Using Machine Learning Technique for Telecom Service Providers in Malaysia to Prioritize Broadband Investment in Urban and Rural Areas

Taik Guan Tan[1], Dino Isa[2], Yee Wan Wong[3], Jasmine PY Chen[4] and Mei Shin Oh[5]

[1] University of Nottingham, Malaysia Campus
Faculty of Engineering, Jalan Broga, Semenyih, Selangor 43500, Malaysia

[2] University of Nottingham, Malaysia Campus
Faculty of Engineering, Jalan Broga, Semenyih, Selangor 43500, Malaysia

[3] University of Nottingham, Malaysia Campus
Faculty of Engineering, Jalan Broga, Semenyih, Selangor 43500, Malaysia

[4] Telekom Malaysia Berhad
TMR&D Innovation Centre, Lingkaran Teknokrat Timur, Cyberjaya, Selangor, Malaysia

[5] University of Nottingham, Malaysia Campus
Faculty of Engineering, Jalan Broga, Semenyih, Selangor 43500, Malaysia

Corresponding author's email: oritalk [AT] gmail.com

ABSTRACT— ***The companies that provide telecommunication services (Telco) in Malaysia commonly use the return on investment (ROI) model to strategize their network investment plans and to deploy their broadband services in their intended markets. The numbers of subscribers and average revenue per user (ARPU) are two dominant contributions to a good ROI. The rural areas are lacking both dominant factors and thus very often fall outside the radar of the Telco's investment plans. The government agencies, therefore, shoulder the responsibility to provide broadband services in rural areas through the implementation of national broadband initiatives and regulated policies and funding for universal service provision. In this paper, we will outline a machine-learning technique which the Telco can use to plan for broadband investments in urban areas and beyond. The proposed technique predicts the socioeconomic potential of a geographical area in correspondence to its local characteristics. This technique is an empirical model that produces a correlation coefficient to quantify the statistical relationships between two or more values of local characteristics and socioeconomic potential. The model can help Telcos to prioritize their investments in urban and rural areas with higher potential for socioeconomic growth. By using this technique as a policy tool, Telcos will be able to prioritize areas where broadband infrastructure can be implemented using a government-industry partnership approach. Both public and private parties can share the initial cost and collect future revenues appropriately as the socioeconomic correlation coefficient improves. The proposed technique functions to formulate an empirical model using a curve-fitting software and to generate sufficient data using Genetic Algorithm to train a Support Vector Machine.***

Keywords— Machine Learning, Curve Fitting, Genetic Algorithm, Support Vector Machine

1. INTRODUCTION

According to the World Bank report (2009), every 10% increase in broadband penetration in developing economies will accelerate economic GDP growth by about 1.38%[1]. According to the International Telecommunication Union's (ITU) 2012 report, a 10% increase in broadband penetration will contribute to a 0.7 percentage increase in Malaysia's GDP[2].

Many types of research (e.g., Röller and Waverman[3], Kuppusamy et al.[4], Shiu and Lam[5], Qiang et al. [6], Czernich et al.[7], Booz & Co.[8], Ericsson[9], Katz and Koutroumpis[10], McKinsey[11], Katz & Berry[12], and so forth) have found that broadband services (BS) have a positive impact on socioeconomic statuses.

In fact, many types of research (e.g., Dewan and Kraemer[13], OECD[14], Gruber and Verboven[15], Daveri[16], Waverman et al.[17], Chakraborty and Nandi[18], Choudrie and Dwivedi[19], Karner and Onyeji[20], Kongout at et al.[21], Sabbagh et al.[22], Lucas[23], Orbicom-ITU[24], Uppal and Mamta[25], Cronin et al.[26][27][28], Wolde-Rufael[29], and so forth) have addressed the different magnitudes of the positive impacts of BS, depending on the current economies of those geographical areas.

However, only the developing and developed economies are normally assumed to be commercially viable for private investments. The underdeveloped economies require legislative support for broadband development because quick profits are unattainable in these areas. In this case, telecommunication investments can enhance economic activity which in turn justifies investments in telecom infrastructure.

Collectively, the various research results provide confidence to the privately owned Telco to provide broadband services in urban and selected suburban (but not rural) areas. Furthermore, there are no empirical models made available to Telcos for the prediction of the economic potential of certain geographical areas so that they can prioritize their network investments in promising rural areas. As a result, the Telcos continue using the return on investment (ROI) model to strategize their network investment plan and to deploy their broadband services (BS) in urban or suburban areas only. Box 1 shows a comparison between ROI model and empirical model.

Box 1: Comparison of ROI model vs. Empirical model

> *ROI* = function {revenue, capital expenditure, operating expenses}
>
> *R* = function {local characteristics such as technical, market & community , government and local authority attributes}
>
> Whereby, R is the correlation coefficient that quantifies the socioeconomic depending on the values of local characteristics of a geographical area.

This paper aims to provide a framework that enables the prediction of the socioeconomic potential in correspondence to the local characteristics of a geographical area.

ROI is a common business term used to identify the potential financial returns, which indicates how successful an investment will be. ROI is also commonly used as a financial performance measure to evaluate the efficiency of different investments. Typically, ROI is expressed as a percentage or a ratio of financial return. A high ROI means the investment's gains compare favorably to its cost. Sometimes ROI is expressed as a number of years to recover the financial investment.

We can test the R hypothesis if the local characteristics could be formulated into an empirical model. The result of R indicates the socio-economic potential as high potential (value 1) or low potential (value 0).

This paper proposes to use a curve-fitting technique to formulate the empirical model by using the limited data from World Bank. The empirical model is then used as a fitness function for GA to generate more data to train an SVM.

It will be ideal to use the data on local characteristics of rural areas to help the curve-fitting to find the empirical model. However, the data on local characteristics of rural areas are lacking or difficult to obtain. Hence, the data on local characteristics of countries from the World Bank database is applied.

The curve-fitting technique is a statistical model useful for identifying patterns of data. The ability to identify the patterns of the data allows better understanding and optimisation of the learning process [30]. Statistical models are empirical expressions of reasoning rather than physical process whereby it can make approximate conclusions with the precision of a computer and the accuracy of a mathematician’s proof. The curve-fitting technique can produce the best curve for any empirical function that best fits a sequence of data points. It examines the relationship between given sets of independent and dependent variable features. The fitted curves obtained can be used in data visualization, and finding relationships among two or more variable features [31].

The availability of data from the World Bank’s databases is limited by the number of countries worldwide. The few hundred data sets in the World Bank’s database provide a small sample available to train and optimize the accuracy of the SVM. The limited sample size will affect the accuracy of machine learning. Hence, large virtual samples are essential to address the issue of insufficient raw data, which will help overcome the problem for the poor and unreliable performance of machine learning and data mining techniques [32].

This paper proposes to use GA to generate more training data by using the model obtained from the curve fitting software as a fitness function in GA.GA can generate large virtual samples when only small amounts of data are available for machine learning [33].

Both the World Bank's data sets and GA generated data sets can be used to train the SVM to classify the R response and to observe the SVM's accuracy in performing the classification with different kernels.

