Faculty of Science and Engineering

Department of Electrical and Electronic Engineering

USING MACHINE LEARNING TECHNIQUE TO CLASSIFY GEOGRAPHIC AREAS WITH SOCIOECONOMIC POTENTIAL FOR BROADBAND INVESTMENT IN MALAYSIA

TAN TAIK GUAN

THESIS SUBMITTED TO THE UNIVERSITY OF NOTTINGHAM
FOR THE DEGREE OF MASTER OF PHILOSOPHY

DECEMBER 2018

LIST OF PUBLICATIONS

Journals:

1. *Using Machine Learning Technique for Telecom Service Providers in Malaysia to Prioritize Broadband Investment in Urban and Rural Areas*, Asian Journal of Engineering and Technology (ISSN: 2321 – 2462) Volume 05 – Issue 06, December 2017. [Published]

Conference Paper

2. *Using Machine Learning Technique to Classify Geographic Areas with Socioeconomic Potential for Broadband Investment*, ISI World Statistics Congress 2019, 18th – 23rd August 2019, Kuala Lumpur Convention Center, Malaysia. [Submitted 5th November 2018]

ABSTRACT

The telecommunication companies (TELCO) in Malaysia commonly use the return on investment (ROI) model for techno-economic analysis to strategize their network investment plan in their intended markets. The number of subscribers and average revenue per user (ARPU) are two dominant contributions to a good ROI. Rural areas are lacking in both dominant factors and thus very often fall outside the radar of 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, regulated policies and funding for universal service provision. This thesis outlines a framework of machine learning technique which the TELCOs and government agencies can use to plan for broadband investments in Malaysia, especially for rural areas. The framework is implemented in four stages: data collection, machine learning, machine testing, and machine application. In this framework, a curve-fitting technique will be applied to formulate an empirical model by using prototyping data from the World Bank databank. The empirical model serves as a fitness function for a genetic algorithm (GA) to generate large virtual samples to train, validate and test the support vector machines (SVM). Real-life field data for geographic areas in Malaysia are then provided to the tested SVM to predict which areas have the socioeconomic potential for broadband investment. By using this technique as a policy tool, TELCOs and government agencies 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.

Keywords— Curve Fitting, Genetic Algorithm, Support Vector Machine, Broadband Investment.

ACKNOWLEDGMENT

During a casual meeting with Professor Dino Isa a couple of years ago, he shared his concern on how a rural Indian boy suffered his death due to poverty. The discussion with Professor Dino has launched a whole new chapter in my life. After working for more than ten years in the telecommunication industry, I decided to study the socioeconomic issues in rural areas with relevance to the telecom industry. Through my working experiences in the telecom industry, I have observed the limitations and dilemmas for TELCOs to invest in rural areas. I am thankful to have Professor Dino as my supervisor. I owe my gratitude to Professor Dino, who has helped me begin a new journey in my life as a MPhil candidate at the University of Nottingham (Malaysia Campus). Professor Dino has been generous with his advice on scholarly research, professional in the subject of system intelligence, and patient in research project management.

Dr. Yee Wan Wong is my second supervisor. She has always been sharp and precise with her input on machine learning techniques such as genetic algorithm and statistical analysis, which has helped me overcome some of the key challenges in this research. Dr. Wong has also been giving invaluable comments that helped to improve the quality and accuracy of my journal paper and thesis. I am indebted to her for her time and patience to enlighten me in these issues.

I would like to extend my appreciation to Jasmine Chen Pei Yi and Mei Shin Oh who have been my mental sparring partners throughout this research. A series of question-and-answer sessions with them have enlightened me to improve the clarity and precision of my research.

I am grateful that a few professionals in the telecom industry have agreed to participate in an interview, helping me to verify the common practices in the TELCO industry. I am particularly grateful to the following experienced professionals who have contributed to this research:

1. Ir. Haji Dr. Ismail Haron, a retiree, formerly served in a senior management position with both wired and wireless telecommunication companies, as well as government agencies.

2. Mr. Jun Jack Yong, currently serving as the director of network planning with one of the TELCOs in Malaysia,
3. Mr. Navin Gopal, an accomplished leader in the area of business strategy, business investment, business development and operations within the media and telecommunications industry.

This research used data provided by two institutes – the World Bank and the Department of Statistics Malaysia (DOSM). Some of the data are available online, while others were provided offline by DOSM upon request. I am thankful to the World Bank and DOSM for their generosity in sharing the data which took a tantamount of time, processes and effort to collect.

Throughout the research period, balancing my family life, office work and academic research were more challenging than I had expected. I am indebted to my parents, my wife, and my three children for their tolerance, motivation, and love in supporting my research.

Above all, I praise the Almighty God for blessing me with strength and knowledge throughout this study period.

TABLE OF CONTENTS

LIST OF PUBLICATIONS II

ABSTRACT III

ACKNOWLEDGMENT IV

TABLE OF CONTENTS VI

LIST OF FIGURES IX

LIST OF TABLES XI

LIST OF BOXES XII

LIST OF ABBREVIATIONS XIII

LIST OF NOTATIONS XV

1 INTRODUCTION 1

1.1 RESEARCH BACKGROUND 5
1.1.1 CURRENT STATE OF BROADBAND INVESTMENT IN THE PRIVATE SECTOR 5
1.1.2 CURRENT STATE OF BROADBAND INVESTMENT IN THE PUBLIC SECTOR 7
1.1.3 CURRENT STATE OF MACHINE LEARNING TECHNIQUE FOR BROADBAND PLANNING 10
1.2 RESEARCH ISSUES 13
1.2.1 GEOGRAPHICAL FEATURES AND SOCIOECONOMIC POTENTIAL 14
1.2.2 DATA REQUIRED 16
1.2.3 TRAINING THE MACHINE 17
1.2.4 TESTING THE MACHINE 17
1.2.5 RESEARCH SYSTEM OVERVIEW 18
1.3 RESEARCH OBJECTIVES AND CONTRIBUTIONS 18
1.3.1 RESEARCH OBJECTIVES 18
1.3.2 CONTRIBUTIONS OF THIS RESEARCH 19
1.4 THESIS ORGANIZATION 24

2 LITERATURE REVIEW 25

2.1 MACHINE LEARNING TECHNIQUE 25
2.2 GENETIC ALGORITHM 28
2.3 SUPPORT VECTOR MACHINE (SVM) 29
2.3.1 PATTERN LEARNING AND RECOGNITION 31

2.3.2 OBJECT, RESPONSE, AND FEATURES 32
2.3.3 FEATURE SPACE 33
2.3.4 DATA AND DATA SETS 35
2.3.5 BINARY AND MULTICLASS CLASSIFICATIONS 36
2.3.6 LINEARLY SEPARABLE VS. NON-LINEARLY SEPARABLE SPACE 37
2.3.7 HYPERPLANE 39
2.3.8 RISK MINIMIZATION AND VC-DIMENSION 45
2.3.9 OVERFITTING 48
2.3.10 ALGORITHM AND KERNEL 49
2.3.11 TRAINING AND TESTING 54
2.3.12 DISADVANTAGES OF SVM 55
2.4 URBAN AND RURAL AREAS 56
2.4.1 DEFINITION OF URBAN AND RURAL AREAS 56
2.4.2 CLASSIFICATION OF AREAS BY ECONOMIC SITUATION 59
2.4.3 PERFORMANCE INDICATORS OF SOCIOECONOMIC AND BROADBAND 60
2.4.4 THE SOCIOECONOMIC IMPACT OF BROADBAND INVESTMENT 61
2.5 BROADBAND 67
2.5.1 THE IMPORTANCE OF BROADBAND 70
2.5.2 THE DIFFERENT STAGES OF BROADBAND DEVELOPMENT 74
2.5.3 FEATURES THAT INFLUENCE BROADBAND DEVELOPMENT 81
2.5.4 PUBLIC POLICIES THAT INFLUENCE BROADBAND DEVELOPMENT 88
2.5.5 BROADBAND INVESTMENT 90
2.6 TECHNO-ECONOMIC ANALYSIS FOR BROADBAND INVESTMENT 98
2.6.1 ROI AS TECHNO-ECONOMIC ANALYSIS FOR BROADBAND INVESTMENT 104
2.6.2 COMPARING ROI TO SVM 110

3 METHODOLOGY 112

3.1 DATA COLLECTION 113
3.2 MACHINE LEARNING 118
3.3 MACHINE TESTING 120
3.4 MACHINE APPLICATION 120

4 RESULTS AND OBSERVATIONS 121

4.1 DATA COLLECTION 121
4.2 MACHINE LEARNING AND TESTING 131
4.3 MACHINE APPLICATION 134
4.3.1 APPLICATION OF REAL-LIFE FIELD DATA FROM 13 STATES IN MALAYSIA 134
4.3.2 APPLICATION OF REAL-LIFE FIELD DATA FOR DISTRICT AREAS IN MALAYSIA 141
4.4 OBSERVATION OF SVM'S BEHAVIORS 142
4.4.1 MACHINE LEARNING APPEARED IN THIS RESEARCH 142
4.4.2 DATA INFLUENCE AND FEATURE SELECTION 144
4.4.3 SVM KERNELS 145
4.4.4 CORRELATION COEFFICIENCY IS CONSISTENT WITH PAST RESEARCH 146

5 CONCLUSION 148

5.1 SIGNIFICANCE OF RESEARCH RESULTS 148
5.2 RELEVANCE OF RESEARCH RESULTS TO THE STAGES OF BROADBAND 150
5.3 RECOMMENDATION FOR FUTURE RESEARCH 151

APPENDIX 1 – TELCO OPERATING MODEL 153

APPENDIX 2 – EMPIRICAL EQUATION COMPUTED BY CURVE-FITTING SOFTWARE (QUADRATIC MODEL WITH WORLD BANK DATA) 154

APPENDIX 3 – EMPIRICAL EQUATION COMPUTED BY CURVE-FITTING SOFTWARE (QUADRATIC MODEL IN MATLAB FORMAT WITH WORLD BANK DATA) 155

APPENDIX 4 – EMPIRICAL EQUATION COMPUTED BY CURVE-FITTING SOFTWARE (QUADRATIC MODEL WITH MALAYSIAN STATE DATA) 156

APPENDIX 5 – MATRIX OF FEATURES AVAILABLE FOR MACHINE LEARNING 157

APPENDIX 6 – CORRELATION COEFFICIENT OF 19 GEOGRAPHICAL FEATURES TO GNI PER CAPITA 158

APPENDIX 7 – CORRELATION COEFFICIENT OF 10 GEOGRAPHICAL FEATURES TO GNI PER CAPITA 159

APPENDIX 8 – CORRELATION COEFFICIENT OF 19 SELECTED FEATURES TO GNI PER CAPITA (1/5) 160

APPENDIX 9 – CORRELATION COEFFICIENT OF 10 SELECTED FEATURES TO GNI PER CAPITA (1/3) 165

APPENDIX 10 – SVM TRAINING AND TESTING RESULTS 168

APPENDIX 11 – TEMPLATE FOR ROI SIMULATION 169

APPENDIX 12 – EXAMPLES OF HYBRID KERNELS 170

REFERENCE 171

LIST OF FIGURES

Figure 1: BB Investment Improves Economies. Source: ZTE 1
Figure 2: SVM maps data in input space to feature space, and separates the data 4
Figure 3: Definition of Digital Divide in Malaysia 9
Figure 4: Next-Generation-Network Financing Model[45] 10
Figure 5: PESTEL Framework 12
Figure 6: Implementation of Support Vector Machine. Source: the author 18
Figure 7: Desired Outcome of USP Symposium 2014, Group 3 Study. Source: MCMC 20
Figure 8: Percentage of Households with Internet Access. Source: ITU 23
Figure 9: Machine Learning with various learning algorithms. Source:[56] 27
Figure 10: GA Work Flow. Source:[57] 28
Figure 11: Linear support vector machine example (modified from Burges, 1998) 30
Figure 12: Model for Statistical Pattern Recognition System. Source:[65] 31
Figure 13: 2-Class Classification of Districts with Single Feature 32
Figure 14: 2-Class Classification of Districts with Dual Features 33
Figure 15: Dilemma in presenting 3 features in a 2-dimensional space 34
Figure 16: A hyperplane separates the districts that meet the observation from others 34
Figure 17: Multiclass Classification of Districts 37
Figure 18: Example of transforming the data nonlinearly. Source:[66] 38
Figure 19: Nonlinear classifier transforms input space into feature space. Source:[67] 39
Figure 20: SVM with maximum-margin separating hyperplane. Source:[68] 40
Figure 21: Linear classifier with hyperplanes (H) and support vectors (v) 41
Figure 22: Structural Risk Minimization. Source:[70] 46
Figure 23: (a) 3-points data that are separable, and (b) 4-points data that are not separable. Source:[71] 47
Figure 24: Over-fitting dilemma. Source:[72] 48
Figure 25: Execution of Kernel-Based Learning Algorithm. Source:[66] 50
Figure 26: Cross-Validation and Cross Testing. Source:[78] 54
Figure 27: The Case of a 5-Fold Cross-Validation with 30 Samples. Source:[79] 55
Figure 28: Underserved Areas Defined for USP Targets. Source: MCMC 58
Figure 29: Digitization impact on the socioeconomic. Source:[21] 62
Figure 30: Growth effects of telecommunications. Source:[6] 63
Figure 31: Telecommunication technologies for BS. Source: the author 68
Figure 32: Bandwidth increments since the 1980s. Source: ZTE 68
Figure 33: Wireless Technologies and Data Rates. Source:[89] 69
Figure 34: National BB Strategies in Selected Countries. Source: ZTE 70
Figure 35: Malaysia National Broadband Implementation Strategy for NBI. Source: MCMC 71
Figure 36: Stages of broadband development of a geographic area 74
Figure 37: Non-Internet users face four categories of barriers. Source:[86] 75
Figure 38: Comparison of BB Affordability by Region 79
Figure 39: Direct Public Policies to Stimulate BB Penetration. Source:[124] 89
Figure 40: Broadband Boosting ICT Applications Development. Source: ZTE 91
Figure 41: Construction Models of BB Network for Government. Source: ZTE 91
Figure 42: Four Geographic Zones for BB Investment. Source:[125] 92
Figure 43: % Rural Populations in Malaysia. Source:[135] 94
Figure 44: Household BB Penetration. Source: MCMC 94

Figure 45: Overview of Architecture for BB Infrastructure. Source: ZTE 95
Figure 46: Understanding the Digital Divide Gap. Source:[138] 96
Figure 47: BB Supply Gap and Demand Gap Outlined by Katz and Berry. Source:[12] 97
Figure 48: 3-Step TEA for Telecommunication Services. Source:[141] 100
Figure 49: Techno-Economic Analysis Model for Telecom Industry. Source:[143] 102
Figure 50: Top-Down vs. Bottom-Up Cost Modelling. Source:[145] 103
Figure 51: Example of LTE Network Architecture. Source: Packet One Networks 105
Figure 52: Data must travel through several networks to reach an end customer[86] 106
Figure 53: Average returns on investment for 78 network operators 108
Figure 54: Comparing BB prices between Malaysia and other ASEAN countries 109
Figure 55: Research Methodology with 4-Stage Implementation. Source: the author. 113
Figure 56: Using three different scales to cross-validate machine learning 119
Figure 57: Correlation between a single variable factor and response 124
Figure 58: Fitness Values vs. Generalization 130
Figure 59: Machine Learning Diagram 143
Figure 60: Delta of Bias and Variance are getting smaller as N increases. Source:[66] 144

LIST OF TABLES

Table 1: Network Coverage by the Malaysian TELCOs 7
Table 2: Example of Geographical Features Tabulated in PESTEL Model 11
Table 3: Example of data, data sets, feature and feature sets 36
Table 4: True and False Classification 56
Table 5: Three Layers of Rural Areas as Target Communities for USP Projects 58
Table 6: Analogy of worldview and country-view of economic states 60
Table 7: Example of KPI to measure socioeconomic and digitalization 61
Table 8: Different Motivations of Having National BB Policy 73
Table 9: Stages of broadband adoption. Source:[45] 79
Table 10: Features that affect ICT development 82
Table 11: Past Research on Determinants of BB Demand 84
Table 12: Past Research on Correlation between ICT and Socioeconomic Impact 85
Table 13: Financial Performance Indicators for Project Management. 88
Table 14: 4-Stage Techno-Economic Analysis for Telecommunication Service 100
Table 15: Comparison of SVM vs. ROI 111
Table 16: Geographical Feature Sets Defined for Data Collection 114
Table 17: Feature Sets with Data Available from World Bank 121
Table 18: World Bank GNI per capita as Socioeconomic Response 122
Table 19: Partial real-life field data for the 19 features and 1 response by country. 123
Table 20: Magnitude of Correlation Coefficient and Its Significance of Linear Correlation 125
Table 21: Correlation Coefficient of Geographical Features to Socioeconomic Response 125
Table 22: Correlation between geographical features and fixed BB penetration 127
Table 23: World Bank and GA Data are Arranged into 3 Groups for Machine Learning 131
Table 24: Training and Cross-Validation Results 132
Table 25: Machine Testing Results 133
Table 26: Global Countries Feature List vs. the Malaysia States Feature List 135
Table 27: Correlation Coefficient for 19 Features vs. 10 Features 136
Table 28: SVM's Performance 138
Table 29: SVM's Prediction on Socioeconomic Potential for States in Malaysia 138
Table 30: ROI Simulation Results for 13 States and 3 Federal Territories in Malaysia 139
Table 31: Difference of Data Availability from Different Database 141
Table 32: Consistency of Use of Features in Relevance to Literature Review 146

