PRESCHOOL_THROUGH_8
0.04966	19	EDU_ELEMENTARY_AND_SECONDARY_TOTAL
0.04574	20	EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_TOTAL
0.04574	3	PUBLIC_SCHOOL_ENROLLMENT
0.03455	18	EDU_TOTAL_ENROLLMENT_ALL_LEVELS
0.03293	14	SAVINGS_RATE_MIN
0.02821	4	PRIVATE_SCHOOL_ENROLLMENT

APPENDIX F – Dimensionality of Original Data

```
@relation data_v2-weka.filters.unsupervised.attribute.Remove-R1

@attribute NUM_PATENTS_APPLICATIONS numeric
@attribute NUM_PATENTS_GRANTED numeric
@attribute POP_PATENT_APPLIED_FOR_RATIO numeric
@attribute POP_PATENT_GRANTED_RATIO numeric
@attribute GRANTED_RATIO_ABOVE_MEAN {NO,YES}
@attribute APPLIED_FOR_RATION_ABOVE_MEAN {NO,YES}
@attribute UNEMPLOYMENT_RATE numeric
@attribute PUBLIC_SCHOOL_ENROLLMENT numeric
@attribute PRIVATE_SCHOOL_ENROLLMENT numeric
@attribute GDP_Q1 numeric
@attribute GDP_Q2 numeric
@attribute GDP_Q3 numeric
@attribute GDP_Q4 numeric
@attribute MORTGAGE_RATE_MAX numeric
@attribute MORTGAGE_RATE_MIN numeric
@attribute MORTGAGE_RATE_MEDIAN numeric
@attribute MORTGAGE_RATE_MEAN numeric
@attribute SAVINGS_RATE_MAX numeric
@attribute SAVINGS_RATE_MIN numeric
@attribute SAVINGS_RATE_MEDIAN numeric
@attribute SAVINGS_RATE_MEAN numeric
@attribute COLLEGE_ENROLLMENT_NUMBER_census numeric
@attribute EDU_TOTAL_ENROLLMENT_ALL_LEVELS numeric
@attribute EDU_ELEMENTARY_AND_SECONDARY_TOTAL numeric
@attribute EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_TOTAL numeric
@attribute EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_PRE_TO_8 numeric
@attribute EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_9_TO_12 numeric
@attribute EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_TOTAL numeric
@attribute EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_PRESCHOOL_TO_8 numeric
@attribute EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_GRADES_9_TO_12 numeric
@attribute EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_TOTAL numeric
@attribute EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PUBLIC numeric
@attribute EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PRIVATE numeric
@attribute US_POPULATION numeric
```