The SVM classifies the maximum margin between two linearly separable classes [34], in this case between 0 (low socioeconomic potential) and 1 (high socioeconomic potential). Training data is provided for the machine to learn the input and output functionality of the data provided. SVM is based on the statistical learning theory and is applicable to various areas [35]. The SVM model is often preferred due to its high computational efficiency and good generalization theory, which prevents over-fitting through the control of hyperplane margins and Structural Risk Minimization [36].

The kernel method is widely used for linear classifier algorithms to solve nonlinear separable problems by mapping the features into a higher dimensional feature space. This separation allows the linear classifier to make the linear classification in a new higher dimensional space equivalent to a nonlinear classification in the original space [37]. The nonlinear SVM can be built by mapping the nonlinear input vector into a high dimensional feature space and constructing an optimal hyperplane to classify the data in the feature space as shown in Figure 1[38].

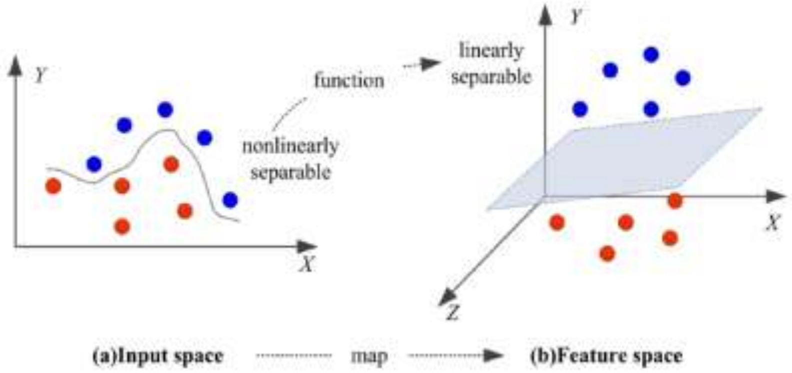


Figure 1: Comparison of ROI model vs. Machine learning model

2. CURRENT STATE OF THE ART OF BROADBAND SERVICES

Typically, Telcos use the ROI (return on investment) model to prioritize their investment in deploying broadband services to benefit a new area or to upgrade their services in such particular areas. As shown in Box 2, the three variable factors in the ROI model are revenue, capital expenditure and the incremental cost of the Telco operating model. (Refer to Appendix I for details of operating model.)

Box 2: ROI Formula

$$\text{ROI} = \frac{\text{Revenue} - \text{Incremental Cost}}{\text{Capital Expenditure}}$$

Whereby,
Revenue = ARPU x number of subscribers
Incremental Cost = operating cost over time
Capital Expenditure = upfront capital investment

The telco will prioritize its BS in geographical areas that are highly affordable and adopt-able, which will, in turn, shorten the ROI period. *Affordability* is a comparison of broadband prices over income level, which indicates the users' ability to pay for the broadband services. *Adopt-ability* is the increase in some subscribers who are willing to pay for the BS for improvements in work-style, lifestyle, and socioeconomic standards.

Both the adopt-ability and affordability of broadband services have domineering impacts on the ROI period. The BS adopt-ability and affordability will generate the number of paying subscribers who will contribute to the revenue.

According to the Malaysian Communications and Multimedia Commission (MCMC), the national broadband penetration has reached 70.2% in the year 2014. In the corresponding period, the World Bank recorded Malaysia's urban population at 74.01%. The network coverage by Malaysian Telcos (Table 1) reflects a relationship between the broadband networks deployed by them and the populations in urban areas.

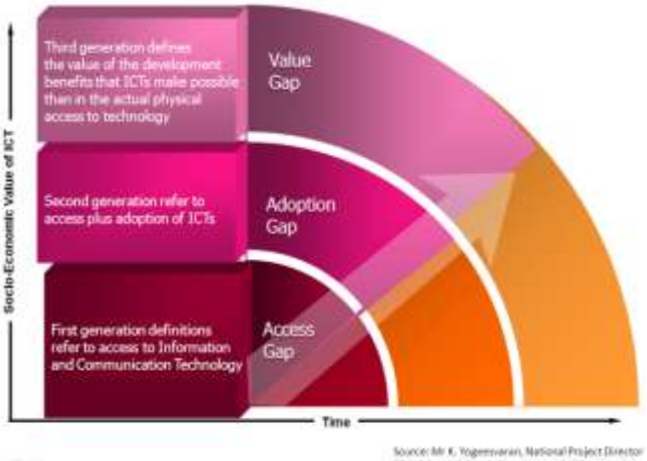
Table 1: Network Coverage by Malaysian Telco

Telco	***Network Coverage by Population***	***Broadband Technology***
Maxis	80% [39]	3G
Celcom	80% [39]	3G
Digi	60% [39]	3G
P1	50% [40]	WiMAX
YES	65% [41]	WiMAX

Beyond urban and suburban areas, the government agencies continue with their national building agenda by executing universal service provision (USP) policies and national broadband initiatives (NBI) to ensure BS *availability* and *accessibility* in unserved or underserved areas.

BS availability means providing broadband infrastructure to enable the users to connect to the Internet world and to prepare them for the accessibility stage. BS accessibility means providing devices, applications, and quality services so that the users can use the broadband services. BS accessibility means enabling users to experience online applications and to realize its benefits to their socio-economic growth. In this stage, the educational level of the people in this geographical context needs to be enhanced, especially in the areas of language proficiency and fundamental ICT skills for the use of broadband services.

The NBI and USP aim to reduce the digital divide between urban and rural areas. Figure 2 illustrates the digital divide defined by the Malaysia Communications and Multimedia Commission (MCMC).


Third generation defines the value of the development benefits that ICTs make possible than in the actual physical access to technology
Value Gap
Second generation refer to access plus adoption of ICTs
Adoption Gap
First generation definitions refer to access to Information and Communication Technology
Access Gap
Socio-Economic Value of ICT
Time

Source: Mr K. Yogeesvaran, National Project Director EPU, Prime Minister's Department, 2007

Figure 2: Definition of Digital Divide (Source: MCMC)

The USP projects fill the “access gap” and nurture the closure of the adoption gap to reach a point that is feasible for Telco’s investments. The government can then form a public-private-partnership with the Telcos to fill up the value gap that stimulates further socioeconomic growth, which will in turn drive BS adopt-ability and affordability that are favorable to the Telco’s ROI.

If rural areas with comparative socioeconomic potential can be identified, then the government can focus its effort on improving the BS availability and accessibility for those selected areas. Public investment can then be bridged with Telco’s investment to expand and upgrade the BS, which will in turn further accelerate the socioeconomic growth of these areas. Consequently, the government can have a sustainable USP program while the Telco can have a continual business plan beyond urban areas. For those rural areas with the least potential for socioeconomic growth, the government can employ a different strategy to reduce the digital divide. (Note: The areas with least potential and the corresponding governmental strategy are not within the scope of this paper.)

It is complex to have an empirical model to address Telco’s desire and national building agenda in rural areas. The complexity varies due to a hybrid of variable attributes influencing socioeconomic growth and BS development. This paper aims to use a machine-learning technique to produce an empirical model for the Telco to prioritize its BS development outside urban areas with the most potential for growth. Consequently, the government can strategize its USP project in collaboration with Telcos to prioritize its BS development with sustainable benefits.