LIST OF BOXES

Box 1: Current State of the Art of Broadband Investment in the Private Sector 5
Box 2**:** ROI Model 6
Box 3: Current State of the Challenges of Broadband Investment in the Public Sector 7
Box 4: Current State of Machine Learning Technique for Broadband Planning 10
Box 5: Difference between regression and classification problems 26
Box 6: Definition of Rural Areas in Malaysia 57
Box 7: Definition of Rural Areas in India 57
Box 8: The Birth of Broadband 67
Box 9: Relevance of the BB Development Stages in this Research 81
Box 10: World Bank Country Classifications by Income Level 2015-2016 117

LIST OF ABBREVIATIONS

A4AI	Alliance for Affordable Internet
AI	Artificial Intelligence
ARPU	Average Revenue Per User
ASEAN	Association of Southeast Asian Nations
BB	Broadband
BS	Broadband Services
BRIC	Brazil, Russia, India, China
CapEx	Capital Expenditure
CCI	Communications Content and Infrastructure
CV	Cross Validation
CVA	Cross-Validation Accuracy
DOSM	Department of Statistics, Malaysia
EPP	Entry Point Projects
ETP	Economic Transformation Programme
FELDA	Federal Land Development Authority
FSR	Florence School of Regulation
GDP	Gross National Products
GNI	Gross National Income
GTP	Government Transformation Programme
HH	Household
ICT	Information and Communication Technology
IT	Information Technology
ITU	International Telecommunication Union
KPI	Key Performance Index
LDC	Least Developed Countries
LTE	Long Term Evolution
MCMC	Malaysian Communications and Multimedia Commission
MCT	Multipurpose Community Telecentre
MDG	Millennium Development Goals
MIMOS	Malaysian Institute of Microelectronic Systems
MLT	Machine Learning Technique

MTA	Machine Test Accuracy
NBI	National Broadband Initiatives
NBP	National Broadband Plan
NEP	National Economy Policy
NDP	National Development Plan
NGA	Next Generation Access
NVP	National Vision Plan
NGO	Non-Governmental Organization
NKEA	National Key Economic Area
NKRA	National Key Result Areas
OECD	Organization for Economic Co-operation and Development
OpEx	Operating Expenditures
PEMANDU	Performance Management & Delivery Unit
PESTEL	Political, Economic, Social, Technological, Environmental, Legal
PIAP	Public Internet Access Center
POPS/pops	Populations
PPP	Public-Private Partnership
RD	Rural Development
SKMM	Suruhanjaya Komunikasi dan Multimedia
SVM	Support Vector Machine
SVC	Support Vector Classification
TEA	Techno-Economic Analysis
TELCO/TELCOs	Telecommunication Companies or Service Providers
USF	Universal Service Funds
USP	Universal Service Provision
ZTE	ZTE Corporation

LIST OF NOTATIONS

N	dimension of feature space
$y \in Y$	output and output space
$x \in X$	input and input space
F	Feature space
$(x \cdot z)$	inner product between x and z
$\phi: X \to F$	mapping data from input space to feature space
Φ	mapping to feature space
$K(x,z)$	kernel $(\phi(x) \cdot \phi(z))$
$f(x)$	real-valued function before thresholding
N, n	dimension of input space
R	radius of the ball containing the data
w	weight factor
b	bias
α	dual variables or Lagrange multipliers
L	primal Lagrangian
$\|\cdot\|p$	*p*-norm
ln	natural logarithm
e	base of the natural logarithm
log	logarithm to the base 2
x', X'	transpose of vector, matrix
$\mathbb{N}, \mathbb{R}$	natural, real numbers
S	training sample
ℓ	training set size
η	learning rate
ε	error probability
δ	confidence
γ	margin
ξ	slack variables
h	VC dimension
$\forall$	For all

1 INTRODUCTION

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

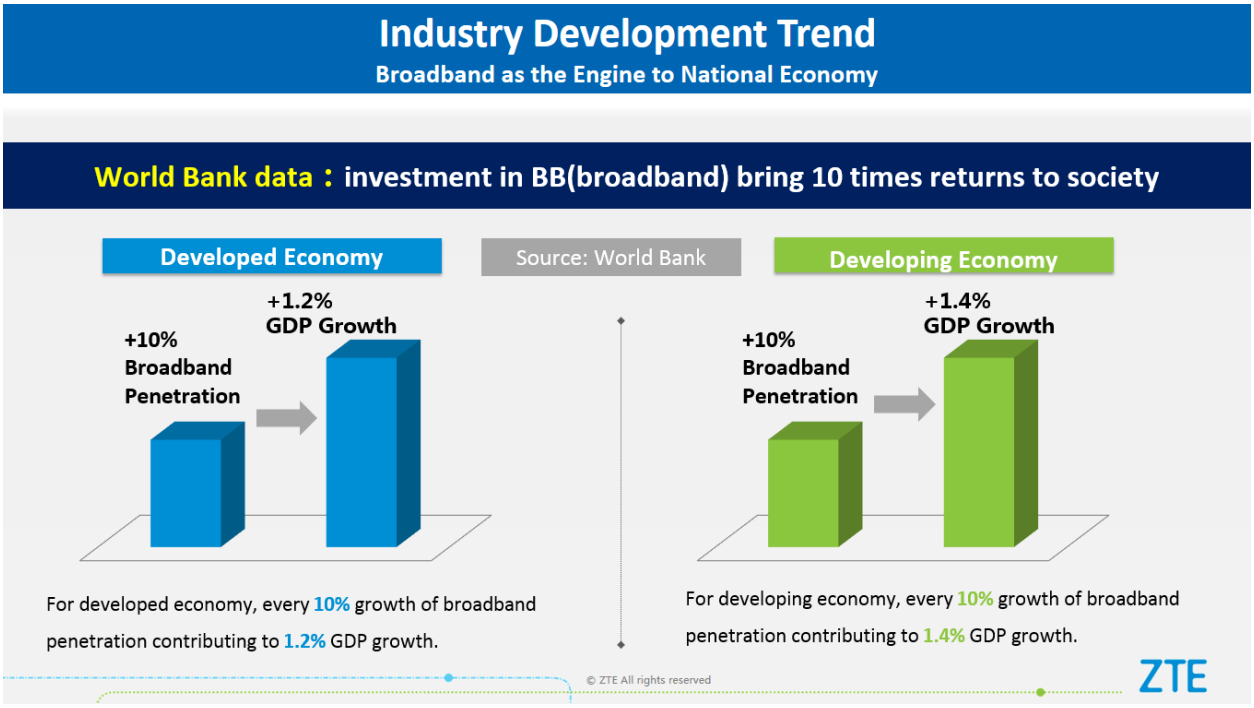


Figure 1: BB Investment Improves Economies. Source: ZTE

Many researches (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.

Many types of research (i.e., Dewan and Kraemer[13], OECD[14], Daveri[15], Waverman et al.[16], Chakraborty and Nandi[17], Choudrie and Dwivedi[18], Karner and Onyeji[19], Kongout et al.[20], Sabbagh et al.[21], Lucas[22], Orbicom-ITU[23], Uppal and Mamta[24], Cronin et al.[25][26][27], Wolde-Rufael[28], to name a few) have found that

the impact of broadband services vary according to economic situations.

However, only the developing and developed economies are normally assumed to be commercially viable for private investments. The underdeveloped economies require legislative support for BB development because quick profits are unattainable in these areas even though 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 BS in urban and selected suburban but not rural areas. Furthermore, there are no empirical models made available for TELCOs to predict the economic potential of certain geographic areas so that they can prioritize their network investments in promising rural areas. As a result, TELCOs continue using the return on investment (ROI) model to strategize their network investment plans and to deploy their BS in urban or suburban areas only.

ROI is a common business term used to identify the potential financial returns which indicate how successful 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 of financial return. Sometimes, ROI is expressed as a payback period regarding some years to recover the financial investment.

This thesis aims to provide a framework of machine learning technique (MLT) which can help TELCOs and government agencies to predict whether a selected geographic area has the socioeconomic potential for BB investments. The proposed machine learning technique is a support vector machine (SVM) which is a classifier that predicts the socioeconomic potential in correspondence to the local features or characteristics of a geographic area.

In this framework, a curve-fitting technique will be applied to formulate the empirical model by using prototyping data from the World Bank databank. The empirical model is then

used as a fitness function for a genetic algorithm (GA) to generate large virtual data to train, validate and test the SVM. Real-life field data for geographic areas in Malaysia is then provided to the SVM to predict which areas have the socioeconomic potential for BB investment.

The curve-fitting technique is a statistical model to formulate the empirical model by establishing the interdependency of the geographical features with the socioeconomic status of geographic areas. Statistical models are empirical expressions of reasoning rather than a physical process whereby they 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 between two or more variable features[29].

It will be ideal to use the real-life field data on local features of rural areas to help in curve-fitting when finding the empirical model. However, the data on local features of rural areas are lacking or difficult to obtain. Hence, the data on local features of countries from the World Bank database is applied instead. 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[30].

This thesis 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[31].

Both the World Bank's data sets and GA-generated data sets can be used to train the

SVM to classify the response (in this case, it is the socioeconomic potential of a geographic area) and to observe the SVM's accuracy in performing the classification with different kernels.

Training data is used as input for the machine (SVM) to learn the pattern of the data. The ability to identify the patterns of the data allows better understanding and optimization of the learning process[32]. SVM is based on the statistical learning theory and applies to various areas[33]. 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[34]. The kernels in SVM 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[35]. 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 2[36].

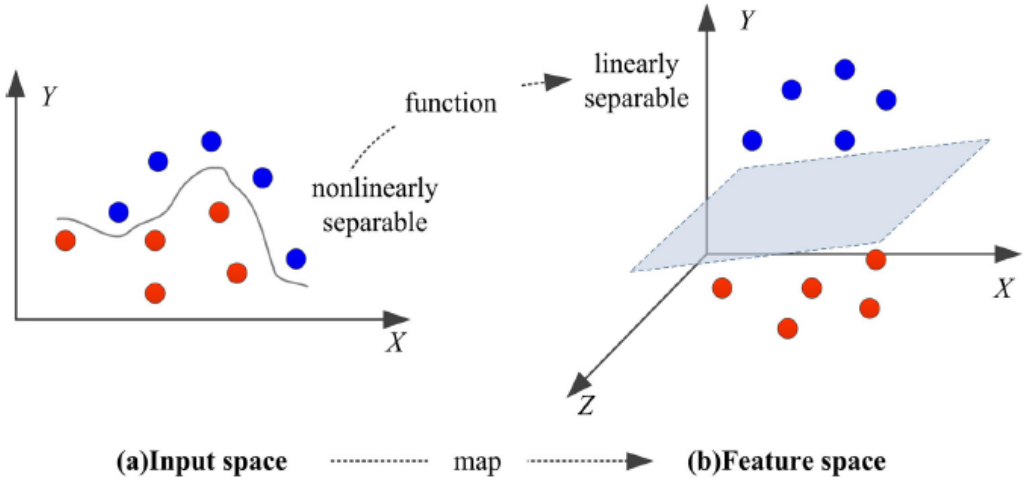


Figure 2: SVM maps data in input space to feature space, and separates the data

1.1 RESEARCH BACKGROUND

1.1.1 CURRENT STATE OF BROADBAND INVESTMENT IN THE PRIVATE SECTOR

Box 1: Current State of the Art of Broadband Investment in the Private Sector

Background Summary:

- TELCOs continue investing in BB services to create BB availability and accessibility in areas with high socioeconomic potential.
- TELCOs continue focusing only on commercially viable areas which have a high potential for BB adoption and affordability.
- BB adoption and affordability will deliver an attractive return on investment to the private sector.
- ROI Model is suitable for urban areas but not rural areas.
- Rural areas are not on the radar of investment by the private sector.

TELCOs in Malaysia have been continuously investing their BS nationwide. As a result, household BB penetration in Malaysia has risen from 10.9%[37] in 2006 to 67.8% in Q3 2014[38]. Between 2010 – 2014, Malaysia's gross national income per capita has increased steadily from RM27,819 to RM34,945 (Malaysian Ministry of Finance). The Internet's impact on the Malaysian economy is one of the highest among aspiring countries, at 4.1 percent of GDP[11].

Through the author's working experience and interviews with professionals in the telecom industry, it is found that the return-on-investment (ROI) model is commonly used for the TELCOs to plan and prioritize their BB investments in Malaysia.

Typically, TELCOs use the ROI model to prioritize their investments in deploying BS to benefit a new area or to upgrade their BB networks in existing areas with BS. 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 1 for details of the operating model.)

Box 2**:** ROI Model

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

where, Revenue = ARPU x number of subscribers

Incremental Cost = operating cost over time

Capital Expenditure = upfront capital investment

TELCOs will invest to create BB availability and accessibility in geographic areas with high socioeconomic potential, in which the BS can be perceived to be adopt-able and affordable.

BB availability is about building the BB network infrastructure and make the network on-air in a geographic area. BB accessibility is about awareness of the service availability and literacy of information technologies that equip the users with devices and know-how to use the BS to access the Internet. Affordability is a comparison of BB prices over income level, which indicates the users' ability to pay for BS. Adoption is the increase in some subscribers who are willing to pay for BS to improve their work-style, lifestyle, and socioeconomic standards.

Both the adoption and affordability of BS have domineering impacts on the ROI. The BS adoption and affordability will generate the number of paying subscribers who will contribute to the revenue, and in turn, will enable on-going BB sustainability.

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

Table 1: Network Coverage by the Malaysian TELCOs

TELCO	***Network Coverage by Population***
Maxis	80%[39]
Celcom	80%[39]
Digi	60%[39]
P1	50%[40]
YES	65%[41]

The numbers of subscribers and average revenue per user (ARPU) are two dominant contributions to a good ROI. The rural areas are lacking in both dominant factors and thus very often fall outside the radar of the TELCO's investment plans. As a countermeasure, Malaysian Communications and Multimedia Commission implemented a universal service provision fund as the government vehicle to invest in telecommunication infrastructure and provide telecom services (broadband included) to rural areas, as well as other specific areas TELCOs would not typically invest.

1.1.2 CURRENT STATE OF BROADBAND INVESTMENT IN THE PUBLIC SECTOR

Box 3: Current State of the Challenges of Broadband Investment in the Public Sector

Background Summary:

- The Gap in the digital divide is a national agenda to be addressed.
- Through National Broadband Initiatives (NBI) and the Universal Service Provision (USP), the government continues building BB infrastructure and developing a value-added program to increase BB access in rural areas.
- The public-private-partnership (PPP) projects are usually funded initially by the USP and then taken over by private investments once the rural areas reach a certain level of socioeconomic standard to stimulate BB adoption with affordability.

- However, without an empirical business model, the USP projects will typically be implemented across rural areas without knowledge of the socioeconomic potential of such areas.

Rural areas remain a key economic focal point for the country due to the vast segment of Malaysians living there. The rural development (RD) is a national key result area (NKRA). RD-NKRA aims to ensure the citizens who choose to live in rural areas can live healthy and sustainable lives[42]. A significant amount of budget and priorities have been set under the Economic Transformation Programme (ETP) by the government in addition to a huge fund accumulated from the Universal Service Provision (USP) contributed by the local TELCO to improve the BB development in rural areas nationwide. The mandatory contributions by the TELCOs have made the Malaysian USP an anchor system to enlarge BB availability and accessibility to use of BS and Internet applications throughout Malaysia[43]. MCMC, as the regulator, has been using the USP fund in encouraging the installation of network facilities as well as the provision of services in underserved areas. The other significant initiative of the Economic Transformation Programme (ETP) is the CCI (Communication, Content, and Infrastructure) which has been scoped in National Key Economic Areas (NKEA) to increase BB household penetration and bridge the digital divide by getting more rural communities online[44].

The NBI and USP aim to reduce the digital divide between urban and rural areas. Figure 3 illustrates the digital divide defined by the 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 investments. The government can then form a PPP with a TELCO to fill up the value gap that stimulates further socioeconomic growth, which will in turn drive BS adoption and affordability that are favorable to the TELCO’s ROI.

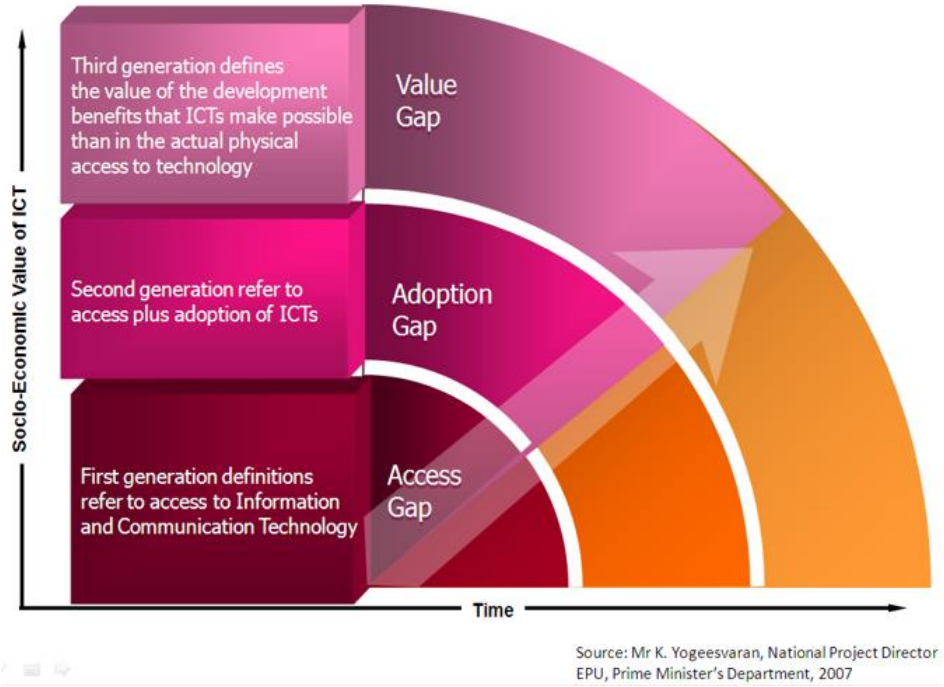


Figure 3: Definition of Digital Divide in Malaysia

The government can focus its effort on improving the BS availability and accessibility for those geographic areas where their comparative socioeconomic potential can be identified. Public investment can then be bridged with TELCO investments 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 socioeconomic potential and the corresponding governmental strategy are not within the scope of this thesis.)