APPENDIX F (cont'd) – Dimensionality of Original Data

Note: The full dataset contains 47 years worth of data (47 rows). For formatting purposes the entire dataset was not shown. Some of the original dataset that was created (i.e. GRANTED_RATIO_ABOVE_MEAN and APPLIED_FOR_RATIO_ABOVE_MEAN) is show below, but was not used during this study's classification analysis.


```
@data
231588,77501,1313.1,3923.8,NO,NO,5.6,49825,6054,11646,11727,11712,11522,6.48,5.33,6.04,6.04,4.8,0,1.55,1.78,
241347,79526,1248.4,3788.6,NO,NO,4.6,49644,6066,11358,11491,11626,11621,6.7,6.1,6.27,6.34,1.3,0.1,0.5,0.55,1
221784,89823,1345.4,3322,NO,NO,4.6,49299,6095,11217,11292,11314,11356,6.76,6.14,6.38,6.41,1.1,0.2,0.8,0.73,1
207867,74637,1422,3960.5,NO,NO,5.1,49113,6073,10878,10954,11050,11086,6.33,5.58,5.79,5.86,4.5,0.1,0.8,1.09,1
189536,84270,1545.3,3475.7,NO,NO,5.5,48795,6087,10612,10672,10729,10796,6.29,5.45,5.75,5.84,4.5,1.4,1.95,2.0
188941,87893,1535.9,3301.7,NO,NO,6,48540,6099,10139,10230,10411,10503,6.26,5.23,5.86,5.81,2.8,1.7,2.1,2.13,1
184245,86971,1561.5,3307.9,NO,NO,5.8,48183,6220,9977,10032,10091,10096,7.01,6.05,6.57,6.53,3.1,1.5,2.4,2.35,
177511,87600,1605.5,3253.4,NO,NO,4.7,47672,6320,9876,9906,9871,9910,7.16,6.62,7.04,6.97,4.2,0.2,1.65,1.8,158
164795,85068,1712.4,3317.3,NO,NO,4,47204,6169,9696,9848,9837,9888,8.52,7.38,8.15,8.06,2.9,1.5,2.4,2.36,15314
149825,83905,1862.1,3325.1,NO,NO,4.2,46857,6018,9316,9393,9502,9671,7.94,6.79,7.59,7.42,4,1.4,2.1,2.37,15203
135483,80289,2036.4,3436.3,NO,NO,4.5,46539,5988,8936,8995,9099,9237,6.72,6.72,6.72,6.72,4.7,3.5,4.4,4.31,155
120445,61708,2263.2,4417.5,NO,NO,4.9,46127,5944,8536,8666,8774,8838,7.1,7.1,7.1,7.1,4,3.3,3.7,3.65,15436,665
106892,61104,2520.3,4408.8,NO,NO,5.4,45611,5933,8169,8303,8373,8471,8.32,7.03,7.92,7.8,4.5,3.1,4.1,4,15226,6
123958,55739,2148.3,4777.6,YES,NO,5.6,44840,5918,7974,7988,8053,8112,9.15,7.2,7.75,7.95,5.9,3.6,4.5,4.65,147
107233,56066,2453.5,4692.6,YES,NO,6.1,44111,5787,7715,7816,7860,7952,9.2,7.06,8.55,8.35,5.8,3.9,5.05,4.82,15
99955,53231,2600.1,4882.4,YES,NO,6.9,43465,5668,7460,7498,7536,7637,8.02,6.83,7.31,7.33,7.6,5,5.6,5.76,11901
92425,52253,2775.2,4908.8,YES,YES,7.5,42823,5677,7228,7298,7370,7451,8.94,7.92,8.37,8.4,9.4,6.9,7.7,7.7,1167
87955,51177,2876.4,4943.6,YES,YES,6.8,42047,5681,7041,7086,7121,7154,9.64,8.5,9.42,9.24,7.9,6.6,7.25,7.25,11
90643,47391,2753.6,5266.8,YES,YES,5.6,41217,5648,7112,7130,7131,7077,10.48,9.67,10.17,10.13,7.3,6.6,7.05,6.9
82370,50184,2996.2,4917.9,YES,YES,5.3,40543,5599,6918,6964,7013,7031,11.05,9.74,10.16,10.32,8.3,6.4,7.1,7.15
75192,40498,3251.6,6037.3,YES,YES,5.5,40189,5242,6639,6724,6759,6849,10.61,9.89,10.43,10.33,7.6,7,7.2,7.28,1
68315,43519,3546.8,5567.6,YES,YES,6.2,40008,5479,6365,6435,6493,6607,11.26,9.04,10.43,10.19,8.8,3.5,7.3,6.96
65487,38126,3666.3,6297.5,YES,YES,7,39753,5452,6207,6232,6292,6323,10.88,9.31,10.11,10.17,9.5,5.9,8.45,8.17,
63874,39556,3724.5,6014.2,YES,YES,7.3,39422,5557,5957,6008,6102,6149,13.2,11.26,12.2,12.42,9.5,0.1,8.4,6.51,
61841,38373,3813,6144.9,YES,YES,7.5,39208,5700,5700,5798,5854,5902,14.67,13.18,13.79,13.87,9.4,0.5,0.9,1.63,
59390,32868,3936.6,7113.3,YES,YES,10.4,39252,5715,5254,5372,5478,5590,13.81,12.63,13.33,13.22,9.9,0,8.75,8.1
63316,33890,3659.4,6836.8,YES,YES,9.7,39566,5600,5177,5205,5185,5190,17.6,13.62,16.69,16.08,9.9,0.3,1.65,2.8
62404,39218,3677.6,5851.9,YES,YES,7.5,40044,5500,5308,5266,5330,5263,18.45,14.9,16.76,16.63,9.9,0,2.5,5,1073
62098,37350,3658.7,6082.9,YES,YES,7.1,40877,5331,5221,5116,5107,5202,16.33,12.19,13.49,13.77,9.8,0,5.1,5.02,
60535,30074,3718.5,7484.8,YES,YES,5.9,41651,5000,5147,5152,5189,5205,12.9,10.39,11.06,11.19,9.5,7.9,8.95,8.8
61441,41250,3622.9,5396.3,YES,YES,6.1,42551,5086,4831,5021,5071,5137,10.35,9.02,9.72,9.63,9.9,8.1,8.85,8.86,
62863,41488,3502.8,5307.5,YES,YES,7.5,43577,5140,4640,4731,4816,4815,8.96,8.67,8.88,8.84,9.4,7.1,8.65,8.7,10
65050,44280,3351.2,4923.2,YES,YES,7.7,44311,5167,4497,4530,4552,4585,9.02,8.73,8.83,8.86,9.9,0.1,9.35,8.56,9
64445,46712,3351.6,4624,YES,YES,8.1,44819,5000,4238,4269,4341,4398,9.43,8.82,9.02,9.04,9.7,0,1.35,3.9,9697,6
```