3. METHODOLOGY

Empirical model: Using a curve-fitting software to build an empirical model of GNI per capita with raw data of local characteristics such as land size, daily temperature, household income, % of labor force, length of tar road, number of schools, etc. This step functions to establish an empirical equation to correlate the selected characteristics to a response

(R), which is the potential index of GNI per capita.

GA for Virtual Samples: Using GA to generate virtual samples. This step functions to produce more data to train a classifier SVM.

SVM for Classification: Using the World Bank's data and GA generated data to train SVM. SVM predicts outcomes of the test data by classifying the data into either class 1 or 0.

3.1 Empirical Model

The three sub-steps to produce an empirical model are:

a. Defining the data

b. Collecting data

c. Generating an empirical equation from the data

Defining the Data: Table 2 illustrates the examples of characteristics obtained from the World Bank's database, and the Department of Statistics Malaysia (DOSM). The characteristics can be divided into four (4) categories according to Telco's industry practices.

Table 2: Category of local characteristics

Category	**Characteristics or features (examples)**
Technical Barriers	Land size, distance from nearest town, availability of adjacent wireless broadband network, average monthly rainfall, average daily temperature
Market Demography and Community Structure	% of land size for agro-economic activity, number of households, average dependency per household, household income, number of populations, population density, % of labour force, % of household with computer access, % of fixed broadband penetration, % of wireless broadband penetration, % of fixed telephony penetration, GDP, GDP per capita, GNI, GNI per capita, distance from nearest wholesales agro-market, average age, gender ratio, birthrate, % of population~~s~~ with secondary education, % of population with post-secondary education, labour force count in last 3 years, household count in last 3 years, GDP contribution % by different industries
Government Initiatives	Length of TAR road, % of households with grid electricity, % of households with piped water, number of secondary schools available, ICT investments in the last 3 years, non-ICT investments in the last 3 years
Local Authority	Number of local authority offices available, number of post offices available, number of years of community broadband center available

Collecting Data: For each pre-defined category, the relevant features of each country are extracted from the World Bank's National Accounts data and OECD National Accounts data files online. The data from the countries with common features are used as samples to build the empirical model. The country's GNI per capita is taken as the outcome of the influence of the features chosen.

Generate an Empirical Model from the Data: A curve-fitting software provided by Design Expert is used to establish a co-relationship among the features of a locality in response to the country's income indicator.

3.2 GA for Virtual Samples

The equation generated by the Design Expert software (from Stat-Ease Inc.) is used as a fitness function of the GA to generate more data that can be used to train the SVM. The GA generated data is divided into different proportions to compare the training and testing of the SVM in the next step. This paper proposes to split the data into different ratios to train and test the SVM. The train-test ratio by default is 50:50. Other proposed ratios are 90:10, 80:20, 70:30 and 60:40.

When training the SVM, the training data is divided into three different scales of cross-validation, for example, 10-fold cross-validation and hold on one cross-validation (cv) as illustrated in Figure 3.

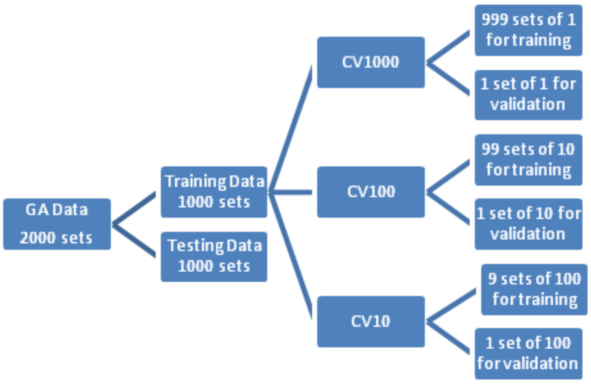

GA Data
2000 sets
Training Data
1000 sets
Testing Data
1000 sets
CV1000
CV100
CV10
999 sets of 1 for training
1 set of 1 for validation
99 sets of 10 for training
1 set of 10 for validation
9 sets of 100 for training
1 set of 100 for validation

Figure 3: Using three different scales to cross-validate the machine learning to ensure consistency of training results.

Each dataset contains 19 local characteristics of a geographical area. 1000 or half of the data sets are used to train the SVM, whereas the other half is used to test the classification accuracy of the SVM. The data sets are also divided into different proportions to compare the classification results.

When the training data is used in scales of CV10 (meaning ten-fold cross-validation), the training data is further divided into ten subsets, of which each subset has 100 datasets. Ninesubsetsare used to train the SVM, and the remaining subset is used to validate the accuracy of the training. This experiment is repeated by rotating a unique data set for validation. The same analogy applies to scales CV100 and CV1000. The cross-validation testing with different scales helps to verify the accuracy of SVM's classification in this research.

3.3 SVM for Classification

The World Bank's data and GA generated data are used to train the SVM. In each test, load only one of the three combination data sets.

- Dataset 1 comprises of virtual samples generated by GA only.
- Dataset 2 comprises of World Bank's data only.
- Dataset 3 comprises of virtual samples and World Bank's data.

Machine training with three different combinations of data sets is required to observe the differences in training results. This paper proposes to use the three (3) commonly used kernels in SVM, namely Linear, Polynomial, and RBF.

Finally, together with the empirical model, the kernel in the SVM that produces the highest accuracy of classification kernel will be the recommended investment model for Telco.

4. RESULTS AND DISCUSSIONS

4.1 Empirical Model

Nineteen (19) features are commonly available for 174 countries in the World Bank's public database. Each country has its corresponding 19 features a.k.a. local characteristics. The nineteen features are shown in Table 3 below.

Table 3: 19 features for 174 countries

Label	Features (also known as locality characteristics)
A	Land size (km^2)
B	Agricultural land(km^2)
C	% of population with electricity access
D	Population size
E	Population density ($perkm^2$)
F	Birth rate
G	Labour force
H	# of students enrolled in secondary schools
J	Fixed broadband per 100 person
K	Wireless broadband per 100 person
L	Telephone lines per 100 person
M	GDP (USD'mill)
N	GDP per capita (USD)

O	Economic Activity – tourism
P	Average monthly rainfall
Q	Average daily temperature
R	Length of tar road (km per million)
S	Life expectancy
T	GNI (USD'mill)
Response (R)	GNI per capita

With the input of data of the 19 features from 174 countries, the curve-fitting software has successfully formulated an equation correlating the local characteristics for a response.

$$R = f\{A, B, C, D, E, F, G, H, J, K, L, M, N, O, P, Q, R, S, T\}$$

The full empirical equation is shown in Appendix 2.

The curve-fitting software produces a correlation coefficient that quantifies the statistical relationships between the local characteristics and GNI per capita. This correlation coefficient, rated between -1 and +1, is shown in Table 4. The correlation coefficient shows if a local feature is in directional (positive coefficient) or reversed directional relationship (negative coefficient) with the GNI per capita. A higher correlation coefficient denotes a higher impact of the features to the response (GNI per capita).