Nevertheless, the NBI and USP projects are typically implemented across rural areas without knowing if one rural area has more socioeconomic potential than the other. With a machine learning model, the public and private sectors can identify those areas with socioeconomic potential for the government to provide BS with USP funding.

Figure 4 shows the different investment models for urban, suburban and rural areas. Public funding (e.g., USP) and cost-sharing are typical models for PPP to develop telecommunication services in rural areas.

		Geographic mix		
		Urban	**Sub-urban**	**Rural**
Financing strategies	**Municipal/Regional**	Municipality as an investor		Public/private credit financing
	Public/Private Partnerships			Public service delegation
	Operator-funded	Incumbent funded Joint venture Multi-fibre		Cost sharing model
	Operator-funded and public policy stimuli		Public funding programme	

Figure 4: Next-Generation-Network Financing Model[45]

Both Malaysian agencies and TELCOs are continuing their BB investment for their national agenda or business plan respectively. The missing link is a network planning technique that can help TELCOs to justify expanding their private investments in urban areas to promising rural areas and to help government agencies prioritize their public investments in promising underserved areas which can be started as a USP project and eventually bridged over to private investments for continuity.

1.1.3 CURRENT STATE OF MACHINE LEARNING TECHNIQUE FOR BROADBAND PLANNING

Box 4: Current State of Machine Learning Technique for Broadband Planning

Background Summary:

- Availability of machine learning technique is unknown to help public and private sectors to classify rural areas with socioeconomic potential for BB investment.
- A new technique is necessary for BB planning to bridge the digital divide.

- It is possible to research a machine learning technique as a planning tool for BB investment to close the digital divide.

Each geographic area is unique according to its characteristics. These characteristics can be complex and vary due to multiple factors like political, economic, social, technological, environmental and legal factors as illustrated in the PESTEL framework.

It is complex to have an empirical model to address TELCO's appetite and the government's nation-building agenda in rural areas. The geographical features can be a complex combination of the PESTEL factors as shown in Table 2.

Table 2: Example of Geographical Features Tabulated in PESTEL Model

Category	**Geographical Features**
Political	Telecommunication acts, policy, ICT investments in the last three years, non-ICT investments in the last three years, number of years of community BB center available, etc.
Economical	GDP contribution % by different industries, GDP, GDP per capita, GNI, GNI per capita, type of economic activities, household income, etc.
Social	% of households with grid electricity, % of households with piped water, number of secondary schools available, number of households, average dependency per household, household income, number of populations, population density, % of labor force, average age, gender ratio, birthrate, % of populations with secondary education, % of population with post-secondary education, labor force count in last three years, household count in last three years, etc.
Technological	Availability of adjacent wireless BB network, length of tarred roads, % of household with computer access, % of fixed BB penetration, % of wireless BB penetration, % of fixed telephony

	penetration, availability of BB technologies, deployment constraints, etc.
Environmental	land size, % of land size for agro-economic activity, average monthly rainfall, average daily temperature, distance from nearest town, distance from nearest wholesales agro-market, building structures, road system, etc.
Legal	number of local authority offices available, number of post offices available, tax on BB, ICT scheme and incentives, etc.

If there was a technique to determine the socioeconomic potential of rural areas for BS, such a technique could potentially help the government agencies to strategize its national BB initiatives. This technique, if available, could also predict areas with socioeconomic potential for the government and TELCO to foster a PPP to invest in BB infrastructure in certain areas where BB is under-served (i.e., areas without BS, areas with only narrowband services, or areas with low BB penetration).

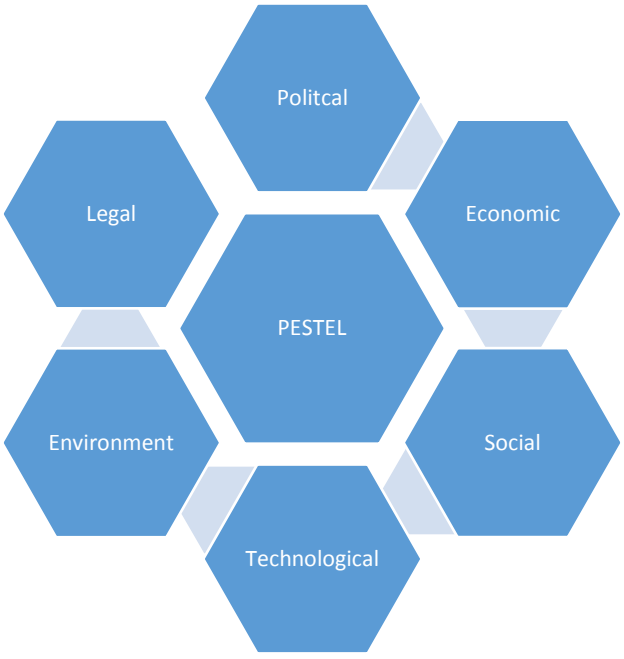


Figure 5: PESTEL Framework

We can then use the PESTEL data to train an artificial intelligence machine (e.g., support vector machine) to learn about the pattern of data sets in correspondence with their response. The data sets are the collection of data for each feature, whereas the responses are the socioeconomic values. A trained machine can separate geographic areas with socioeconomic potential from those areas without socioeconomic potential.

Conceptually, machine learning can be executed through 4 stages:

1. Input the PESTEL data in input space into a machine
2. Machine transforms the data in a high-dimensional feature space
3. Machine learns the pattern of the data and separates the data
4. Machine classifies the data according to the groups pre-defined

In the event of insufficient data, we can use statistical processing software to produce an empirical model to quantify the statistical relationships between the values of those features and the socioeconomic potential of a particular area. This empirical model can serve as a fitness function to generate virtual data to train an artificial intelligence machine (e.g., support vector machine) to learn the pattern of data sets in correspondence with their response. The data sets are the collection of data for each feature, whereas the responses are the socioeconomic values. A trained machine can classify a geographic area that has socioeconomic potential.

1.2 RESEARCH ISSUES

The ROI-based planning model is not good enough to address the current state of BB investment in the private and public sectors, especially on planning for BS beyond urban and sub-urban areas. This thesis aims to provide a framework of machine learning technique which the TELCOs and government agencies can use to plan for BB investments in Malaysia, especially for rural areas.

The research is to establish a framework of machine learning technique that has the artificial intelligent capability to study the geographical features of different geographic areas and classify those areas that have socioeconomic potential.

There are five (5) key technical challenges or uncertainties affecting the development of a framework of machine learning technique. The challenges are:

i. To search for appropriate features of geographic areas as indicators of their measurable socioeconomic potential
ii. To obtain sufficient data or generate virtual samples to train the machine
iii. To train the machine with training data and validate the training accuracy
iv. To test the machine with testing data and observe its performance
v. To apply the machine with real field data

The first two issues are the input stages in this research. The third issue is the processing stage whereas the fourth issue is the output stage of this research. The final stage is to verify if the output is applicable.

The following sub-chapters elaborate each of these technical challenges.

1.2.1 GEOGRAPHICAL FEATURES AND SOCIOECONOMIC POTENTIAL

1.2.1.1 MEASURING GEOGRAPHICAL FEATURES

The classification of urban-rural areas might vary in different countries. The variation could be a result of the differences in social structure and economic status or even political ideologies. The understanding of this issue will be addressed via a literature review.

Many information and communication technology (ICT) organizations provide or recommend influencing features that affect BB deployment and adoption. To support its member countries in planning for BB development, global organizations (e.g., ITU, World Bank) provide some universal reference guides such as ICT Regulation Toolkit, ICT

Indicators, and the Millennium Development Goals. Similarly, regional organizations (e.g., OECD, ASEAN) have also provided guidelines such as A Digital Economy Toolkit and Regulatory Models for ASEAN Telecommunications that are suitable for use in the respective region, with some variances from the universal guideline.

At a national level, national organizations (e.g., Malaysian Communications and Multimedia Commission) drew a clear guideline in the National Broadband Plan (NBP) and Universal Service Provision (USP) within the provision of Communications and Multimedia Act 1998.

The local Malaysian business organizations (e.g., Maxis, Celcom, Digi, Umobile, WEBE, Telekom Malaysia, Altel, etc.) adopt certain business models with specific features for planning. Some consulting firms (e.g., McKinsey, BCG, and Ericsson) may have also provided reference guides for use by its customers in the private or public sectors.

Wherever possible and available, literature from various sources will be reviewed in this research.

1.2.1.2 SOCIOECONOMIC MEASURES

Since the 1990s, many scholarly researchers have been studying the impact of ICT on the individual consumer, business, as well as governmental and socioeconomic development. The research done in the 20th century is mostly focused on telephony services and narrowband services. For studies of BB development on socioeconomic impact, more scholarly results are found in the 21st century.

Aside from scholarly research, many nonprofit organizations and non-governmental organizations have also reported ICT progress using common indicators to measure ICT progress and its socioeconomic impact over time.

Socioeconomic status is commonly measured by GDP or GDP per capita, which is the mother of all economic indicators in the USA[46]. GDP could be one of the common parameters being used in research on the socioeconomic impact of BB development. Alternatively, GNI or GNI per capita could be another common measure of economic status.

There is a wide range of socioeconomic indicators that could be used by certain organizations or countries. The Conversation[46] highlighted some other welfare measures that could be used as alternative national indicators, such as:

- Index of sustainable economic welfare (ISEW)
- Genuine progress indicator (GPI)
- Genuine savings
- Inclusive wealth index
- Australian unity well-being index
- Gallup-Healthways well-being index
- Gross national happiness
- Human development index
- Happy planet index
- OECD better life index

Wherever possible and available, literature from various sources will be reviewed in this research.

1.2.2 DATA REQUIRED

Two different geographic areas could be different from one another regarding demography, basic amenities available and socioeconomic activities. It is difficult to tell which area will fare better if BB is introduced because real life geographical data is either not available or difficult to be obtained. It is difficult to quantify the variations caused by economic activities (e.g. types of crops or cottage industry), demographic factors (population, gender, education level, locality, etc.) and amenities (electricity, water, school,

health, community & religious centers, etc.) in the process of determining the level of socioeconomic impact that a specific area can benefit from BS.

Wherever possible, real field data are to be obtained as input. If the data is insufficient, the limited real field data will be used as prototyping data to develop an empirical model which will be a fitness function to produce more virtual samples.

This research proposes to use a socioeconomic measure that is available in a database whereby the corresponding data for local features are also available. The data sets for local features and the socioeconomic measure will form the basis to formulate an empirical model. The data might be obtained from World Bank, Malaysian government agencies, and past scholarly academic research.

1.2.3 TRAINING THE MACHINE

This research proposes to use a support vector machine (SVM) as the machine to be trained. Support vector machine, which is based on the statistical learning theory, has arguably outperformed most other predicted algorithms[47][48][49].

Real life data is expected to be scarcely available to train the SVM. Thus, determining the accuracy of the SVM could be a major uncertainty in this research. Virtual samples can be the solution to overcome the issue of limited data (socioeconomic indicators and geographical features) available in the real field. Besides, uncertainties of these features are reasons for the need to have more virtual samples to help the machine learn.

1.2.4 TESTING THE MACHINE

This research proposes to split the available data sets into training sets and testing sets for use in machine training and testing respectively.

To verify the application of the machine, this research proposes to use the tested machine to predict the socioeconomic potential of geographic areas in Malaysia, with real-life field data provided by the Department of Statistics Malaysia (DOSM).

1.2.5 RESEARCH SYSTEM OVERVIEW

The block diagram in Figure 6 shows the relationship among research issues and how the SVM will be implemented in this research.

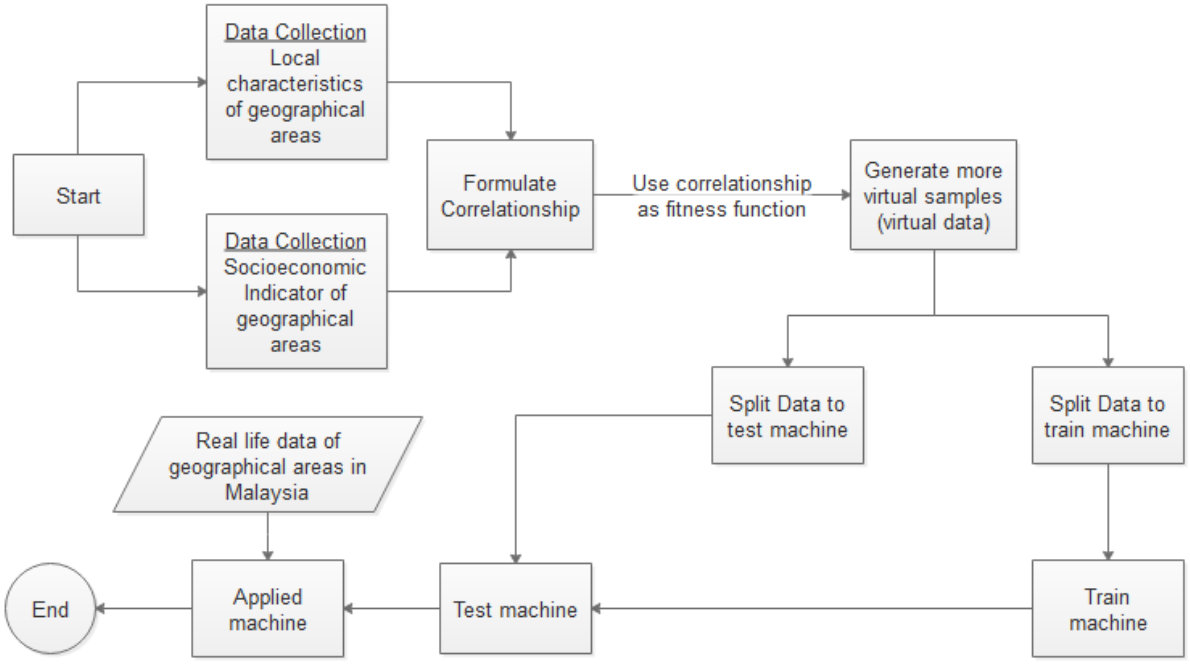


Figure 6: Implementation of Support Vector Machine. Source: the author

1.3 RESEARCH OBJECTIVES AND CONTRIBUTIONS

1.3.1 RESEARCH OBJECTIVES

The primary objectives of this research project are to:

1. Develop an empirical model with local geographical features (i.e., amenities, demography, and BB barriers) that affect the socioeconomic potential of a particular geographic area.

2. Use the empirical model as a fitness function to generate virtual samples to overcome the situation of limited real-life data to train a machine.

3. Test the machine and provide it with real-life data in Malaysia as a technique for classification of geographic areas according to its socioeconomic potential for BB investment.

1.3.2 CONTRIBUTIONS OF THIS RESEARCH

Malaysia has 30% of its population located across approximately 24,000 villages in 2014[50]. The United Nations reported[51] that 46% of the world population resides in rural areas in 2014. The machine learning technique for BB planning, if available, could open up new research areas to enhance public policies, improve PPP models, and to prioritize investment for BB development in rural areas in Malaysia. This technique could also be applied outside of Malaysia.

The outcome of the research is recommended for applications in public and private sectors as a national development planning tool, business planning tool or project management tool as explained below.

A. Application as National Development Planning Tool in the Public Sector

In Malaysia, TELCOs with net revenue exceeding the minimum revenue threshold of RM2 million are to contribute 6% of their net revenue to the USP Fund. This fund is used by the Malaysian government to finance nationwide community BB projects. A group study in the 2014 USP Symposium called for effective use of

USP funds to ensure sustainability for the continuity of USP. Three factors were deemed important for the continuation of USP (see Figure 7).

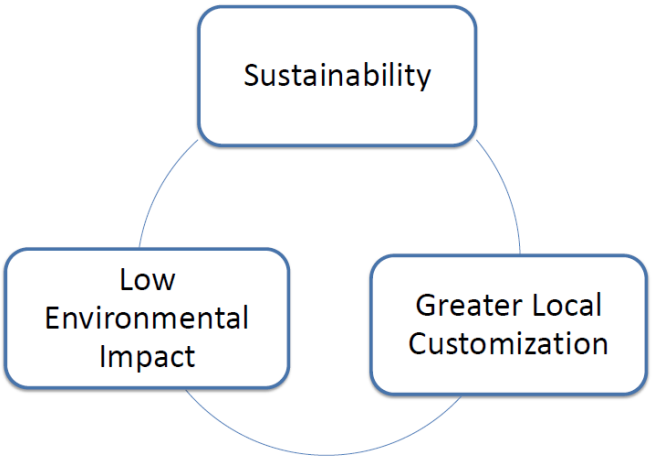


Figure 7: Desired Outcome of USP Symposium 2014, Group 3 Study. Source: MCMC

This thesis aspires to create a machine learning technique that enables to calculate the factors mentioned above. Government agencies can use the technique to classify the underserved or unserved geographic areas according to their socioeconomic potential. Through a PPP, government agencies can apply the empirical model to formulate a game theory for relevant stakeholders to implement corresponding public policies to influence the features of target geographic areas to increase their socioeconomic potential. In turn, the improved socioeconomic potential will improve the BB adoption for service sustainability.

In other words, the machine learning framework is primarily able to classify geographic areas according to their socioeconomic potential, and it is also able to suggest policies or ideas to enhance the socioeconomic situation of a certain area.

The government agency can also apply the machine learning technique to evaluate the benefits of any BB project before a contract is awarded to the selected TELCO for implementation.