APPENDIX G – Feature Selection Results

=== Run information for Patent Applications ===

Evaluator: weka.attributeSelection.ReliefFAttributeEval -M -1 -D 1 -K 10
Search: weka.attributeSelection.Ranker -T -1.7976931348623157E308 -N -1
Relation: data_v2-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R2-6
Instances: 49
Attributes: 29

NUM_PATENTS_APPLICATIONS
UNEMPLOYMENT_RATE
PUBLIC_SCHOOL_ENROLLMENT
PRIVATE_SCHOOL_ENROLLMENT
GDP_Q1
GDP_Q2
GDP_Q3
GDP_Q4
MORTGAGE_RATE_MAX
MORTGAGE_RATE_MIN
MORTGAGE_RATE_MEDIAN
MORTGAGE_RATE_MEAN
SAVINGS_RATE_MAX
SAVINGS_RATE_MIN
SAVINGS_RATE_MEDIAN
SAVINGS_RATE_MEAN
COLLEGE_ENROLLMENT_NUMBER_census
EDU_TOTAL_ENROLLMENT_ALL_LEVELS
EDU_ELEMENTARY_AND_SECONDARY_TOTAL
EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_TOTAL
EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_PRESCHOOL_THROUGH_8
EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_GRADES_9_TO_12
EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_TOTAL
EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_PRESCHOOL_TO_8
EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_GRADES_9_TO_12
EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_TOTAL
EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PUBLIC
EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PRIVATE
US_POPULATION

Evaluation mode: evaluate on all training data

APPENDIX G (cont'd) – Feature Selection Results

=== Attribute Selection on all input data ===

Search Method:
    Attribute ranking.

Attribute Evaluator (supervised, Class (numeric): 1 NUM_PATENTS_APPLICATIONS):
    ReliefF Ranking Filter
    Instances sampled: all
    Number of nearest neighbours (k): 10
    Equal influence nearest neighbours

Ranked attributes:

```
 0.0885   27 EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PUBLIC
 0.0669   26 EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_TOTAL
 0.0669   24 EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_PRESCHOOL_TO_8
 0.0608   17 COLLEGE_ENROLLMENT_NUMBER_census
 0.0593   13 SAVINGS_RATE_MAX
 0.0588   22 EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_GRADES_9_TO_12
 0.0527    5 GDP_Q1
 0.0516    6 GDP_Q2
 0.0506    7 GDP_Q3
 0.0504   28 EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PRIVATE
 0.0465    8 GDP_Q4
 0.0443   18 EDU_TOTAL_ENROLLMENT_ALL_LEVELS
 0.0345   29 US_POPULATION
 0.0187   19 EDU_ELEMENTARY_AND_SECONDARY_TOTAL
 0.0153   20 EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_TOTAL
 0.0153    3 PUBLIC_SCHOOL_ENROLLMENT
-0.016    16 SAVINGS_RATE_MEAN
-0.0211    2 UNEMPLOYMENT_RATE
-0.0277   21 EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_PRESCHOOL_TO_8
-0.0332   25 EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_GRADES_9_TO_12
-0.0352   15 SAVINGS_RATE_MEDIAN
-0.0365   14 SAVINGS_RATE_MIN
-0.0395    4 PRIVATE_SCHOOL_ENROLLMENT
-0.0395   23 EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_TOTAL
-0.0701   10 MORTGAGE_RATE_MIN
-0.0784   11 MORTGAGE_RATE_MEDIAN
-0.0807   12 MORTGAGE_RATE_MEAN
-0.0948    9 MORTGAGE_RATE_MAX
```

APPENDIX H – Decision Table Results

=== Run information for Educational Dataset ===

Scheme: weka.classifiers.rules.DecisionTable -X 1 -R -S "weka.attributeSelection.BestFirst -D 1 -N 5"
Relation: data_v2-weka.filters.unsupervised.attribute.Remove-R1-weka.filters.unsupervised.attribute.Remove-R2-7,10-21-weka.filters.unsupervised.attribute.Remove-R16
Instances: 49
Attributes: 15
NUM_PATENTS_APPLICATIONS
PUBLIC_SCHOOL_ENROLLMENT
PRIVATE_SCHOOL_ENROLLMENT
COLLEGE_ENROLLMENT_NUMBER_census
EDU_TOTAL_ENROLLMENT_ALL_LEVELS
EDU_ELEMENTARY_AND_SECONDARY_TOTAL
EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_TOTAL

EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_PRESCHOOL_THROUGH_8
EDU_PUBLIC_ELEMENTARY_AND_SECONDARY_SCHOOLS_GRADES_9_TO_12
EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_TOTAL
EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_PRESCHOOL_TO_8
EDU_PRIVATE_ELEMENTARY_AND_SECONDARY_GRADES_9_TO_12
EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_TOTAL
EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PUBLIC
EDU_POSTSECONDARY_DEGREE_INSTITUTIONS_PRIVATE
Test mode: split 66.0% train, remainder test