Table 4: Correlation coefficient of different features on GNI per capita

Local Characteristics	**Label**	**Coefficient**
GDP per capita (USD)	N	0.981
Fixed broadband per 100 person	J	0.746
Wireless broadband per 100 person	K	0.712
Telephone lines per 100 person	L	0.665
Life expectancy	S	0.616
Length of tar road (km per million)	R	0.454
Economic Activity – tourism	O	0.373
GDP (USD'mill)	M	0.275
GNI (USD'mill)	T	0.273
% of population with electricity access	C	0.271
Population density ($perkm2$)	E	0.197
Land size ($km2$)	A	0.105
Agricultural land($km2$)	B	0.083
Average monthly rainfall	P	-0.015
Labour force	G	-0.028
Population size	D	-0.041
# of students enrolled in secondary schools	H	-0.041
Average daily temperature	Q	-0.122
Birth rate	F	-0.532

Note: For a full view of co-relationship, refer to Appendix 3.

In response to GNI per capita, it has been observed that GDP per capita has the highest-impact (0.981) directional relationship, followed by broadband penetration. The fixed telephony penetration and life expectancy have a relatively moderate co-relationship impact with ratings above 0.500. The birth rate also has a relatively moderate co-relationship impact to GNI per capita but in the reverse direction.

The characteristics of the length of tar road, tourism activity, GDP, GNI and % of the population with electricity access have some co-relationship impact, with impact ratings ranging between 0.271 and 0.454. The characteristics of the agriculture land size, population size, population density, labor force, number of secondary school students, average monthly rainfall and average daily temperature, have low co-relationship impacts (below 0.200) to GNI per capita.

Fixed broadband (0.746), wireless broadband (0.712) and telephony services (0.665) are key elements in ICT ecosystems that boost the socioeconomic status. The research results show that these features have a high impact on GNI per capita, and the results are in line with the World Bank and ITU reports which indicate that broadband penetration will accelerate economic GDP growth.

Life expectancy and birth rate are the two features of social nature with the highest impact on GNI per capita. Life expectancy has a directional impact whereas birth rate has reversed directional impact to GNI per capita.

When changing the response R from GNI per capita to fixed broadband penetration, it is observed that each feature also possess different magnitudes of impact towards the response.

Table 5: Correlation coefficient of different features on fixed broadband penetration

Local Characteristics	**Label**	**Coefficient**
Telephone lines per 100 person	L	0.881
GNI per capita	J	0.746
GDP per capita (USD)	N	0.738
Life expectancy	S	0.735
Wireless broadband per 100 person	K	0.678
Length of tar road (km per million)	R	0.499
Economic Activity – tourism	O	0.395
% of population with electricity access	C	0.347
GDP (USD'mill)	M	0.276
GNI (USD'mill)	T	0.274
Population density ($per km2$)	E	0.206
Land size ($km2$)	A	0.089
Agricultural land($km2$)	B	0.042
Labour force	G	0.004
Average monthly rainfall	P	-0.015
Population size	D	-0.016
# of students enrolled in secondary schools	H	-0.022
Average daily temperature	Q	-0.211
Birth rate	F	-0.741

Note: For a full view of co-relationship, refer to Appendix 3.

It is observed that telephony penetration has the highest-impact (0.881) directional relationship in response to fixed broadband penetration.GNI per capita (0.746), GDP per capita (0.738), life expectancy (0.735) and wireless broadband penetration (0.678) have moderately high co-relationship impacts to broadband penetration. Birthrate (-0.741) also has a moderately high impact co-relationship to fixed broadband penetration but in the reverse direction.

The length of tar road, tourism activity, % of the population with electricity access, GDP, and GNI have some co-relationship impact to broadband penetration, with the impact ratings ranging between 0.206 and 0.499.The characteristics with relatively low impacts to fixed broadband penetration are population density, land size, agriculture land size, labor force, average monthly rainfall, population size, number of secondary school students and average daily temperature.

Life expectancy and birth rate are the two features of social nature with the highest impact on fixed broadband penetration. Birthrate is the only feature that has a significant impact in the reversed direction.

The co-relationship results can serve as a guideline when applying the empirical model in a real field study. The data for local characteristics with high or medium impact, if missing, will affect the distortion of the empirical model. On the other hand, the data with low impact, if missing will have minimal effect on the empirical model.

4.2 GA for Virtual Samples

GA can generate more virtual data (2000 sets) using the fitness function formulated by the curve-fitting technique. All the available data sets are arranged into three combinations to train the SVM.

- DGA - Dataset 1 is comprised of 2000 sets of virtual samples generated by GA machine only.
- DWB - Dataset 2 is comprised of 174 sets of data obtained from the World Bank's database only.
- DGAWB - Dataset 3 is comprised of 2174 sets of data obtained from the World Bank's database and GA.

4.3 SVM for Classification

All the three different combinations of data sets are used to train the SVM. The SVM training has been validated according to the pre-defined cross-validation scales of 10, 100 and 1000. The accuracy of SVM's classification has been tested with three kernels - linear, polynomial and RBF. Table 6.1, Table 6.2, and Table 6.3 show the cross-validation accuracy (CVA) in response to different cross-validation scales of 10, 100 and 1000.

The CVA is best at 100% for training and cross-validating with DGA data. The CVA with DGAWB data is high ranging from 95.9% - 99.0% on different kernels. The CVA with DWB data is wide-ranging at 71.3% - 90.8%. It is observed that cross-validation accuracy is relatively higher when the number of training data sets is sufficient. In this case, DGA has 2000 sets, and DGAWB has 2174 sets; whereas DWB has only 174 sets of data. This pattern is seen on the SVM with linear, polynomial and RBF kernels.

Table 6.1: Cross-validation accuracy (CVA) with DGA data. Training-Testing ratio is 50-50

SVM	CV10		CV100		CV1000	
Kernel	CVA	CTA	CVA	CTA	CVA	CTA
Linear	100.0%	90.8%	100.0%	90.8%	100.0%	90.8%
Polynomial	100.0%	93.6%	100.0%	93.6%	100.0%	93.6%
RBF	100.0%	92.3%	100.0%	92.3%	100.0%	92.3%

Table 6.2: Cross-validation accuracy (CVA) with DGAWB Data. Training-Testing ratio is 50-50.

SVM	CV10		CV100		CV1000	
Kernel	CVA	CTA	CVA	CTA	CVA	CTA
Linear	98.8%	77.2%	98.4%	77.2%	98.4%	77.2%
Polynomial	95.9%	91.0%	96.0%	91.0%	96.0%	91.0%
RBF	99.0%	86.5%	99.0%	86.5%	99.0%	86.5%

Table 6.3: Cross-validation accuracy (CVA) with DWB Data. Training-Testing ratio is 50-50.