B. Application as a Project Management Tool

Under the Communications Content and Infrastructure (CCI) National Key Economic Area (NKEA), which is a key component of the Malaysian Economic Transformation Programme (ETP), more projects are expected to be implemented to realize the country's aspiration to become a developed country by 2020. "Extending Reach" is one of the entry point projects (EPP) to be implemented under the CCI NKEA.

In Malaysia, BB penetration in urban areas is high (60%) whereas the penetration is low in suburban (25%), rural (20%) and remote (15%) areas. Malaysia aims to increase BB penetration in non-urban areas to as much as 90% of households in 2020[44]. The EPP also underlines the integration of BS and economic growth in rural areas. All projects that are deployed under the EPP are expected to have its specific objectives or desired outcome for the communities in rural areas.

This thesis proposes a fitness function with the statistical capability to formulate a correlation between geographical features and the socioeconomic potential of a geographic area. The function could be used to evaluate the potential of the individual project, and also to simulate data for project performance management.

C. Application as a Business Planning Tool in the Private Sector

Rural areas are generally low in population density as compared to urban areas. The low population limits the potential number of service subscribers. The household income per capita in rural areas is generally low when compared to the income levels in urban areas, leading to an impact on the ARPU (average revenue per user) potential. Generally, the business revenue for service providers is a multiplication of its number of subscribers by the ARPU.

Urban population has increased from 28.8 percent in 1970 to 74% in the year 2014 and will reach 75% by 2020[52].

Statistics show that the broadband investment for network coverage in the private sector is in synchronization with the urban population. As the growth of BB penetration in an urban area is becoming saturated, the TELCO can use the machine learning technique to find non-urban areas with socioeconomic potential and prioritize its BB investment beyond urban areas.

D. Application Beyond Malaysia

Globally, BB services are becoming an integral driver of improved socioeconomic performance. Though all three applications mentioned above were mentioned in the context of Malaysia, they are also applicable globally.

ITU (International Telecommunication Union) has adopted *Connect 2020* for Global Telecommunication/ICT Development. Connect 2020 has multiple targets which include: -

- In the developing world, 50% of households should have access to the Internet by 2020
- In the least developed countries (LDCs), 15% of households should have access to the Internet by 2020
- Worldwide, 90% of the rural population should be covered by BS by 2020
- The affordability gap between developed and developing countries should be reduced by 40% by 2020

The first three targets are similar to EPP's "Extending Reach" that was undertaken by the Malaysian government through CCI NKEA; whereas the fourth target deals more with economic growth in rural areas to bridge the digital divide.

With 193 member-states on board, the ITU has all the countries worldwide to support Connect 2020. In other words, many of those countries would have national agendas similar to CCI NKEA.

ITU reported that 53% of the world's population were not using the Internet by the end of 2016 [53]. ITU also reported that 79% of the Europeans were online. In the Americas and Commonwealth of Independent States (CIS) regions, two-thirds of the population was online. In contrast, almost three-quarters of people in Africa were still offline. In the Asia Pacific and the Arab States, the percentage of the population not using the Internet was approximately 58%.

Regarding household BB subscription, the digital divide gap remains wide in developing and least-developed countries as shown in Figure 8. The machine learning technique will be essential to government agencies or government-linked agencies to reduce the gap of the digital divide.

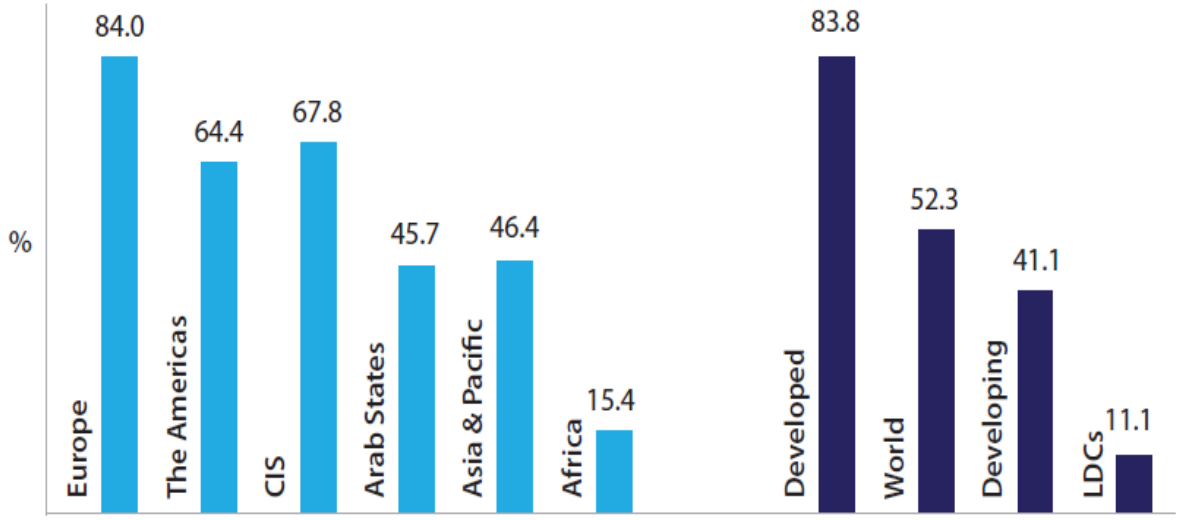


Figure 8: Percentage of Households with Internet Access. Source: ITU

1.4 THESIS ORGANIZATION

Chapter 1 provides an overview of the research background on the relevance of socioeconomic impact to BS. The chapter explains the research issues as well as the technical challenges regarding the machine learning technique. The objectives and contributions are explained with illustration on how the research results could be applied in different fields that involve the partnership of the public sector, the private sector and people in communities.

Chapter 2 covers the findings of a literature review on key elements as underlined in the research topic, "USING MACHINE LEARNING TECHNIQUE TO CLASSIFY GEOGRAPHIC AREAS WITH SOCIO-ECONOMIC POTENTIAL FOR BROADBAND INVESTMENT IN MALAYSIA." Machine learning technique and the data requirement for machine learning are reviewed in the first three sections. The next three sections contain literature reviews on rural areas (which is the targeted geographic area in this research), BB development and techno-economic analysis for BB investment.

Chapter 3 entails the methodology used in the research. The overall methodology is first explained in a block diagram. The method starts with how real-life data is obtained and formulated into a mathematical model. This chapter also explains how the empirical model is developed and used to generate more virtual samples to train an SVM, which is eventually tested for use with actual field data.

Chapter 4 shows the results of the execution of the research methodology and the analysis of the results. This chapter also discusses the results of research experiments and simulation with different datasets and machine algorithm.

Chapter 5 concludes the findings of this research project; and recommends the results for real application in relevant industries. Detected limitations of this research are recommended for future research.

2 LITERATURE REVIEW

2.1 MACHINE LEARNING TECHNIQUE

As early as 1957, Arthur Samuel defined machine learning as a field of study that gives computers the ability to learn without being explicitly programmed[54].

Tom Mitchell (1998) used more up-to-date terms to define machine learning[55]. Tom Mitchell defined the machine learning technique as a computer program which learns from experience (E) concerning some task (T) and some performance measure (P), if its performance on (T), as measured by (P), improves with experience (E). In other words, the performance (P) to do a certain task (T) improves with experience (E). For example:

A. Marking emails as spam or not spam
   - T = Classifying emails as spam or not spam
   - E = Watching you label emails as spam or not spam
   - P = The number of emails correctly classified as spam or not spam

B. Predicting the probability of winning in playing checkers
   - T = Playing checkers again and again
   - E = Experience the results of playing many games of checkers
   - P = Predicting the probability that the program will win the next game

C. Grouping rural areas according to different levels of socioeconomic potential
   - T = Classifying rural areas as having high potential or low potential
   - E = Learning the pattern of geographical features and classified results
   - P = Predicting the potential of another rural area according to its geographical features

There are two types of machine learning techniques, namely supervised learning and unsupervised learning.

In supervised learning, a known response (output) is labeled according to the data sets given (input); and the relationship between the input and the output can be

computationally correlated. This supervised learning technique can solve regression and classification problems. In solving regression problems, a supervised learning machine predicts results within a continuous output. In solving classification problems, a supervised machine predicts results in a discrete output. Box 5 shows an example illustrating the difference between the regression and classification problem in the context of this thesis.

Box 5: Difference between regression and classification problems

1. How was the BB impact on socioeconomic growth across 150 districts in Malaysia in the last three years?
2. Out of the 150 districts, which of those districts have the socioeconomic potential for BB investment now?

Problem 1 is a regression problem whereas problem 2 is a classification problem.

Unsupervised learning deals with data without knowing its output or results. An unsupervised learning machine is capable of establishing a structure from the data by clustering the data based on the pattern of the data.

Machine learning can be applied when there is the existence of a certain pattern in the data, but the pattern cannot be described in a mathematical equation. *Data*, the *pattern in data* and *without a mathematical model* are the three characteristics of the machine learning technique.

Figure 9 shows the various type of machines that are used for different purposes of generalization in a supervised or unsupervised learning model. For example, a support vector machine is a machine that can be used for classification in a supervised learning model.

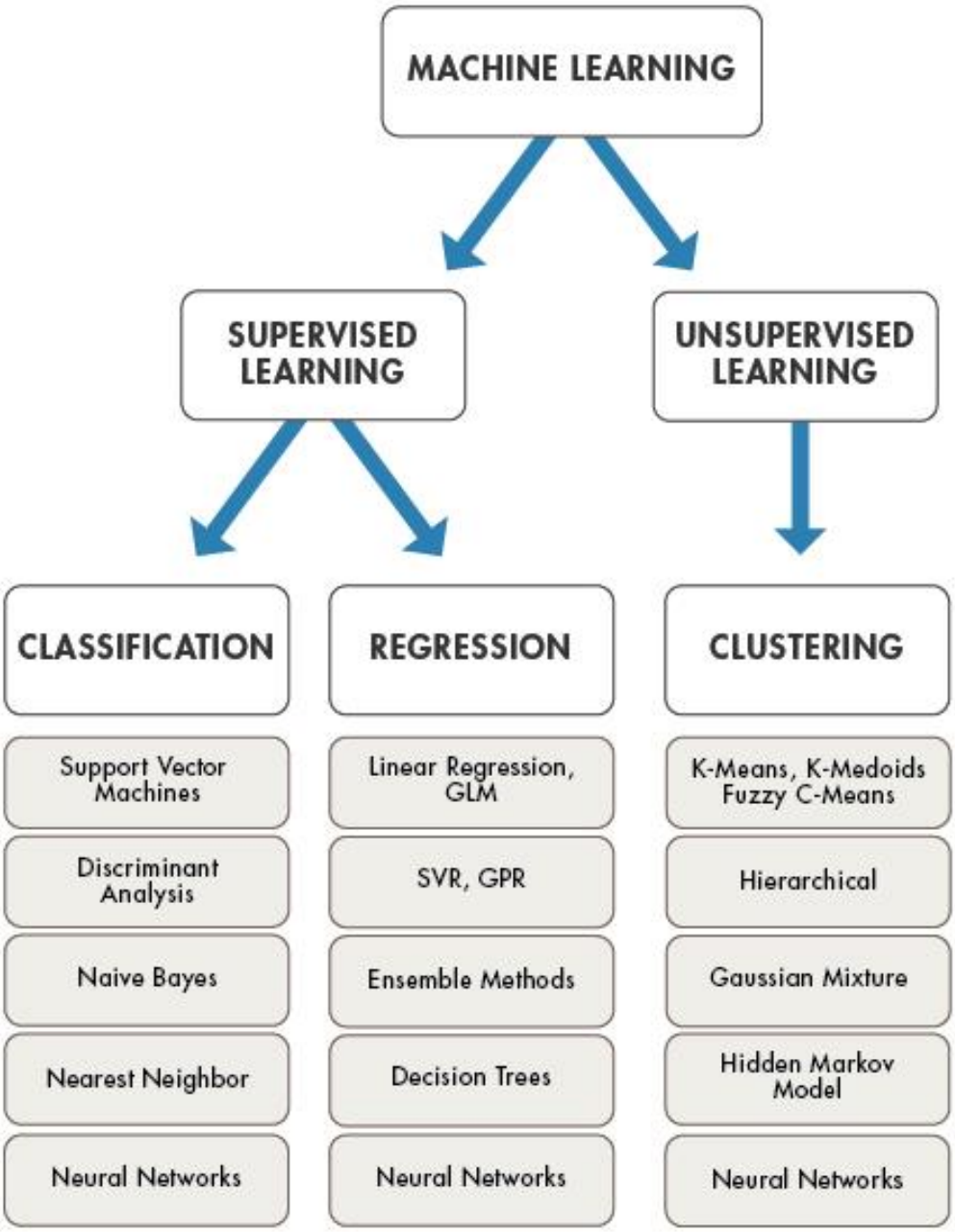


Figure 9: Machine Learning with various learning algorithms. Source:[56]

Various machine learning models are available for research projects. This research focuses on mainly two models, which are the Genetic Algorithm (GA) and Support Vector Machine (SVM).

2.2 GENETIC ALGORITHM

In this research, a genetic algorithm (GA) is used to generate virtual data to overcome the issue of having insufficient real-life data. The virtual data generated by GA can emulate the properties of the real-life prototyping data according to the application of appropriate constraints relevant to the geographical features of rural areas.

Technically, MathWorks defines genetic algorithm (GA) as a method for solving both constrained and unconstrained optimization problems based on a natural selection process that mimics biological evolution. To solve the optimization problems, GA continuously adjusts a population of individual solutions by randomly choosing individuals from the present population and uses them as parents to produce the children for the next generation. Over many generations, the population evolves into a more optimal solution[56]". Figure 10 shows the general flow chart of GA and the with steps involved from the beginning until the termination conditions met.

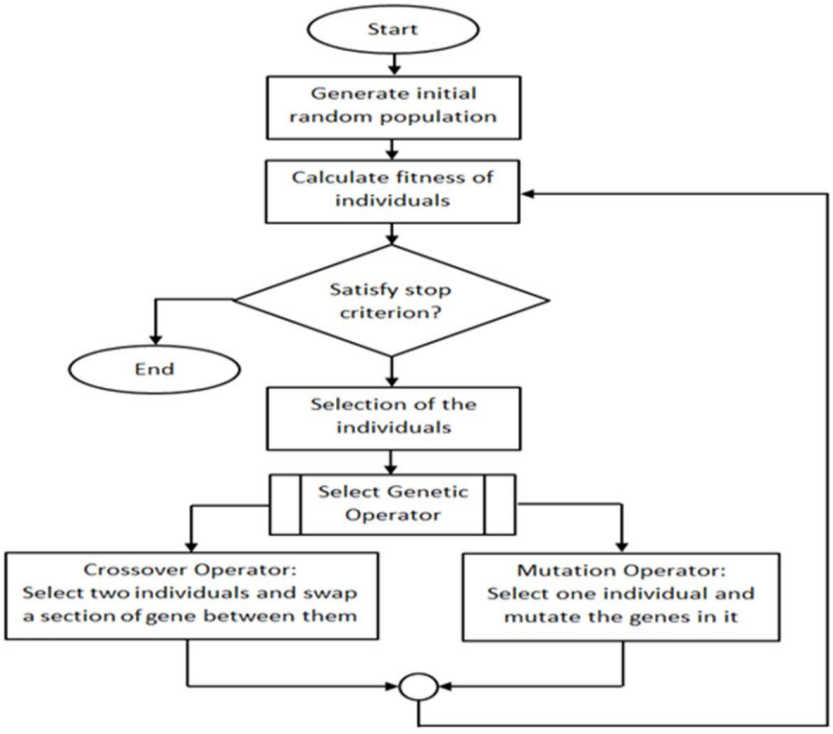


Figure 10: GA Work Flow. Source:[57]

Coley explains GA as a numerical optimization algorithm based on the natural selection process which mimics biological evolution[58]. GA is implemented by initializing a random population of chromosomes that are undergoing selection, crossovers, and mutations to obtain the optimum solution to the problem. The fitness of the chromosomes is determined by the objective function, which in turn determines its ability to reproduce offspring for the next generation. This algorithm is used in various fields as GA is capable of solving large complex problems while other methods may experience some difficulties.

Past research[59][60][61] have been successful in creating large virtual examples by incorporating real-life data as prior knowledge or prototyping data. The artificially generated virtual samples are then used to train the machine learning algorithm for reliable results.

As mentioned earlier, the performance of a machine to do certain tasks improves with experience. A machine can gain more experience when it is trained with large samples. In this research, GA is used to generate large virtual samples to train the support vector machine.

2.3 SUPPORT VECTOR MACHINE (SVM)

In this research, a support vector machine (SVM) is used as a classifier to separate the geographic areas with socioeconomic potential from those without socioeconomic potential.

A support vector machine (SVM) is a supervised learning machine that performs as a discriminative classifier defined by a separating hyperplane. Given labeled training data sets with specific features of an object, the algorithm analyses data and outputs an optimal hyperplane for classification. SVM outputs a map of the sorted data sets with the margins between the two as far apart as possible. Support vectors are the data points that lie closest to the decision surface (or hyperplane). They are the data points most

difficult to classify. They have a direct bearing on the optimum location of the decision surface. Figure 11 illustrates an example of SVM [62].