=== Classifier model (full training set) ===

Decision Table:

Number of training instances: 46
Number of Rules : 11
Non matches covered by Majority class.
Best first.
Start set: no attributes
Search direction: forward
Stale search after 5 node expansions
Total number of subsets evaluated: 61
Merit of best subset found: 15087.07
Evaluation (for feature selection): CV (leave one out)
Feature set: 5,1

APPENDIX H (cont'd) – Decision Table Results

Rules:

```
=====================================================================
EDU_TOTAL_ENROLLMENT_ALL_LEVELS NUM_PATENTS_APPLICATIONS
=====================================================================
'(-inf-58842.9]'            64245.2
'(58842.9-60535.8]'         68037.0
'(60535.8-62228.7]'         81014.33333333333
'(62228.7-63921.6]'         96190.0
'(63921.6-65614.5]'         115595.5
'(65614.5-67307.4]'         120940.0
'(67307.4-69000.3]'         157310.0
'(69000.3-70693.2]'         177511.0
'(70693.2-72386.1]'         187574.0
'(72386.1-inf)'             225646.5
=====================================================================
```

Time taken to build model: 0.16 seconds

=== Predictions on test split ===

```
 inst#,    actual, predicted, error
    1 120445     121187.5      742.5
    2  61651     68266.75    6615.75
    3 221784     236467.5    14683.5
    4 207867     236467.5    28600.5
    5  60535     65254.667   4719.667
    6 189536     186593      -2943
    7  68243     69107.286    864.286
    8  65487     65254.667   -232.333
    9 164795     149825     -14970
   10  61841     65254.667   3413.667
   11  67013     68266.75    1253.75
   12  66935     69107.286   2172.286
   13  62404     65254.667   2850.667
   14  62098     65254.667   3156.667
```

=== Evaluation on test split ===
=== Summary ===

```
Correlation coefficient                  0.9905
Mean absolute error                   6229.898
Root mean squared error               9891.5053
Relative absolute error                 11.914  %
Root relative squared error             16.4414 %
Total Number of Instances               14
```

APPENDIX I – Source and Links to Data

Educational Data Sources

- U.S. Department of Education
    - http://www.ed.gov/about/landing.jhtml

- National Center for Educational statistics - U.S. Department of Education Institute of Educational Sciences
    - http://nces.ed.gov/programs/digest/d08/tables/dt08_003.asp

Economic Data Sources

- U.S. Gross Domestic Product
    - Department of Commerce (DOC), Bureau of Economic Analysis
        - http://www.eia.doe.gov/emeu/aer/txt/ptb1601.html

- U.S Unemployment Rate
    - Department of Labor, Bureau of Labor Statistics
        - http://www.bls.gov/cps/tables.htm
    - The Wall Street Journal
        - http://online.wsj.com/public/resources/documents/JOBSHISTORY09.html

- U.S. Savings Rate
    - U.S. Department of Commerce, Bureau of Economic Analysis
        - http://www.bea.gov/national/nipaweb/Nipa-Frb.asp
            - NIPATable.csv

- U.S. Mortgage Rate
    - Board of Governors of the Federal Reserve System
        - Federal Reserve Bank of St. Louis
            - http://research.stlouisfed.org/fred2/series/MORTG/downloaddata?cid=114
                - MORTG.xls

APPENDIX I (cont'd) – Source and Links to Data

U.S. Population Data Source

- Department of Commerce (DOC), U.S. Bureau of the Census.
    - http://www.eia.doe.gov/emeu/aer/txt/ptb1601.html

U.S. Patent Data

- U.S. Patent and Trademark Office, Electronic Information Products Division Patent Technology Monitoring Team (PTMT), U.S. Patent Statistics Chart Calendar Years 1963 – 2008.
    - http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.htm
    - http://www.uspto.gov/web/offices/ac/ido/oeip/taf/us_stat.pdf

VITA

Canon Edward Fazenbaker was born on April 25, 1981 in Cumberland, MD and grew up in New Creek, WV. He earned his Eagle Scout in the fall of 1997 and graduated from Keyser High School in 1999. He then attended Potomac State College for one year before transferring to West Virginia University Institute of Technology. There he majored in Computer Science and participated in the co-op program. He received a co-op position at with Dominion Power working at North Anna Nuclear Power Station in the component engineering department. He served as vice-president for the Association for Computing Machinery (ACM) chapter at WVU Tech - and still continues his ACM membership today. After receiving his Computer Science Bachelor of Science degree in 2005 he moved to Richmond, VA to continue his employment with Dominion Power and further his education by attending Virginia Commonwealth University and working towards a Master of Science in Computer Science.