SVM	CV10		CV100		CV1000	
Kernel	CVA	CTA	CVA	CTA	CVA	CTA
Linear	90.8%	85.1%	90.8%	85.1%	90.8%	85.1%
Polynomial	71.3%	60.9%	71.3%	60.9%	71.3%	60.9%
RBF	89.7%	81.6%	89.7%	81.6%	89.7%	81.6%

The observation mentioned above is based on a 50:50 data split for training and testing. The equal split is a default setting in this experiment. When the split becomes 60:40, 70:30, 80:20 or 90:10, the same observation maintains. Nevertheless, the CVA accuracy is lower (Appendix 4 referred) than the default setting.

Table 6.1, Table 6.2, and Table 6.3 also show the SVM classification test accuracy (CTA) in response to three different data sets – DGA, DGAWB, and DWB.

The CTA is best at 93.6% for the classification test using GA data on polynomial kernels. The CTA with GA has a high range from 90.8% - 92.3%. The CTA with GAWB data is high and wide-ranging from 77.2% - 96.0% on different kernels. The CTA with WB data is relatively low and wide-ranging from 60.9% - 85.1%. It is observed that CTA is relatively higher when the number of training data sets is sufficient. This pattern is seen on SVM with linear, polynomial and RBF kernels.

The observation mentioned above is based on a 50:50 data split for training and testing. The equal split is a default setting in this experiment. When the split becomes 60:40, 70:30, 80:20 or 90:10, the observation maintains. Nevertheless, the CTA accuracy is higher (Appendix 4 referred) than the default setting, and this observation is antithetical to the findings in CVA.

5. CONCLUSION AND FUTURE RESEARCH

This paper fulfilled its objective to propose a machine learning technique to predict the socioeconomic potential of one geographical area as compared to another. The empirical model has been successfully constructed in response to the various local characteristics of a geographical area. As the land size and populations have a low co-relationship impact to GNI per capita and fixed broadband penetration, the model is, therefore, applicable to urban and rural areas regardless of its geographical location.

GA machine can generate large virtual samples to train a support vector machine. GA has produced 2000 sets of virtual samples to train the SVM.

The SVM has classified the data with high accuracy, and with consistency across linear, polynomial and RBF kernels. If the raw data could be obtained from local rural areas and with a larger sample, the empirical model can be further improved, which can subsequently improve the SVM classification accuracy as well.

The proposed machine learning technique can be a promising solution for Telco to analyze the socioeconomic potential of the thousands of rural areas in Malaysia. This technique can correlate the local characteristics in response to the socioeconomic potential of a particular geographical area. This correlation coefficient can serve as valuable information for investment beyond urban areas. Government agencies can, therefore, put a higher priority on those promising rural areas in joint-effort with Telcos to deploy broadband services. Those rural areas with higher potential indexes can be the success stories to encourage Telcos with private investments for continuous growth.

The experiment results also coincide with many other types of research, which conclude that broadband services have a positive impact on socioeconomic statuses; and the impact on socioeconomic growth varies in magnitude depending on the current economies of the geographical areas. Besides GDP per capita, the broadband penetration is found to have the highest co-relationship with GNI per capita.

Additional experiments are recommended for future research from three perspectives:

- To improve the empirical model by using larger raw data sets, especially those data with medium to high co-relationship impact.
- To verify and improve the cross-validation accuracy with a larger pool of virtual samples
- To further research on SVM kernels and to modify the kernels, if necessary to improve classification accuracy.

6. REFERENCES

[1] "Broadband: A Platform for Progress", a report by the Broadband Commission for Digital Development, International Telecommunications Union, June 2011.

[2] "11th Malaysia Plan", Malaysian Economic Planning Unit, Prime Minister's Department, 2015.

[3] Roller LH, Waverman L, "Telecommunications infrastructure and economic development: A simultaneous approach", American Economic Review, 91: 909-923, 2001.

[4] Kuppusamy, Raman & Lee, "Whose ICT Investment Matters to Economic Growth: Private or Public? The Malaysian Perspective", The Electronic Journal of Information Systems in Developing Countries, EJISDC, pp. 37, 7, 1-19, 2009.

[5] Alice SHIU, Pun-Lee LAM, "Causal Relationship between Telecommunications and Economic Growth: A Study of 105 Countries", The Hong Kong Polytechnic University, http://www.imaginar.org/taller/its2008/192.pdf

[6] C Qiang, C Rossotto, K Kimura, "Economic impacts of broadband", Information and Communications for Development: Extending Reach and Increasing Impact, pp. 35–50, 2009.

[7] N Czernich, O Falck, T Kretschmer, L Woessman, "Broadband infrastructure and economic growth", CESifo Working Paper No. 2861, 2009.

[8] Booz & Company, "Digitization for Economic Growth and Job Creation: Regional and Industry Perspectives", In The Global Information Technology Report, World Economic Forum, 2013.

[9] Johan Bergendahl, "Broadband changes society", 2010, from Ericsson website: http://www.ericsson.com/news/101027_broadband_bergendahl_244218599_c

[10] Katz, Raul L. and Koutroumpis, Pantelis, "Measuring Socio-Economic Digitization: A Paradigm Shift", TRPC, 2012. Available at SSRN: https://ssrn.com/abstract=2031531 or http://dx.doi.org/10.2139/ssrn.2031531

[11] Nottebohm, Manyika, Bughin, Chui, Syed, "Online and upcoming: The Internet's impact on aspiring countries", McKinsey & Company, 2012.

[12] R.L. Katz, T.A. Berry, Driving Demand for Broadband Networks and Services, Springer and Communication Technology, DOI: 10.1007/978-3-319-07197-8_4, 2014.

[13] Melih Kirlidog, "Financial Aspects of National ICT Strategies", In Sherif Kamel, E-Strategies for Technological Diffusion and Adoption: National ICT Approaches for Socioeconomic Development, p. 279, Information Science Reference, USA, 2010.

[14] Tony Irawan, "ICT and Economic Development: Conclusion from IO Analysis for Selected ASEAN Member States", EUROPEAN INSTITUTE FOR INTERNATIONAL ECONOMIC RELATIONS, 2013.

[15] Tony Irawan, "ICT and Economic Development: Conclusion from IO Analysis for Selected ASEAN Member States", EUROPEAN INSTITUTE FOR INTERNATIONAL ECONOMIC RELATIONS, 2013.

[16] Mudiarasan Kuppusamy, Bala Shanmugam, "Information-Communication Technology and Economic Growth in Malaysia, International Association for Islamic Economics", Review of Islamic Economics, Vol. 11, No. 2, pp. 87-100, 2007.

[17] L Waverman, M Meschi, M Fuss, "The impact of telecoms on economic growth in developing countries", In Africa: The impact of mobile phones, The Vodafone Policy Paper Series, Number 2, pp. 10-23, 2005.

[18] C Chakraborty, B Nandi, "Privatization, telecommunications and growth in selected Asian countries: An econometric analysis", Communications and Strategies 52: 31-47, 2003.

[19] Peter Stenberg, Mitchell Morehart, "Toward Understanding U.S. Rural-Urban Differences in Broadband Internet Adoption and Use", In Yogesh Dwivedi, Adoption, Usage, and Global Impact of Broadband Technologies, p. 157. Information Science Reference, USA, 2011.

[20] Karner J., Onyeji R., "Telecom private investment and economic growth: The case of African and Central & East European countries", Unpublished thesis, Jönköping University, Jönköping International Business School, JIBS, Economics, 2007.