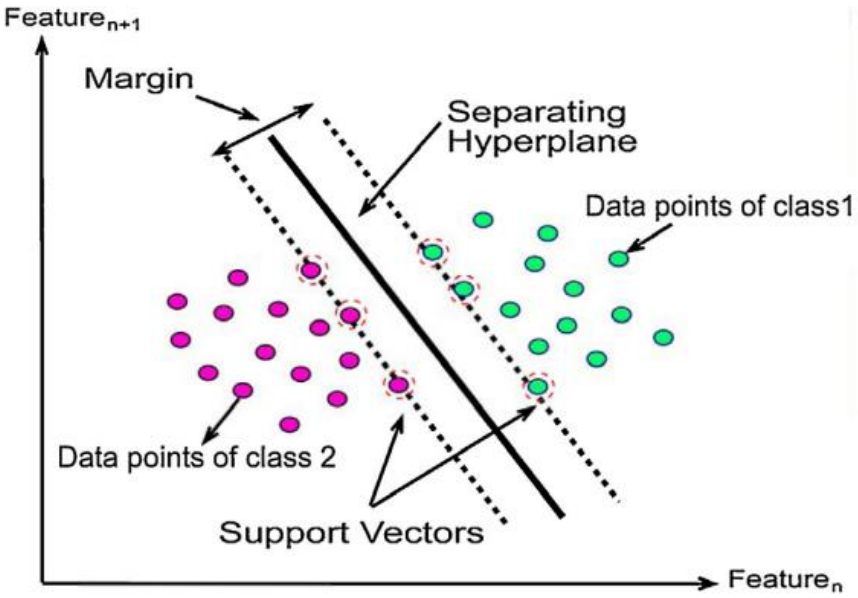


Figure 11: Linear support vector machine example (modified from Burges, 1998)

In other words, SVM is a linear classifier with the capability to detect separable patterns in a target data sets. SVM functions by constructing a separate hyperplane as the decision surface where the margin separating the positive and negative outputs is maximized. The process of maximizing the separation gap is executed according to structural risk minimization – SRM, which is a statistical learning theory in machine learning. The SRM is further explained in section 2.3.9.

The goal of supervised machine learning is to find an algorithm that accurately separates the data into predefined classes. This research proposes to use an SVM as the supervised machine learning technique to classify the socioeconomic potential of certain geographic areas which correspond to the pattern of features of those areas.

Furthermore, SVM is the best machine learning technique for classification.

2.3.1 PATTERN LEARNING AND RECOGNITION

SVM uses statistical pattern recognition. Pattern recognition is about assigning an event or data point to one of the categories, based on features derived to emphasize commonalities[63]. Among the various frameworks for pattern recognition, the statistical approach has been most widely used in practice. The statistical approach is completely domain-independent and computationally inexpensive[64].

An SVM could be used to perform the task of pattern recognition to automate decision-making processes, i.e., decide if a geographical area is suitable for BB investment. Nevertheless, recognizing patterns is challenging when dealing with qualitative or subjective attributes impacting urban and rural development, in both social and economic aspects.

Statistical pattern recognition is performed in two stages: training (learning) and classification (testing) as shown in Figure 12.

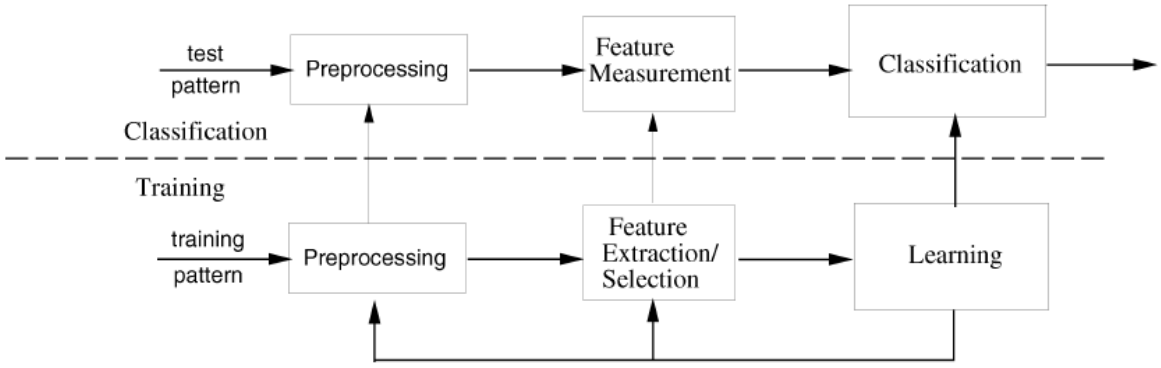


Figure 12: Model for Statistical Pattern Recognition System. Source:[65]

In the context of this research, SVM with different kernels will be provided with input (data) for the kernels to learn about the pattern of socioeconomic data in response to the data of the geographical features for each geographic area. Later, new sets of data of the geographical features will be provided to the trained machine for it to classify the socioeconomic responses.

2.3.2 OBJECT, RESPONSE, AND FEATURES

Data is a requirement for machine learning. Before collecting the data required for machine training, an object (e.g., rural areas, districts, states) and the corresponding features (e.g., GNI per capita, population size, education level, labor force, weather, etc.) need to be defined. The threshold of one feature (selected among the features) will be treated as the response (or phenomenon) being observed. Then the data can be collected according to the features pre-defined.

A feature is an individual measurable characteristic of observation on an object. It is important to select features which are informative, discriminating and independent to create effective algorithms in pattern recognition, classification, and regression.

For example, assuming there are 17 districts in Malaysia and each district has a certain number of households with different income levels in each household. For a district to be classified as a high-income area, its gross national income (GNI) per capita needs to reach $10,000 and above. We can, therefore, categorize the districts into two groups: high-income districts and low-income districts as per Figure 13.

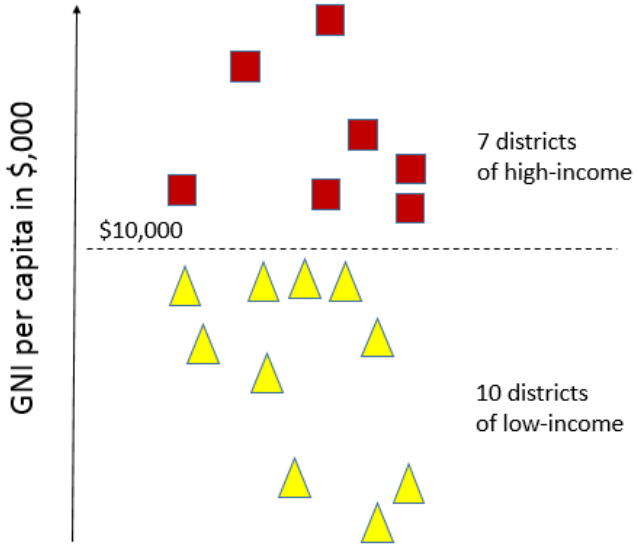


Figure 13: 2-Class Classification of Districts with Single Feature

Based on the case illustrated above, the districts are the objects. The GNI per capita is the geographical feature (or individual measurable property) of the district, and the income level of $10,000 is the response (or phenomenon) being observed.

Another geographical feature for the 17 districts is the average headcount in each household, with the individual measurable property being the number of people in the household. The problem is to identify which high-income households have 3 or more people. Figure 14 below illustrates how the districts are grouped into two different classes according to its features (high income with 3 or more people in the household).

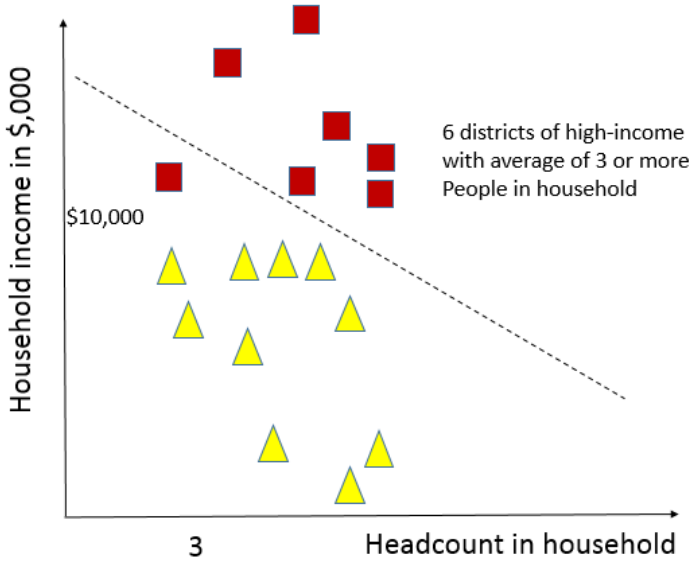


Figure 14: 2-Class Classification of Districts with Dual Features

2.3.3 FEATURE SPACE

As more features are added for observation, the machine training and classification becomes more complex. Feature space is therefore required to observe the subject with multiple features. Feature space is an n-dimensional vector of numerical features that represent some objects.

Extended from the case illustrated in Figure 14, let us take districts with 75% BB penetration as a third feature to be observed. It is difficult to identify all three geographical features on a 2-dimensional feature space as shown in Figure 15. Therefore, a 3-dimensional feature space (Figure 16) will allow a hyperplane to be created to separate the districts that meet all the 3 features (with average income of $10,000 or higher, and with average of 3 people or more in household, and with average of 75% household BB penetration) from other districts.

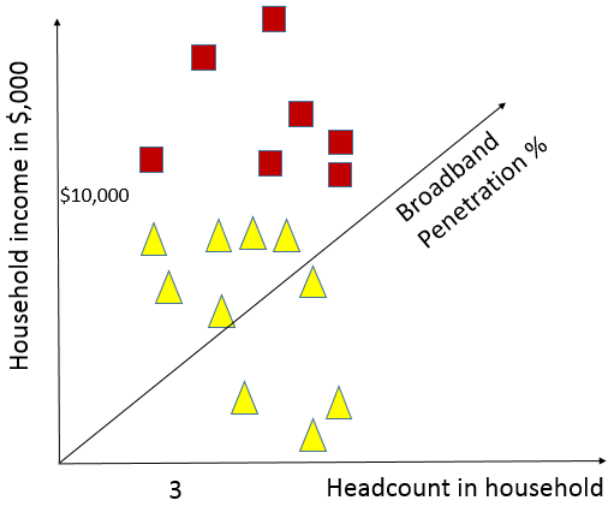


Figure 15: Dilemma in presenting 3 features in a 2-dimensional space

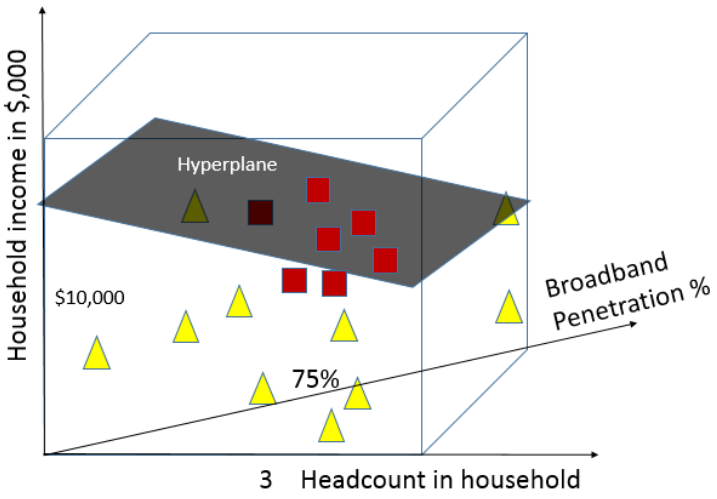


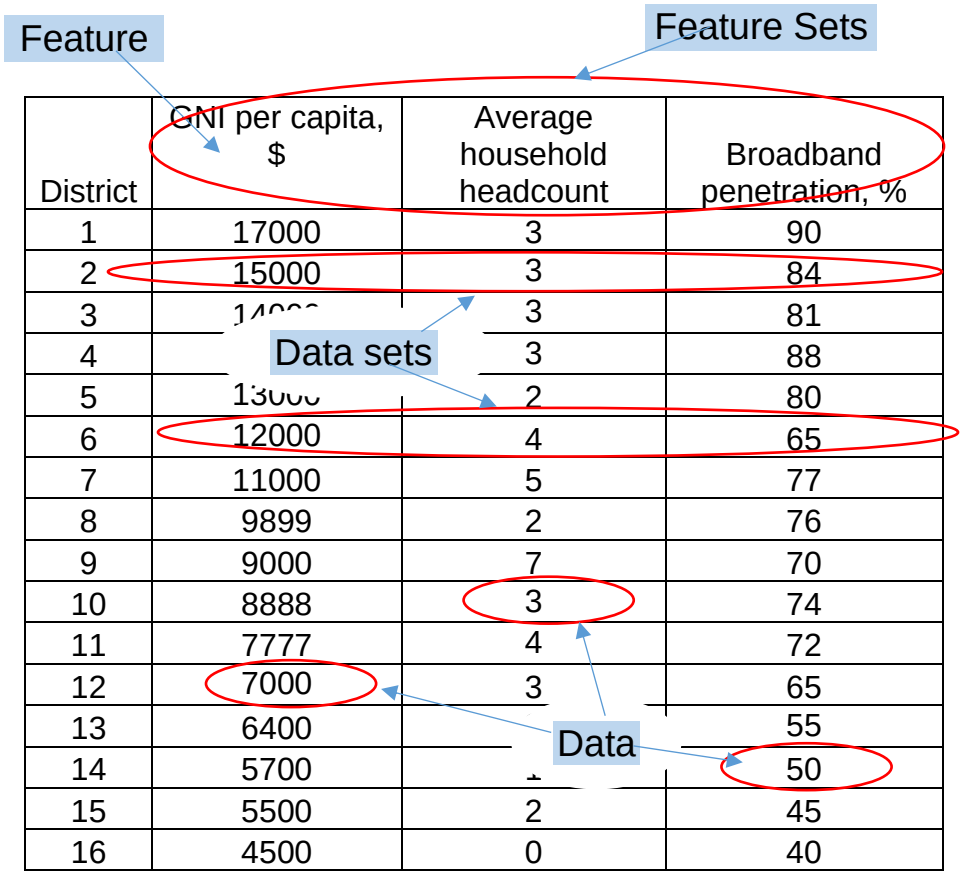
Figure 16: A hyperplane separates the districts that meet the observation from others

Further details on hyperplane are elaborated in section 2.3.7.

2.3.4 DATA AND DATA SETS

Data are facts and statistics of selected features collected for reference or analysis. Dataset is a collection of data. For illustration purpose, let us take a look at a hypothetical case study.

Table 3 below shows an example of data, data sets, feature and feature sets for 17 districts. The three features are GNI per capita, household headcount, and BB penetration. Together, they are called feature sets, and the GNI per capita is selected as the socioeconomic response being observed. In this example, each district has three types of data, and the collection of the 3 data forms a data set for each district. For example, the GNI per capita for District 1 is $17,000. It has an average household headcount of 3 people, and the BB penetration in District 1 has reached 90%.



District	GNI per capita, $	Average household headcount	Broadband penetration, %
1	17000	3	90
2	15000	3	84
3	14000	3	81
4	13000	3	88
5	13000	2	80
6	12000	4	65
7	11000	5	77
8	9899	2	76
9	9000	7	70
10	8888	3	74
11	7777	4	72
12	7000	3	65
13	6400	1	55
14	5700	1	50
15	5500	2	45
16	4500	0	40

17	3788	5	30

Table 3: Example of data, data sets, feature and feature sets

Machine learning techniques are dependent on real-life data in the training phase. Insufficient data is well known as the main problem causing poor performance in machine learning. But sometimes it is labor-intensive, difficult and expensive to collect the required data[30]. In the absence of real-life data or insufficient data, virtual data needs to be generated to train the target machine to determine whether a geographic area has sufficient socioeconomic potential for BB investment. In this research, real-life data will be collected and treated as prototyping data (or prior knowledge) for the creation of virtual samples to increase the training data sets.

Real-life data from the World Bank database is publicly available and commonly used in much past research (refer to section 2.5.3) on issues related to BB development. This research suggests using the data and data sets from the World Bank database. The feature sets in the database will be cross-referenced with the features that are reviewed in section 2.5.3 (features that influence BB development).

The literature review in section 2.5.3 illustrates the various significant geographical features of BB adoption and socioeconomic status. The real-life data for use in this research will be verified against the finding of the literature review.

2.3.5 BINARY AND MULTICLASS CLASSIFICATIONS

Classifying the observations into one of the two classes is called binary classification or dual-class classification. Classifying the observations into one of the three or more classes is called multiclass classification.

For the same 17 districts that are used as examples in the previous case study, let us assume there are four different responses to household income:

1. High-income districts, GNI per capita of $12,236 or more

2. Upper-middle-income districts, GNI per capita between $3,956 and $12,235
3. Lower-middle-income districts, GNI per capita between $1,006 and $3,955
4. Low-income districts, GNI per capita of $1,005 or less

The 17 districts can now be grouped into multi-classes of high-income, upper-middle-income, lower-middle-income and low-income groups as shown in Figure 17.

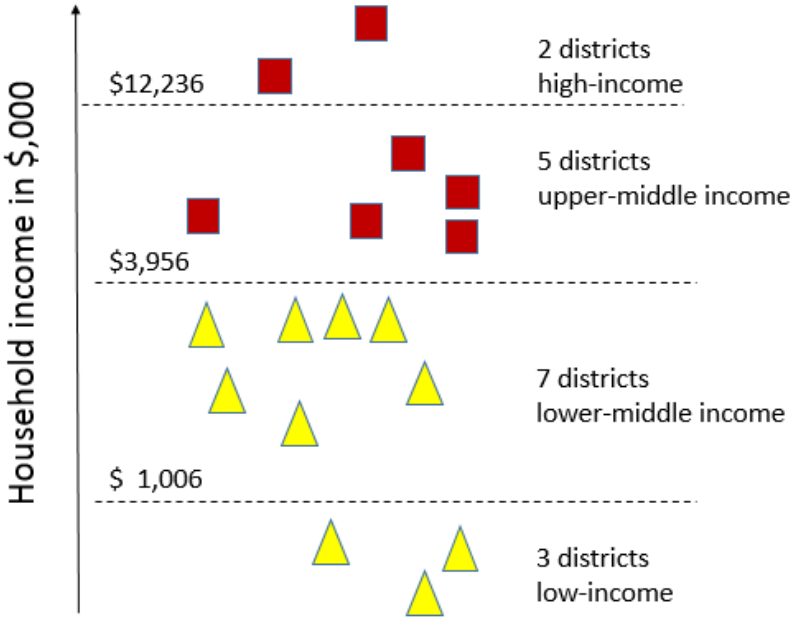


Figure 17: Multiclass Classification of Districts

Cristianini and Shawe-Taylor (2000) summarized that "*a learning problem with binary outputs is referred as a binary classification problem, one with a finite number of categories as multi-class classification, while for real-valued outputs the problem becomes known as regression.*"

2.3.6 LINEARLY SEPARABLE VS. NON-LINEARLY SEPARABLE SPACE

A linear method can solve a problem with data sets that are linearly separable. The separation can be done with a line or a hyperplane which is the classification threshold. If the linear method is not workable (i.e., the data sets are not linearly separable), the next

step is to use a nonlinear method, which involves applying a technique to transform the data sets in such a way that they become linearly separable.

A nonlinear method transforms the data into a new representational space and then applies classification techniques. During the data transformation process, the features in the data sets are described in a structured pattern as compared to the original space. Classification algorithms can then be used to create a more reliable prediction in the new space. Classification techniques that reform (transform to new space) data features before application of the classifier are called nonlinear methods.

The nonlinear transformation can be written as:

$$X = (x_0, x_1, x_2, \ldots\ldots, x_n) \overset{\Phi}{\rightarrow} Z = (z_0, z_1, z_2, \ldots\ldots, z_n)$$
$$Each\ z_i = \phi_i(X)\ \ and \quad Z = \Phi(X)$$

where feature X is transformed to feature Z (or vector x_i is transformed to vector z_i).
A data point is viewed as an n-dimensional vector. For illustration purposes, a simple example of such transformation is shown in Figure 18 below.

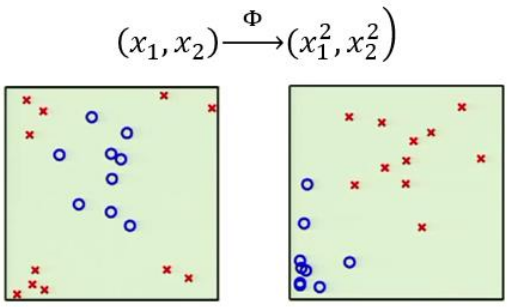


Figure 18: Example of transforming the data nonlinearly. Source:[66]

A linear classifier is used when the number of features is very high, e.g., document classification. A linear SVM has almost similar accuracy as a nonlinear SVM, but the linear machine performs faster in such cases. However, a linear classifier gives very poor results (accuracy) in a nonlinearly separable problem. Nonlinear SVM has capabilities

(kernels) which transform (or map) the input data (Input Space) to a higher dimensional space (Feature Space) where a linear hyperplane can be created to separate the different classes. Figure 19 illustrates an example of an SVM classifier with a nonlinear kernel transforming the input space into a high-dimensional feature space where a linearly separable data classification takes place[67].

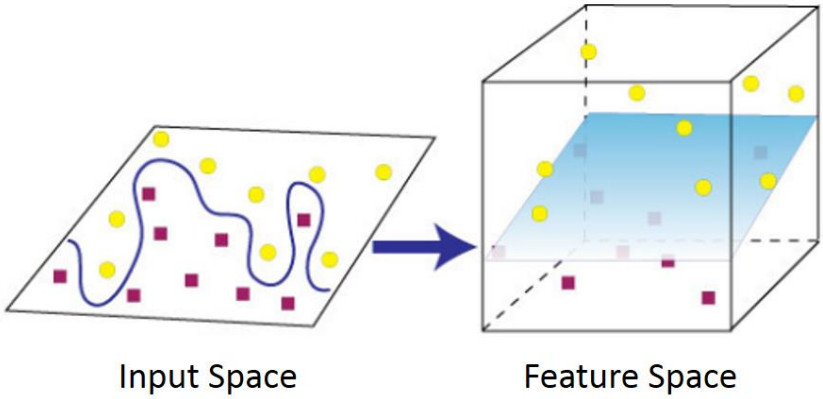


Figure 19: Nonlinear classifier transforms input space into feature space. Source:[67]

In real-world situations, some data points in the two classes might fall into an ambiguous area that is not easily separated by a linear hyperplane. To overcome the ambiguous situations, SVM can introduce a user-defined parameter that specifies the trade-off between the minimization of the misclassifications and maximization of margin; and use kernel functions to add more dimensions to the low dimensional space so that two classes become separable in the high dimensional space.

2.3.7 HYPERPLANE

A hyperplane of n-dimensional space is a flat subset with dimension $(n-1)$ which separates the space into two half-spaces. The hyperplanes in 3-dimensional space are 2-dimensional planes; the hyperplanes in 2-dimensional space are 1-dimensional lines. In other words, SVM constructs an (n-1) dimensional hyperplane to separate two classes in an n-dimensional space. For example, two variables in a data set will create a 2-

dimensional space; the separating hyperplane would be a 1-dimensional straight line dividing the space.

When more dimensions are involved, SVM will search for an optimal separating hyperplane, $w^T\varphi(X) + b = 0$, also known as the maximum-margin separating hyperplane. The nearest data points on each side of the hyperplane are called support vectors, w. The distance between the nearest data points on each side is called margin, as shown in Figure 20.

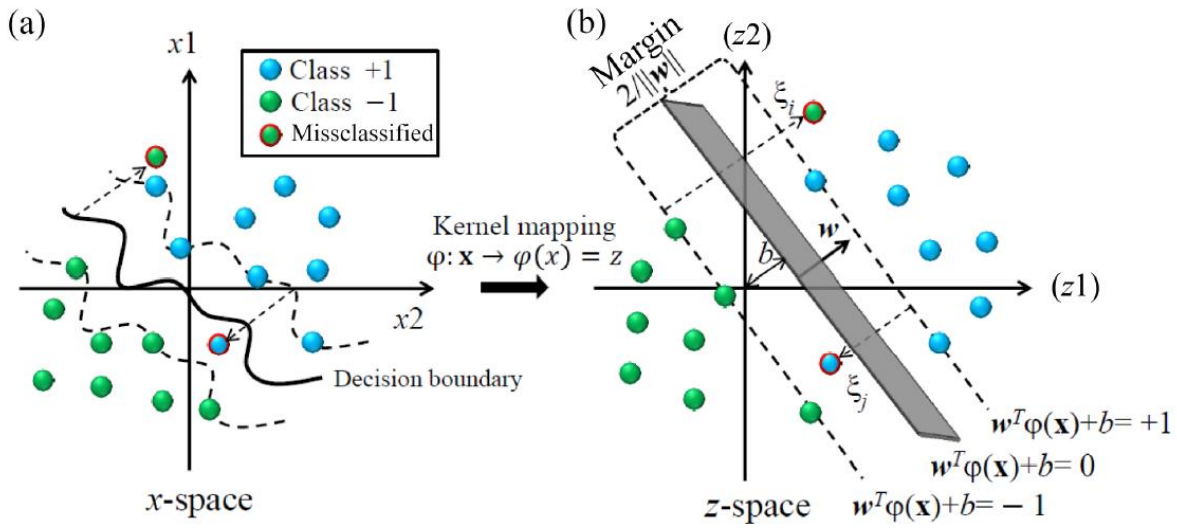


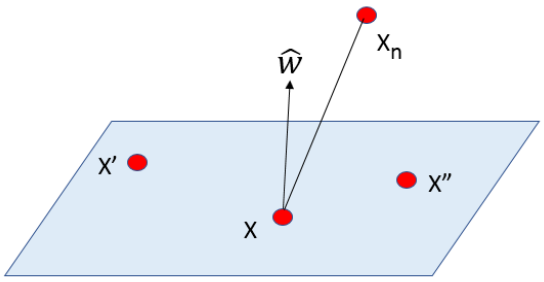
Figure 20: SVM with maximum-margin separating hyperplane. Source:[68]

The equation of a hyperplane can be written in the form of vectors:

$$W^T X + b = 0 \text{ [sometimes, written as } \sum_{i=0}^{d}(w_i x_i) + b = 0]$$

The plane $W^T X + b = 0$, where $|w^T x_n + b| = 1$, the vector w is 1 to the plane in X space.

If x' and $x^{"}$ are 2 data points sitting on the plane, then $w^T x' + b = 0$ and $w^T x^{"} + b = 0$. Hence $w^T(x' - x^{"}) = 0$.



Assuming x_n is any point above the plane, the projection of $x_n - x$ on vector w is:

$$\widehat{w} = \frac{w}{\|w\|}$$

Distance between x_n and the hyperplane

$= |\widehat{w}^T(x_n - x)| = \frac{1}{\|w\|}|w^T x_n - w^T x| = \frac{1}{\|w\|}|w^T x_n + b - w^T x - b| = \frac{1}{\|w\|}|0 - 0| = \frac{1}{\|w\|}$

The total margin on both sides of the hyperplane is $2 \times \frac{1}{\|w\|} = \frac{2}{\|w\|}$

Figure 21 shows an SVM with its hyperplanes (H) and support vectors (v). The three circled coordinates are the input factors or the tips of the vectors.

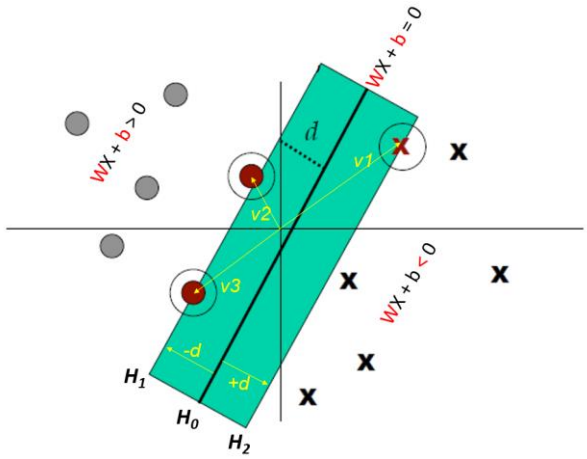


Figure 21: Linear classifier with hyperplanes (H) and support vectors (v)

$w^T x + b \geq 0$ for $d_i = +1$

$w^T x + b < 0$ for $d_i = -1$

where the margin of separation d is the separation between the hyperplane and the closest data point for a given weight vector w and bias b.

Given a function $f: X \subseteq \mathbb{R}^n \rightarrow \mathbb{R}$ and input $\boldsymbol{x} = (x_1, \ldots, x_n)$ is assigned to the +1 if $f(x) \geq 0$ otherwise to -1, a linear classification function can be written in the form of vectors:

$$f(x) = \ sign\ (w^T \boldsymbol{x_n} + b), \text{ where } \ w \in \mathbb{R}^d, b \in \mathbb{R}$$

where $sign$ can be positive (+) or negative (-) for binary classification.
Note: $w^T x_n = w \cdot \boldsymbol{x} = (w_1, w_2, w_3, \ldots) \cdot (x_1, x_2, x_3, \ldots) = w_1 x_1 + \ w_2 x_2 + w_3 x_3 + \ldots$

Concerning Figure 21, the hyperplanes H are defined as:

$$w^T \emptyset(x) + b \geq +1 \ when \ y = +1$$
$$w^T \emptyset(x) + b < -1 \ when \ y = -1$$

In the formula for hyperplanes H, w is the weigh factor to feature x, b is the bias and y is the labelled response. H_1 and H_2 are the hyperplanes, where,

$$H_1: w^T \emptyset(x) + b1 = +1$$
$$H_2: w^T \emptyset(x) + b1 = -1$$

For the computed decision function, the data is correctly classified if

$$y_n(\boldsymbol{w^T x_n} + b) > 0 \ , where \ n = 1,2, \ldots, N$$

or, $y_n(\boldsymbol{w} \cdot \boldsymbol{x_n} + b) > 0 \ \forall n$, when presented in quadratic programming format.

The nearest positive points to the positive plane are labeled as $\boldsymbol{w} \cdot \boldsymbol{x} + b = 1$ whereas the nearest negative points to the negative plane is labelled as $\boldsymbol{w} \cdot \boldsymbol{x} + b = -1$. Together the labels form the support vector (also known as canonical hyperplanes).

Hyperplane = $w^T \boldsymbol{x} + b = 0$

And, sometimes it is also written as $\boldsymbol{w} \cdot \boldsymbol{x} + b = 0$, or $w^T \emptyset(x) + b1 = 0$

Normal vector for hyperplane = $\frac{\boldsymbol{w}}{\|\boldsymbol{w}\|_2}$, where $\|w\|_2 = \sqrt{w^T x}$

Hence, the projection of two points from the margins onto the normal vector

= $\frac{2}{\|w\|_2}$

Lagrangian is introduced to solve the constrained minimization problem. The Primal Lagrange function is defined as:

L_p = minimize $\frac{1}{2} w^T w$, subject to $y_n(w^T x_n + b) \geq 1$ for $n = 1,2, \ldots, N$

= $Min\ \frac{1}{2} \|\boldsymbol{w}\|_2^2$, subject to $y_n(w \cdot x_n + b) \geq 1\ \forall n$

(in terms of Quadratic Programming optimization).

To maximize the margin = maximize $\frac{1}{\|w\|}$, subject to $\min_{d=1,2,\ldots,N} |w^T x_n + b| = 1$

Note: $|w^T x_n + b| = y_n(w^T x_n + b)$

Lagrange multipliers is an optimization technique to find the optimal hyperplane for SVM analytically. The optimization algorithm will generate only the support vectors which determine the weight vectors and the boundary. In other words, the optimization technique will only move those support vectors that optimize the decision boundary of a hyperplane. Lagrange multiplier, Karush-Kuhn Tucker (KKT) conditions, and duality formulate the objective function to optimize the margin and constraints. The formula is presented as:

Minimize $\mathcal{L}(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{n=1}^{N} \alpha_n (y_n(w^T x_n) + b) - 1$

Concerning w and b and maximize with respect to each $\alpha_n \geq 0$

α_n are the Lagrange multipliers. For $\alpha_n \geq 0$, the partial derivatives of the other variables in the primal form is computed and substituted to obtain a dual formulation.

The Lagrangian function is differentiated concerning w and b, imposing stationarity, to obtain the solution for the optimization problem.

$$\frac{\partial L_p}{\partial w} = 0 => w = \sum_{n=1}^{N} \alpha_n y_n x_n$$

By substituting: $w = \sum_{n=1}^{N} \alpha_n y_n x_n$ and $\sum_{n=1}^{N} \alpha_n y_n = 0$

in the Lagrangian $\mathcal{L}(w, b, \alpha) = \frac{1}{2} w^T w - \sum_{n=1}^{N} \alpha_n (y_n (w^T x_n) + b) - 1$

we get $\mathcal{L}(\alpha) = \sum_{n=1}^{N} \alpha_n - \frac{1}{2} \sum_{n=1}^{N} \sum_{m=1}^{N} \alpha_n \alpha_m y_n y_m \boldsymbol{x_n^T x_m}$

which needs to be maximized concerning α, subject to $\alpha_n \geq 0$ for $n = 1,2, \dots N$
and $\sum_{n=1}^{N} \alpha_n y_n = 0$

In the abovementioned formula, w is a weight vector, x is input vector and b is a “bias”. The weight vector increases according to the importance of the input vector. If the input vector has a negative effect to the outcome, the weight vector will become negative.

The Lagrangian can then be solved with quadratic programming:

$$\min_{\alpha} \quad \frac{1}{2} \alpha^T \begin{pmatrix} y_1 y_1 & x_1^T x_1 & y_1 y_2 & x_1^T x_2 & \dots & y_1 y_N & x_1^T x_N \\ y_2 y_1 & x_2^T x_1 & y_2 y_2 & x_2^T x_2 & \dots & y_2 y_N & x_2^T x_N \\ \dots & & \dots & & \dots & & \dots \\ y_N y_1 & x_N^T x_1 & y_N y_2 & x_N^T x_2 & \dots & y_N y_N & x_N^T x_N \end{pmatrix} \alpha + (-1^T)\alpha$$

subject to the linear constraint of: $y^T \alpha = 0$
And, $0 \leq \alpha \leq \infty$, where 0 is lower bounds and ∞ is upper bounds

Subsequently, the quadratic programming will find the value of α.

For $\alpha = \alpha_1, \alpha_2, \dots, \alpha_N$

$w = \sum_{n=1}^{N} \alpha_n y_n x_n$

Under KKT condition: for $n = 1,2, \dots, N$

$\alpha_n(y_n(w^T x_n) + b) - 1 = 0$

When $\alpha > 0$, x_n becomes a support vector

2.3.8 RISK MINIMIZATION AND VC-DIMENSION

SVM is well known for its high computational efficiency and good generalization theory which prevents overfitting through control of hyperplane margin and risk minimization[34][69]. SVM applies Empirical Risk Minimization (ERM) on the training data to reduce the risk of misclassification according to the law of large numbers[69]. In other words, the empirical risk will be minimal when the training sample is infinitely large[35].

Given the data set: $(X, Y) = [(x_1 y_1), (x_2 y_2), (x_3 y_3), \dots, (x_N y_N)]$, the error of classification is formulated as:

$$R(w) = \int \mathcal{L}(y, f(x, w) dP(x, y)$$

where, $y = f(x)$, $\mathcal{L}(y, f(x, w)$ is a predetermined loss function and $P(x, y)$ is the probability distribution function.

The probability distribution function $P(x, y)$ is unknown as it is data dependent. Hence, the empirical risk is evaluated by averaging the error function on the training data sets. The risk calculation is compromised by forming a new risk function:

$$R_{emp}(w) = \frac{1}{2N}\sum_{i=1}^{N}|y - f(x,w)|$$

In which, w is the optimal weight vector which provides the lowest risk.

While ERM mitigates the risk of misclassification by the law of large numbers, structural risk minimization (SRM) minimizes the upper bound on expected risk and undergoes steps to generate a model based on the balance between hypothesis space complexity and training error.