[21] Chatchai Kongaut, Ibrahim Kholilul Rohman, Erik Bohlin, "The economic impact of broadband speed: Comparing between higher and lower income countries", In Research project between the European Investment Bank (EIB) and the Institute for Management of Innovation and Technology (IMIT), Gothenburg, Sweden, 2012.

[22] Karim Sabbagh, Bahjat El-Darwiche, Roman Friedrich, Milind Singh, “Maximizing the impact of digitization”, Strategy&, PwC, 2012.

[23] Melih Kirlidog, Stephen Little, “Regional – National ICT Strategies”, In Sherif Kamel, E-Strategies for Technological Diffusion and Adoption: National ICT Approaches for Socioeconomic Development, p. 66. Information Science Reference, USA, 2010.

[24] Background paper by the UNCTAD secretariat, Expert Meeting in Support of the Implementation and Follow-up of WSIS: USING ICTs TO ACHIEVE GROWTH AND DEVELOPMENT, United Nations Conference on Trade and Development, Geneva, 4-5 December 2006.

[25] Peter Curwen, Jason Whalley, Telecommunications in a High Speed World – Industry structure, Strategic Behavior and Socio-Economic Impact, UK Gower Publishing, p211, 2010.

[26] Cronin F.J., Parker E.B., Colleran E.K., Gold M.A., “Telecommunications infrastructure and economic growth: An analysis of causality”, Telecommunications Policy 15, 529-535, 1991.

[27] Cronin F.J., Parker E.B., Colleran E.K., Gold M.A., “Telecommunications infrastructure investment and economic development”, Telecommunications Policy 17, 415-430, 1993a.

[28] Cronin F.J., Colleran E.K., Herbert P.L., Lewitzky S., “Telecommunications and growth: The contribution of telecommunications infrastructure investment to aggregate and sectoral productivity”, Telecommunications Policy 17, 677-690, 1993b.

[29] Wolde-Rufael Y., “Another look at the relationship between telecommunications investment and economic activity in the United States”, International Economic Journal 21, 199-205, 2007.

[30] R. Baker, G. Siemens, Educational data mining and learning analytics, Cambridge University Press, UK, 2014.

[31] Sandra Lach Arlinghaus, PHB Practical Handbook of Curve Fitting. CRC Press, 1994.

[32] F. Hu, Q. Hao, Intelligent Sensor Networks: The Integration of Sensor Networks, Signal Processing and Machine Learning, CRC Press, 2013.

[33] Der-Chiang Li, Wen, I-Hsiang Wen, “A genetic algorithm-based virtual sample generation technique to improve small data set learning”, Neurocomputing. 143. pp. 222–230, 2014.

[34] S. Haylin, Neural Networks A Comprehensive Foundation, Prentice Hall International, New Jersey, 1999.

[35] X. Li, D. Lord, Y. Zhang, Y. Xie, "Predicting motor vehicle crashes using Support Vector Machine models”, Accident Analysis & Prevention, pp. 1611-1618, 2008.

[36] N. Cristianini, J. Shawe-Taylor, An Introduction to Support Vector Machines and Other Kernel-based Learning Methods, Cambridge University Press, 2000.

[37] Heng Fui Liau, Dino Isa, “Feature selection for support vector machine-based face-iris multimodal biometric system”, Expert Systems with Applications, 2011.

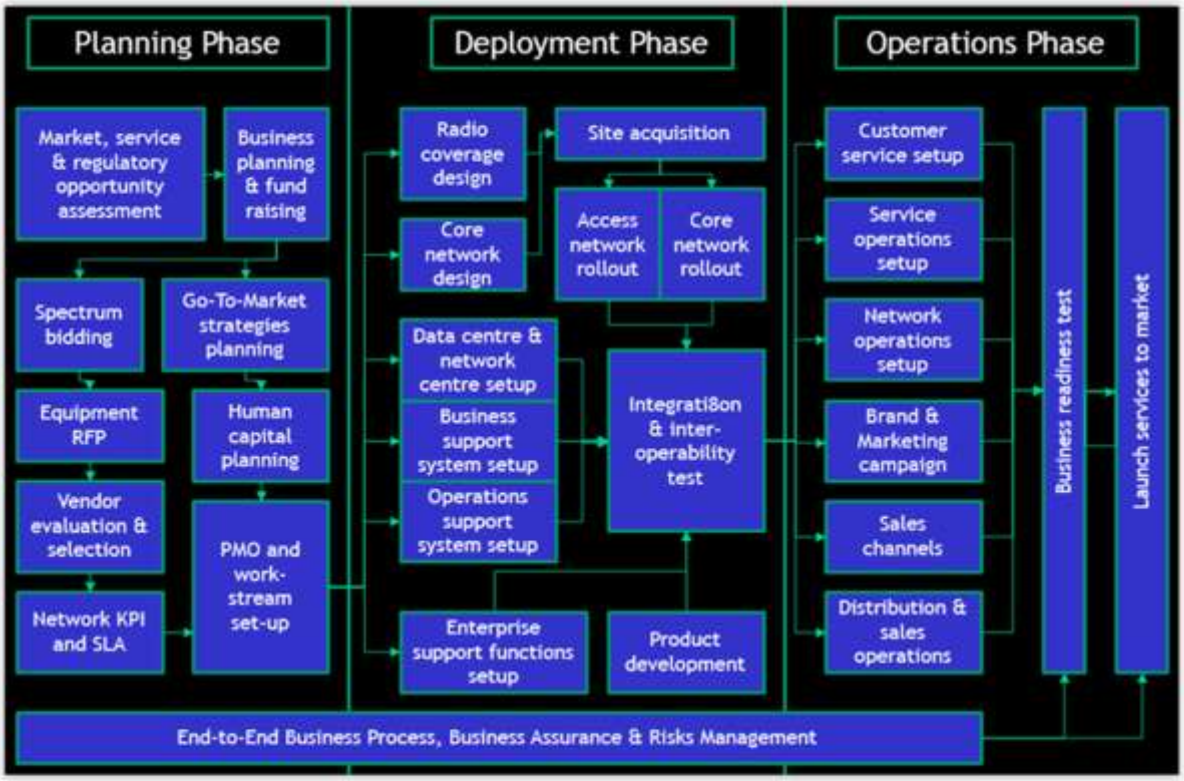
[38] Cheng et al., “Heuristic Methods for Reservoir Monthly Inflow Forecasting: A Case Study of Xinfengjiang Reservoir in Pearl River China”, MDPI AG, Switzerland, 2015.

[39] Kugan, “DiGi LTE Ready Tomorrow but Maxis & Celcom LTE ready now”, from Malaysian Wireless website: https://www.malaysianwireless.com/2013/01/digi-lte-ready-maxis-celcom/, 2013.

[40] P Prem Kumar, “Green Packet to reveal P1 buyer this month”, from Free Malaysian Today website: https://www.freemalaysiatoday.com/category/business/2014/02/13/green-packet-to-reveal-p1-buyer-this-month/, 2014.