Figure 22 illustrates an example of SRM, with an upper bound on the true risk and the empirical risk as a function of VC-dimension h (for fixed sample size N).

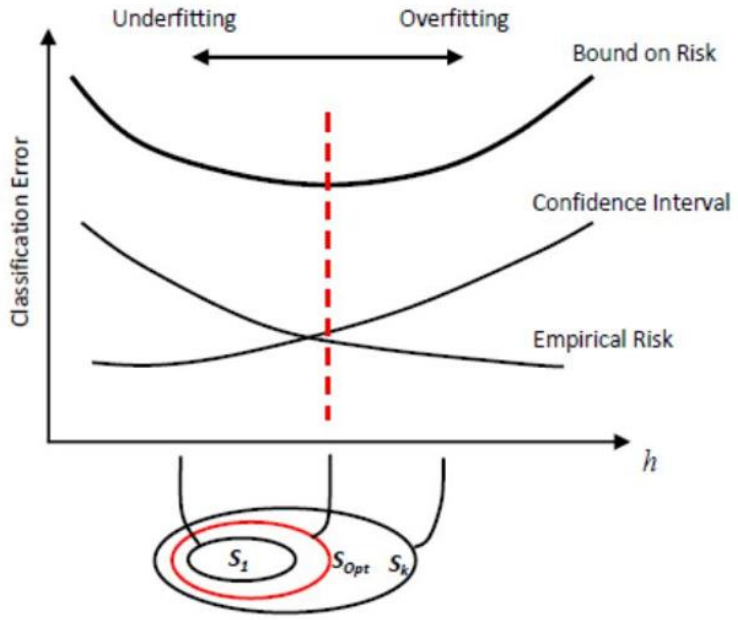


Figure 22: Structural Risk Minimization. Source:[70]

The hypotheses of VC-dimension (h) are divided into a hierarchy of nested structures according to their complexity. The training error decreases as the complexity of the hypotheses increases.

Vapnik-Chervonenkis (VC) dimension is a mathematical description of the complexity of the data[34]. VC dimension is used to measure the complexity of a hypothesis space of

a learning machine and to classify data into separate spaces. When VC-dimension increases, the training error decreases but the risk of overfitting of data increases. SVM is unable to classify data when VC-dimension is infinite.

A simple example is given in Figure 23. On picture (a), all training points can be separated by a hyperplane regardless how the training data are labeled. Once the training data go beyond 3 points, certain labels of the training data are not separable by hyperplanes as shown in picture (b) with 4 points.

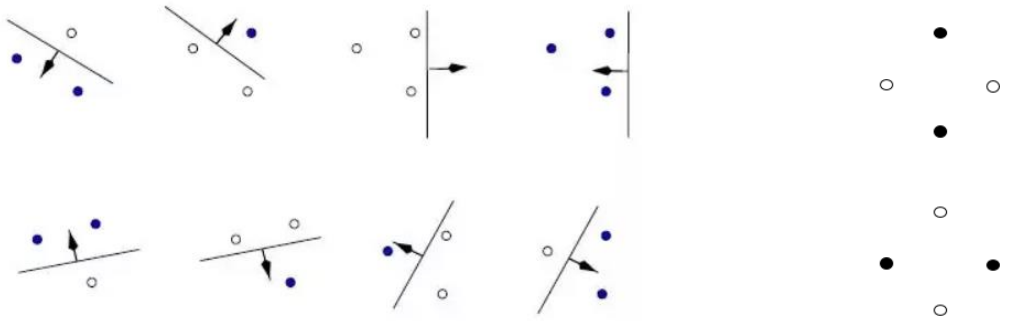


Figure 23: (a) 3-points data that are separable, and (b) 4-points data that are not separable. Source:[71]

The VC-dimension for a set of function $f(x)$ is defined as the maximum number of training points that can be arranged so that $f(x)$ can shatter those points. The theory of VC-dimension suggests that,

$$Testing\ error\ \leq Training\ error + VC\ confidence$$

That is: $$R(w) \leq R_{emp}(w) + \sqrt{h\left(\frac{\log\left(\left(\frac{2N}{n}+1\right)-\log\left(\frac{n}{4}\right)\right)}{N}\right)}$$

In the abovementioned formula, h is the VC-dimension, and N is the number of training samples. When h approaches zero, N approaches infinity.

2.3.9 OVERFITTING

In contrast to the law of large numbers, small samples cannot guarantee the minimization of the actual risk and may lead to the problem of overfitting.

Over-fitting is a scenario of fitting the data more than warranted of fitting the noise in the machine learning process. The data become noise in machine learning, and the noise is harmful to the recognition of the right pattern. The data that creates the noise is an outlier. An outlier is a data point that is distant from other data points of similar observations. An outlier can cause serious problems in statistical analyses. To reduce the noise level, we can use the cross-validation technique.

Figure 24 illustrates an example of an over-fitting dilemma in classifying cats and dogs. The linear classifier (straight line) is achieved with some errors of misclassification. The nonlinear classifier (curly line) is achieved with less misclassification, but the classifier has also become unfriendly data that is unseen. The methods of risk minimization mentioned in the previous chapter can measure the complexity when computing the risk function.

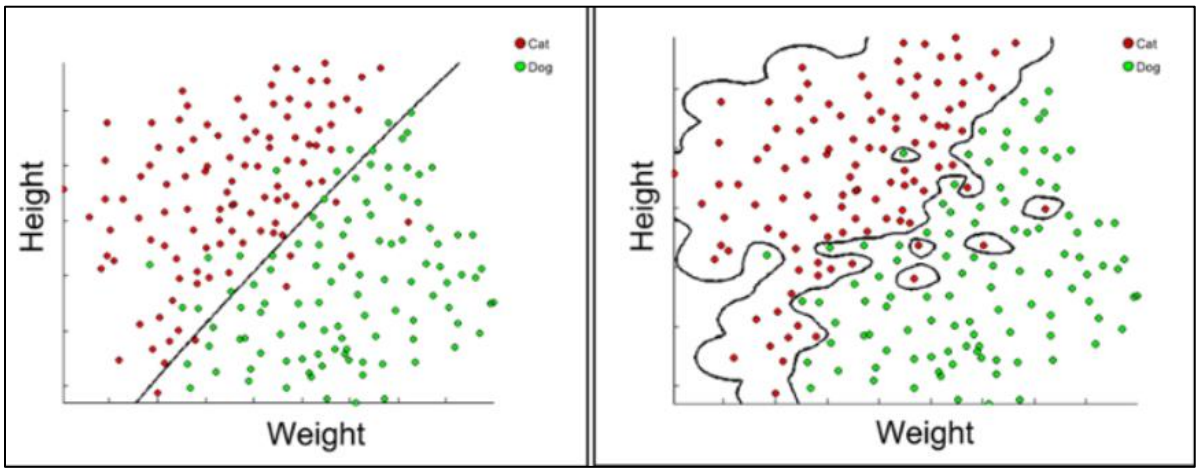


Figure 24: Over-fitting dilemma. Source:[72]

2.3.10 ALGORITHM AND KERNEL

Computer scientist Arthur Samuel introduced the term of machine learning in 1959. He defined machine learning as "*computer's ability to learn without being explicitly programmed*." The ability to learn is the essence of machine learning which evolves from the study of statistical learning theory and computational learning theory. As long as training data are available, a machine or computer with algorithms can learn from the data and predict unseen data. Algorithms are sets of mathematical instructions that are programmed in a computer. The algorithms that enable the computer to learn is called the learning algorithms. With sample data given, the algorithms can learn to build a model to overcome strictly static program instructions by making data-driven predictions or decisions[73].

As mentioned in section 2.1, the supervised learning machine and unsupervised learning machine are two common models in machine learning. SVMs are supervised learning machines with supervised learning algorithms.

For an SVM to transform data from input space to feature space (so that the machine can create a hyperplane for generalization), it requires kernels. The kernel is a way of computing the dot product of two vectors x and y in some (possibly very high dimensional) feature space. A kernel function is also sometimes called "generalized dot product". This research suggests the usage of an SVM as a learning algorithm that uses some common kernels to classify geographic areas with socioeconomic potential for BB investment.

C-SVM classification and nu-SVM classification are two common types of SVM learning algorithms for classification.

- In C-SVM classification, the error function is minimized through training. The larger the capacity constant, the more the error will be penalized. Thus, the capacity constant should be chosen carefully to avoid overfitting.

- In nu-SVM classification, Lagrange construction solves the minimization problem by finding a saddle point. This type of SVM is comparatively difficult to optimize and often takes a longer run time as compared to C-SVM.

This research proposes to use C-SVM classification.

Figure 25 shows the operations of a kernel-based learning algorithm. First, the training data is provided in the input space. The kernel functions will transform the data in a high-dimensional feature space and separate the data according to the data pattern that it recognizes. Finally, the data is classified into one (1) of the two (2) categories (e.g., areas with socioeconomic potential and areas without socioeconomic potential).

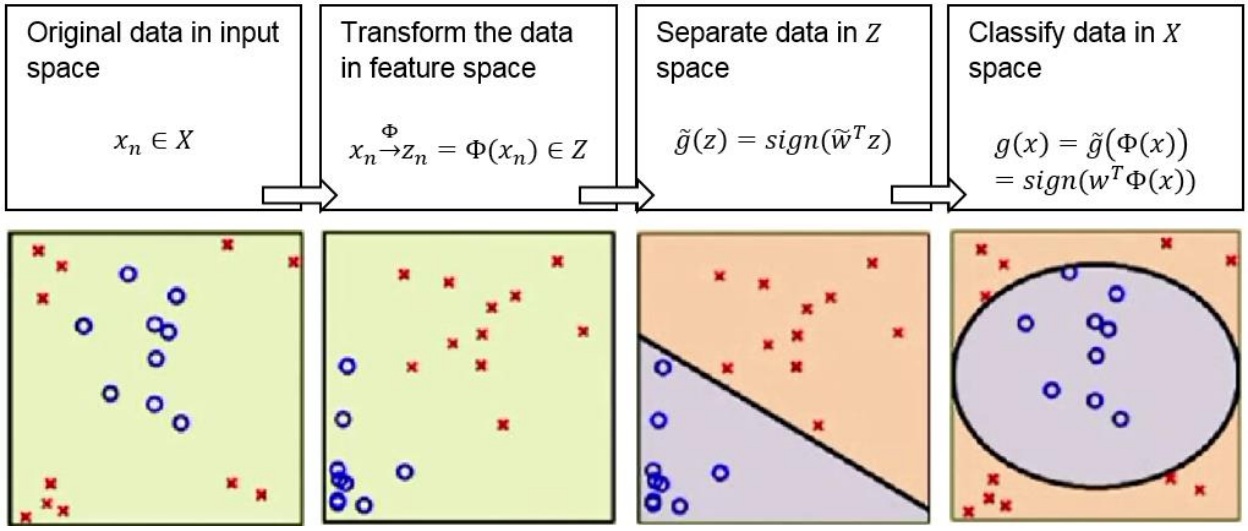


Figure 25: Execution of Kernel-Based Learning Algorithm. Source:[66]

Linear, polynomial, radial basis function (RBF) and sigmoid are the kernels commonly used in SVM[74]. The performance of a support vector machine is very much dependent on its kernels[75]. All that four kernels can be found in LIBSVM which is an integrated software made available online[76]. This research uses LIBSVM as the platform to execute the SVM training and testing.

A kernel is a function k that for all $x, z \subset X$ satisfies the following equation:

$$k(x,z) =< \phi(x), \phi(z) >$$

where, ϕ is a nonlinear (or sometimes linear) map from the input space X to the feature space F, and <·,·> is an inner product.

From the symmetry of the inner product, a kernel must be symmetric:

$$k(x,z) = k(z,x)$$

A kernel shall also satisfy the Cauchy-Schwartz inequality[77]:

$$k(x,z) = \big(\phi(x) \cdot \phi(z)\big) = \big(\phi(z) \cdot (x)\big) = k(z,x)$$

$$\begin{aligned} k(x,z)^2 &= (\phi(x) \cdot \phi(z))^2 \\ &= (\phi(x) \cdot \phi(z))(\phi(z) \cdot \phi(x)) \\ &= k(x,x)k(z,z) \end{aligned}$$

$$(\phi(x) \cdot \phi(z))^2 \leq \big||\phi(x)|\big|^2 \big||\phi(z)|\big|^2$$

Hence $k^2(x,z) \leq k(x,x)k(z,z)$

For all functions $z(x)$, $z(y)$ satisfying the inequality:

$$\int z^2(x)dx \leq \infty$$

A kernel that is symmetric and meets the Cauchy-Schwartz inequality does not guarantee the existence of feature space. Mercer (1909) showed that a kernel must be a positive definite [77]. According to Mercer's Theorem, for any set of training examples $x_1, x_2, \dots, x_N$

and any set of real numbers $\lambda_1, \lambda_2, \ldots, \lambda_N$, the function k must be symmetric for positive definite functions which satisfy:

$$\sum_{i=1}^{N}\sum_{j=1}^{N}\lambda_i\lambda_j k(x_i, x_j) \geq 0$$

Or, $$\int K(x,z)g(x)g(z)dxdz \geq 0$$

A kernel that meets the Mercer's Theorem requirement is fit to be an SVM kernel. A hybrid kernel would still meet Mercer's Theorem requirement.

Symmetric positive definite functions are called covariances, and kernels are essentially covariances. New kernels can be created from existing kernels. First, if k_1, k_2 are two kernels, and a_1, a_2 are two positive real numbers, then:

$$k(x,z) = a_1 k_1(x,z) + a_2 k_2(x,z)$$

The multiplication of kernels k_1 and k_2 will form a new kernel:

$$k(x,z) = k_1(x,z)k_2(x,z)$$

Properties of $k(x,z) = a_1 k_1(x,z) + a_2 k_2(x,z)$ and $k(x,z) = k_1(x,z)k_2(x,z)$ imply that any polynomial with positive coefficients, $pol^+(x) = \{\sum_1^n \alpha_i x^i | n \in \mathbb{N},\ \alpha_1, \ldots, \alpha_n \in \mathbb{R}\}$, evaluated at a kernel k_1, yields a kernel:

$$k(x,z) = pol^+(k_1(x,z))$$

By taking the limit of the series expansion of the exponential function, the kernel will become:

$$k(x, z) = exp(k_1(x, z))$$

If g is a real-valued function on x, then:

$$k(x, z) = g(x)g(z)$$

If ϕ is an $\mathbb{R}^p$-valued function on x, and k_3 is a kernel on $\mathbb{R}^p \times \mathbb{R}^p$, then:

$$k(x, z) = k_3(\phi(x), \phi(z))$$

Finally, if A is a positive definite matrix of size $d \times d$, then:

$$k(x, z) = x^T A z_3$$

Or, in Gram Matrix: $$k = \begin{bmatrix} x_1^T x_1 & x_1^T x_2 & \dots \\ x_2^T x_1 & \dots & \\ \dots & & \end{bmatrix} = XX^T$$

As mentioned earlier, a function $k(x_i, x_j)$ is a valid kernel if the matrix k complies to Mercer's Theorem. This research proposes to use the common kernels in its study. The common kernels are linear, polynomial and Radial Basis Function (RBF).

Linear kernel : $$k(x_i, x_j) = (x_i \cdot x_j)$$

or, $$K(x_i, x_j) = x_i^T x_j$$

Polynomial kernel: $$k(x_i, x_j) = (x_i \cdot x_j + c)^d$$

or, $$K(x_i, x_j) = (\gamma x_i^T x_j + c)^d, \ \gamma > 0$$

Radial Basis Function: $k(x_i, x_j) = \exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right), \quad \gamma > 0,$

If $\gamma = \frac{1}{2\sigma^2}$, then it is called a Gaussian RBF kernel.

2.3.11 TRAINING AND TESTING

The supervised machine learning process involves training, validating and testing. An SVM needs to be trained with the data sets for the classifier to find the best parameters for classification. The process of validating is essential in confirming the best parameters for the classifier. The trained and validated SVM with its optimized parameters can then be applied to a separate test set for generalization.

Generally, the data will be divided into two parts: cross-validation set and testing set. The data for cross-validation set will be used iteratively for training and cross-validation to optimize the parameters of the chosen classifier. The data for testing will be used to validate the accuracy of the generalization performance of the final classifier.

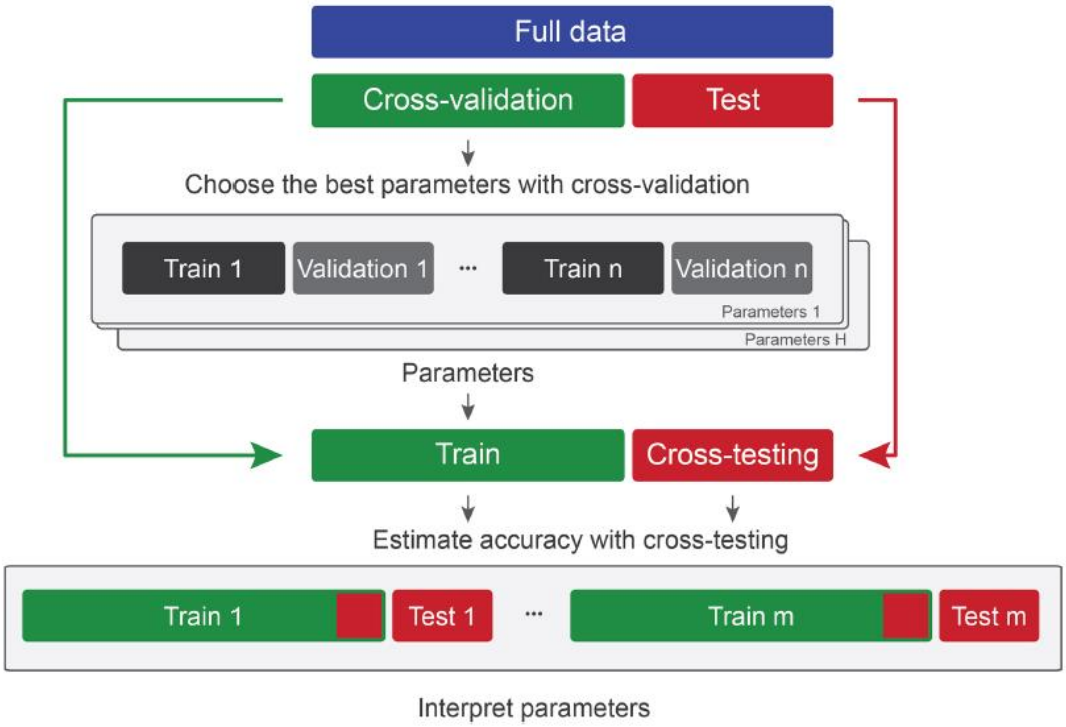


Figure 26: Cross-Validation and Cross Testing. Source:[78]

For example, in X-fold cross-validation, the data are divided randomly into X subsets of equal size. In cross-validation, one of the subsets is used to test the machine which has been trained by X-1 subsets. The process will repeat until every unique subset has been used to test the machine as shown in Figure 27. The cross-validation accuracy is defined as the percentage of data that is correctly classified at the end of the iteration process. Cross-validation process helps the machine to search for optimum kernel parameters.