[41] Soyacincao, “Yes to roll out 4G WiMAX in East Malaysia”, from Soyacincao website: http://www.soyacincau.com/tag/yes-coverage/, 2011.

APPENDIX 1 – TELCO OPERATING MODEL (SOURCE: AUTHOR)


Planning Phase
Deployment Phase
Operations Phase
Market, service & regulatory opportunity assessment
Business planning & fund raising
Spectrum bidding
Go-To-Market strategies planning
Equipment RFP
Human capital planning
Vendor evaluation & selection
PMO and work-stream set-up
Network KPI and SLA
Radio coverage design
Site acquisition
Core network design
Access network rollout
Core network rollout
Data centre & network centre setup
Business support system setup
Operations support system setup
Integrati8on & inter-operability test
Enterprise support functions setup
Product development
Customer service setup
Service operations setup
Network operations setup
Brand & Marketing campaign
Sales channels
Distribution & sales operations
Business readiness test
Launch services to market
End-to-End Business Process, Business Assurance & Risks Management

APPENDIX 2 – EMPIRICAL EQUATION COMPUTED BY CURVE-FITTING SOFTWARE

```
R =
- 53003.12 + 19482.23*A + 1208.34*B + 45242.04*C - 55108.44*D + 12152.05*E + 3949.75*F
+ 39626.27*G - 13441.39*H + 3126.98*J + 1750.12*K - 3815.07*L + 74879.88*M - 38351.13*N +
7755.48*O + 262.06*P + 2777.10*Q - 15230.85*R + 1393.57*S
- 1.060E+005*T - 99.46*A*B - 513.78*A*C - 98315.78*A*D + 1240.30*A*F + 86901.40*A*G
+ 37148.31*A*H - 890.18*A*J - 604.35*A*K+ 1689.63*A*L + 1.020E+005*A*M
+ 975.96*A*N + 850.02*A*O - 417.25*A*P - 367.68*A*Q- 282.38*A*R + 8.51*A*S
- 1.105E+005*A*T + 174.22*B*C + 75682.74*B*D + 22975.17*B*E- 1413.09*B*F
- 63971.80*B*G - 34206.30*B*H + 1233.66*B*J + 349.68*B*K - 143.14*B*L
- 1.436E+005*B*M - 275.10*B*N - 4360.30*B*O+ 253.71*B*P + 171.00*B*Q - 573.14*B*R
- 786.52*B*S + 1.485E+005*B*T - 5153.16*C*D - 776.99*C*E + 57.80*C*F + 4920.95*C*G
- 75.27*C*H - 123.31*C*J + 34.59*C*K + 284.86*C*L - 2.324E+005*C*M - 725.62*C*N
- 1558.03*C*O + 9.03*C*P - 41.31*C*Q + 242.00*C*R - 44.23*C*S+ 2.810E+005*C*T
+ 35180.49*D*E + 717.81*D*F - 1529.94*D*G + 16947.67*D*H - 21093.04*D*J
+ 9622.49*D*K + 9422.54*D*L + 7.653E+005*D*M- 35976.62*D*N - 4818.83*D*O
+ 651.88*D*P - 139.97*D*Q + 45906.81*D*R + 7668.78*D*S- 8.509E+005*D*T
+ 314.88*E*F -13259.68*E*G - 25836.41*E*H - 1195.76*E*J + 331.20*E*K
+ 623.94*E*L + 33117.50*E*M + 414.35*E*N - 908.40*E*O - 91.08*E*P + 241.31*E*Q
+ 1846.76*E*R + 736.70*E*S - 41681.14*E*T - 1673.17*F*G + 1690.65*F*H + 1.55*F*J
- 88.10*F*K - 21.11*F*L - 599.39*F*M - 288.33*F*N+ 1787.45*F*O - 18.27*F*P - 15.55*F*Q
+ 144.20*F*R - 34.46*F*S + 2204.01*F*T - 22327.62*G*H + 11790.30*G*J - 6366.37*G*K
- 13596.52*G*L - 5.010E+005*G*M + 17029.84*G*N + 17720.81*G*O - 1346.58*G*P
- 1142.60*G*Q - 52342.23*G*R - 5466.25*G*S + 5.805E+005*G*T+ 8598.65*H*J
- 2690.13*H*K + 1806.51*H*L - 3.975E+005*H*M - 19138.76*H*N + 1483.43*H*O
+ 2265.16*H*P + 1397.58*H*Q - 4327.59*H*R- 1994.54*H*S + 4.241E+005*H*T - 62.50*J*K
- 6.44*J*L + 1557.38*J*M - 121.85*J*N + 1731.15*J*O - 13.88*J*P - 17.04*J*Q - 45.81*J*R
- 19.30*J*S + 1441.49*J*T + 20.49*K*L - 13562.16*K*M - 44.02*K*N - 599.96*K*O
- 11.91*K*P - 4.97*K*Q + 203.80*K*R + 9.57*K*S+ 15279.57*K*T + 23107.61*L*M
+ 0.35*L*N + 331.33*L*O + 13.65*L*P - 20.20*L*Q - 156.75*L*R - 18.07*L*S
- 26613.10*L*T - 2482.64*M*N - 5666.71*M*O - 1041.78*M*P + 2698.99*M*Q
- 5696.09*M*R + 1273.24*M*S + 6044.75*M*T+ 239.31*N*O + 17.85*N*P - 73.18*N*Q
+ 414.67*N*R - 43.23*N*S
```

APPENDIX 3 – CORRELATOIN COEFFICIENT OF FEATURES (LOCAL CHARACTERISTICS) TO GNI PER CAPITA

Features		Land Size	Agro Land	Electric Access	Pop Size	Pops Dens	Birth Rate	Labor	Second School	Fixed BB	Wireless BB	Phone	GDP	GDP /cap	Tourism	Rain	Temp	Road	Life	GNI	GNI/cap
	Label	**A**	**B**	**C**	**D**	**E**	**F**	**G**	**H**	**J**	**K**	**L**	**M**	**N**	**O**	**P**	**Q**	**R**	**S**	**T**	**Respond**
Land size ($km2$)	**A**		0.772	0.068	0.433	-0.084	-0.086	0.459	0.415	0.089	0.094	0.109	0.544	0.082	0.378	-0.054	0.218	0.152	0.075	0.537	0.105
Agricultural land($km2$)	**B**			0.053	0.631	-0.08	-0.042	0.666	0.597	0.042	0.046	0.06	0.687	0.062	0.468	-0.061	0.242	0.102	0.028	0.679	0.083
% of population with electricity access	**C**				0.077	0.068	-0.573	0.077	0.086	0.347	0.279	0.374	0.11	0.266	0.146	-0.068	-0.031	0.202	0.532	0.11	0.271
Population size	**D**					-0.006	-0.061	0.981	0.975	-0.016	-0.047	-0.023	0.549	-0.048	0.327	-0.009	0.109	-0.062	0.013	0.541	-0.041
Population density ($perkm2$)	**E**						-0.173	-0.007	-0.006	0.206	0.291	0.242	-0.011	0.192	0.115	0.175	-0.171	-0.121	0.202	-0.012	0.197
Birth rate	**F**							-0.082	-0.071	-0.741	-0.566	-0.748	-0.212	-0.524	-0.304	0.034	0.09	-0.365	-0.862	-0.209	-0.532
Labour force	**G**								0.923	0.004	-0.037	-0.004	0.591	-0.035	0.343	-0.005	0.099	-0.053	0.034	0.583	-0.028
Number of secondary school students enrolled	**H**									-0.022	-0.051	-0.026	0.506	-0.047	0.319	-0.002	0.129	-0.056	0.023	0.499	-0.041
Fixed broadband per 100 people	**J**										0.678	0.881	0.276	0.738	0.395	-0.015	-0.211	0.499	0.735	0.274	0.746
Wireless broadband per 100 people	**K**											0.617	0.226	0.69	0.351	-0.062	-0.104	0.332	0.597	0.224	0.712
Telephone lines per 100 people	**L**												0.282	0.664	0.401	-0.022	-0.173	0.408	0.738	0.279	0.665
GDP (USD’mill)	**M**													0.25	0.877	0.022	0.045	0.123	0.214	0.999	0.275
GDP per capita (USD)	**N**														0.35	-0.005	-0.138	0.436	0.606	0.247	0.981
Economic Activity – tourism	**O**															0.014	-0.017	0.142	0.317	0.877	0.373
Average monthly rainfall	**P**																-0.186	-0.102	-0.047	0.02	-0.015
Average daily temperature	**Q**																	0.026	-0.117	0.04	-0.122
Length of tar road (km per million)	**R**																		0.359	0.122	0.454
Life expectancy	**S**																			0.21	0.616
GNI (USD’mill)	**T**																				0.273

APPENDIX 4

– SVM TRAINING AND TESTING RESULTS

NOTES:

- CV10, CV100 and CV1000 are the cross validation scales used to train the SVM and cross validate the training accuracy. For example: CV10 means 10-folds CV and hold one on CV.
- CVA denotes cross validation accuracy. CTA denotes classification testing accuracy.

Table A: GA Data. Training-Testing ratio is 50-50

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	CTA
Linear	100.0%	90.8%	100.0%	90.8%	100.0%	90.8%
Polynomial	100.0%	93.6%	100.0%	93.6%	100.0%	93.6%
RBF	100.0%	92.3%	100.0%	92.3%	100.0%	92.3%

Table B: GAWB Data. Training-Testing ratio is 50-50

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	CTA
Linear	98.8%	77.2%	98.4%	77.2%	98.4%	77.2%
Polynomial	95.9%	91.0%	96.0%	91.0%	96.0%	91.0%
RBF	99.0%	86.5%	99.0%	86.5%	99.0%	86.5%

Table C: WB Data. Training-Testing ratio is 50-50

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	CTA
Linear	90.8%	85.1%	90.8%	85.1%	90.8%	85.1%
Polynomial	71.3%	60.9%	71.3%	60.9%	71.3%	60.9%
RBF	89.7%	81.6%	89.7%	81.6%	89.7%	81.6%

Table D: GA Data. Training-Testing ratio is 60-40

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	CTA
Linear	98.6%	95.3%	98.6%	95.3%	98.6%	95.3%
Polynomial	98.8%	92.6%	98.8%	92.6%	98.8%	92.6%
RBF	98.8%	97.4%	99.0%	97.4%	99.0%	97.4%

Table E: GAWB Data. Training-Testing ratio is 60-40

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	CTA
Linear	96.4%	93.8%	95.9%	93.8%	95.7%	93.8%
Polynomial	95.6%	91.5%	95.4%	91.5%	95.4%	91.5%
RBF	97.7%	95.6%	97.8%	95.6%	97.6%	95.6%

Table F: WB Data. Training-Testing ratio is 60-40

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	CTA
Linear	94.2%	87.1%	94.2%	87.1%	94.2%	87.1%
Polynomial	70.2%	67.1%	70.2%	67.1%	70.2%	67.1%
RBF	89.4%	82.9%	87.5%	82.9%	87.5%	82.9%

Table G: GA Data. Training-Testing ratio is 70-30

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	CTA
Linear	99.1%	93.7%	98.9%	93.7%	98.8%	93.7%
Polynomial	98.9%	90.2%	99.0%	90.2%	99.0%	90.2%
RBF	98.9%	96.7%	98.9%	96.7%	98.9%	96.7%

Table H: GAWB Data. Training-Testing ratio is 70-30

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	CTA
Linear	96.3%	92.9%	96.1%	92.9%	96.0%	92.9%
Polynomial	95.5%	89.4%	95.7%	89.4%	95.7%	89.4%
RBF	97.8%	95.6%	98.0%	95.6%	98.0%	95.6%

Table I: WB Data. Training-Testing ratio is 70-30

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	CTA
Linear	91.0%	86.5%	93.4%	86.5%	93.4%	86.5%
Polynomial	70.5%	75.0%	69.7%	75.0%	69.7%	75.0%
RBF	86.9%	82.7%	87.7%	82.7%	87.7%	82.7%

Table J: GA Data. Training-Testing ratio is 80-20

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	CTA
Linear	99.1%	96.4%	98.9%	96.4%	98.9%	96.4%
Polynomial	99.1%	96.4%	99.1%	96.4%	99.1%	96.4%
RBF	98.7%	96.3%	98.8%	96.3%	98.7%	96.3%

Table K: GAWB Data. Training-Testing ratio is 80-20

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	CTA
Linear	96.4%	95.8%	96.4%	95.8%	96.3%	95.8%
Polynomial	95.7%	94.6%	95.8%	94.6%	95.8%	94.6%
RBF	97.5%	96.0%	97.7%	96.0%	97.8%	96.0%

Table L: WB Data. Training-Testing ratio is 80-20

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	CTA
Linear	90.6%	91.4%	89.9%	91.4%	89.9%	91.4%
Polynomial	74.1%	82.9%	74.1%	82.9%	73.4%	82.9%
RBF	86.3%	82.9%	86.3%	82.9%	86.3%	82.9%

Table M: GA Data. Training-Testing ratio is 90-10

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	CTA
Linear	97.8%	100.0%	97.8%	100.0%	97.8%	100.0%
Polynomial	97.7%	100.0%	98.2%	100.0%	98.2%	100.0%
RBF	97.9%	100.0%	98.1%	100.0%	98.0%	100.0%

Table N: GAWB Data. Training-Testing ratio is 90-10

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	TA
Linear	95.2%	99.5%	95.2%	99.5%	95.2%	99.5%
Polynomial	94.8%	97.2%	94.7%	97.2%	94.8%	97.2%
RBF	97.0%	99.1%	97.1%	99.1%	97.0%	99.1%

Table O: WB Data. Training-Testing ratio is 90-10

SVM Kernel	CV10		CV100		CV1000	
	CVA	CTA	CVA	CTA	CVA	CTA
Linear	89.2%	94.1%	89.8%	94.1%	89.2%	94.1%
Polynomial	78.3%	88.2%	78.3%	88.2%	78.3%	88.2%
RBF	86.6%	88.2%	86.0%	88.2%	86.0%	88.2%