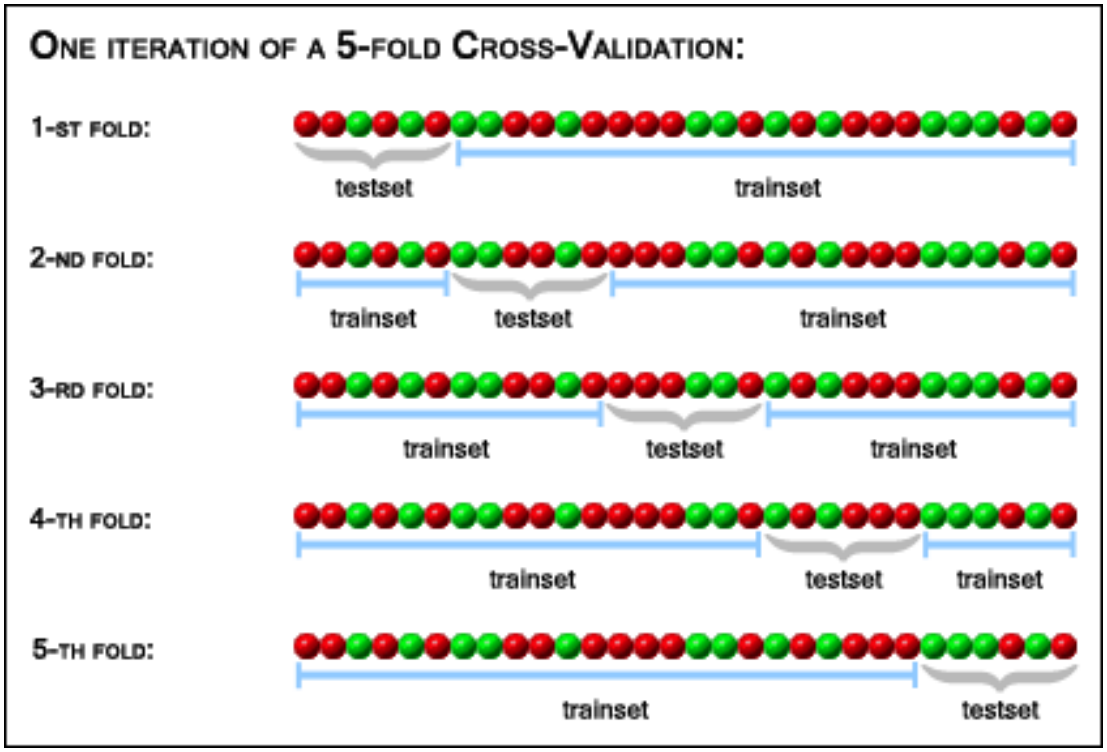


Figure 27: The Case of a 5-Fold Cross-Validation with 30 Samples. Source:[79]

2.3.12 DISADVANTAGES OF SVM

The performance of an SVM is affected by the data that it requires for training. SVM requires large samples for training. However, sometimes real-life data is difficult to obtain results in poor machine performance. This thesis proposes to use GA to generate large virtual samples to overcome the limitation of data availability.

SVM can be sensitive to a small number of mislabeled samples. These mislabeled samples are noises that affect SVM learning and decrease its performance.

Training and testing SVM can be time-consuming. SVM performs with complex algorithms and extensive statistical programming with large samples[80][81].

SVM could be used to perform pattern recognition to solve a problem or generalize a situation for decision making. The optimal kernel functions, hyperparameters and penalty coefficient vary to deal with problems or decision-making of different nature. Kernel, hyperparameters and penalty coefficient will need to be determined during SVM modeling. [82].

Two possibilities may cause the machine to make a wrong classification as elaborated in Table 4 below. A wrong classification is costly and a waste of an opportunity to invest with potential for growth.

Table 4: True and False Classification

Binary Classification	Classification Result	
Class 1: The rural area has the socioeconomic potential for BB investment	True: Correctly classified as a potential area	False: Wrongfully classified as a potential area. [False accept]
Class 0: The rural area is not socioeconomically viable for BB investment	False: Wrongfully classified as a non-potential area [False reject]	True: Correctly classified as a non-potential area

2.4 URBAN AND RURAL AREAS

2.4.1 DEFINITION OF URBAN AND RURAL AREAS

There is no single universal definition of rural areas, and the definition varies from country to country. According to ITU (2014), *countries that have a notable rural/urban divide or a*

strong regional structure may be interested in using a geographic classification; and there is no internationally comparable definition of rural or urban, and countries have their definitions based on the size, density or administrative status of localities[83].

Box 6 and Box 7 show a comparison of the definition of rural areas applied in Malaysia and India.

Box 6: Definition of Rural Areas in Malaysia

Areas with population less than 10,000 people having agriculture and natural resources in which its population is either clustered, linear or scattered. Source: Department of Statistics Malaysia

Box 7: Definition of Rural Areas in India

India: definition of rural and urban areas *The Ministry of Statistics and Programme Implementation in India uses several demographics, administrative and socio-economic variables to define urban and rural areas. Urban areas are defined as (a) all places with a Municipality, Corporation of Cantonment and places notified as town area, (b) all other places that satisfy the following criteria: a minimum population of 5000, at least 75 percent of the male working population is non-agriculturist and (iii) a density of population of at least 400 per square kilometer. However, there are urban areas which do not possess all of the above characteristics uniformly. Certain areas are treated as urban by their possessing distinct urban characteristics, overall importance and contribution to the urban economy of the region. The rural sector covers areas other than the urban areas.* *Source: Ministry of Statistics and Programme Implementation.* http://mospi.nic.in/Mospi_New/upload/nsso/concepts_golden.pdf?status=1&menu_id=49

In Malaysia, MCMC defines underserved areas (Figure 28) as the target area for ICT development through USP funding. BB penetration in underserved areas is below the national BB penetration rate.

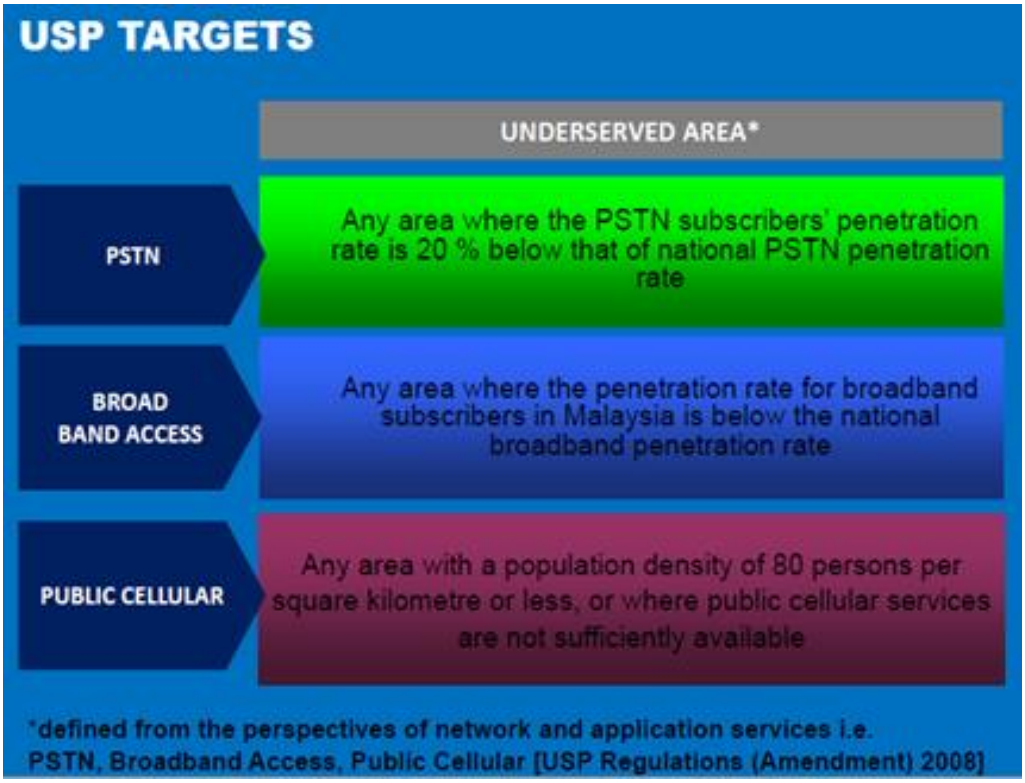


Figure 28: Underserved Areas Defined for USP Targets. Source: MCMC

For USP projects, MCMC defines rural areas into three layers (Table 5) in the context of ICT development[43].

Table 5: Three Layers of Rural Areas as Target Communities for USP Projects

Target	Description	Target Universal Service
Layer 1: Sub-Urban	• An area with a population of more than 50,000 people • Its distance from nearest town is within a 15 km radius • Electricity is available • Basic infrastructure and facilities are available	• Public cellular services • BB community applications • Collective telephony access
Layer 2: Rural	• An area with a population between 5,000 to 50,000 people	• Public cellular services

	• Its distance from nearest town is between a 15 to 50 km radius • Electricity is partially accessible • It is accessible via dirt roads or timber tracks • It is a semi-forest area	• BB community applications • Collective telephony access
Layer 3: Remote	• An area with a population of fewer than 5,000 people • Its distance from nearest town exceeds a 50 km radius • Grid electricity is not available • Backhaul is not available • It is accessible via jungle tracks or river	• Public cellular services • Collective telephony access

2.4.2 CLASSIFICATION OF AREAS BY ECONOMIC SITUATION

World Bank classifies the economic situation of each country into one of the four categories:

- High-income economies
- Upper-middle-income economies
- Lower-middle-income economies
- Low-income economies

In the telecommunication industry, it is a common practice for TELCOs and government agencies to define the geographic areas as one of the four categories:

- Urban areas
- Sub-urban areas
- Rural areas
- Remote areas

There seems to be a logical analogy between the terminologies used by the World Bank and the practice of TELCOs.

Table 6: Analogy of worldview and country-view of economic states

Country Economy Situation Defined by Income Level (World Bank)	*Geographic Area for BB Development (Industry Practice)*
High-income economies	Urban Areas
Upper-middle-income economies	Sub-Urban Areas
Lower-middle-income economies	Rural Areas
Low-income economies	Remote Areas

For this research, the assumption is that the size of the geographic area is not limited. Hence, the life data available from a worldview or country-view may be applicable, depending on whichever is available.

2.4.3 PERFORMANCE INDICATORS OF SOCIOECONOMIC AND BROADBAND

There are many key performance indexes (KPI) that are used by an industrial analyst, governmental and international bodies to measure the value of BS, social improvement, and economic growth. Table 7 provides some examples of those KPI.

Notes: In this section, words that appear in *Italic red* are partial features used in research experiments.

Table 7: Example of KPI to measure socioeconomic and digitalization

Social KPI	Economical KPI	Digitalization KPI
Better Life Index	Purchasing Power	Broadband Composite Index (BCI)
Quality of Time	Basic Income Growth	$ per Mbps
Convenience	Saving	% of $ per Mbps over GNI
Productivity	Investment Rate	Network Readiness Index (NRI)
Life Enrichment	GDP Growth – Income approach	The Global Gender Gap Index (GGGI)
Unemployment Rate	GDP Growth – Product approach	Digital Opportunity Index, DOI
Happiness	GDP Growth – Expenditure approach	Digital Effectiveness Framework (DEF)
Fatality	*GNI / per capita*	Broadband quality
Years of Education	Poverty Index	Mobile quality
% of Knowledge Workers	*Land Utilization*	Internet Support Method
Gender Equality	Inflation Rate	Network coverage
Health Index		*Broadband diffusion*
Human Development Index		Broadband affordability
Aging Phenomenon		Internet Barriers Index

The socioeconomic indicator used in this research is GNI per capita which is explained in section 2.3.4 on data and data sets.

2.4.4 THE SOCIOECONOMIC IMPACT OF BROADBAND INVESTMENT

Socioeconomic is a big word, and its measurement can be very complex. However, *no single indicator can provide a complete picture of what the economy is up to, nor is there*

a simple combination of measurements that provide a connect-the-dots path to the future[84]. Specifically for socioeconomic measurement about ICT development, Strategy& (2015)[21] proposed a framework for measuring digitalization impact on the socioeconomic as shown in Figure 29.

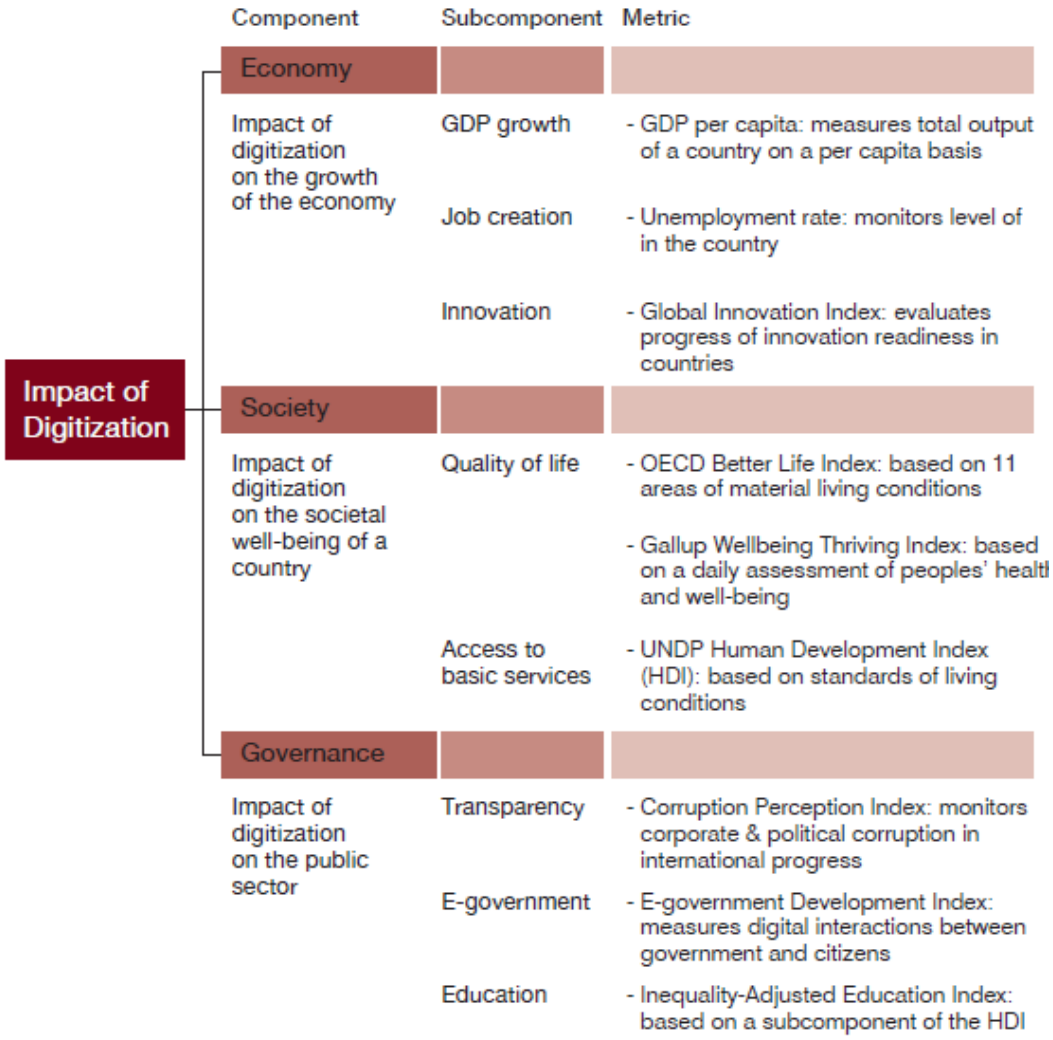


Figure 29: Digitization impact on the socioeconomic. Source:[21]

The important role of ICT in the economy and also the magnitude of its positive impact differs across countries with different economic situations. The OECD reported that developed countries benefit more than developing countries and LDCs from ICT development due to economic structure and policy issues[14].

In their respective studies, Koutroumpis (2009), Czernich et al. (2011) and Qiang et al. (2009) found that BB adoption improves economic growth. Figure 30 illustrates an example of findings from past research[85].

Qiang and Rossotto (2009) found that GDP per capita increases for every 10% increase in ICT penetration[6]. Nevertheless, the increase of GDP per capita differs in a different region as shown in Figure 30.

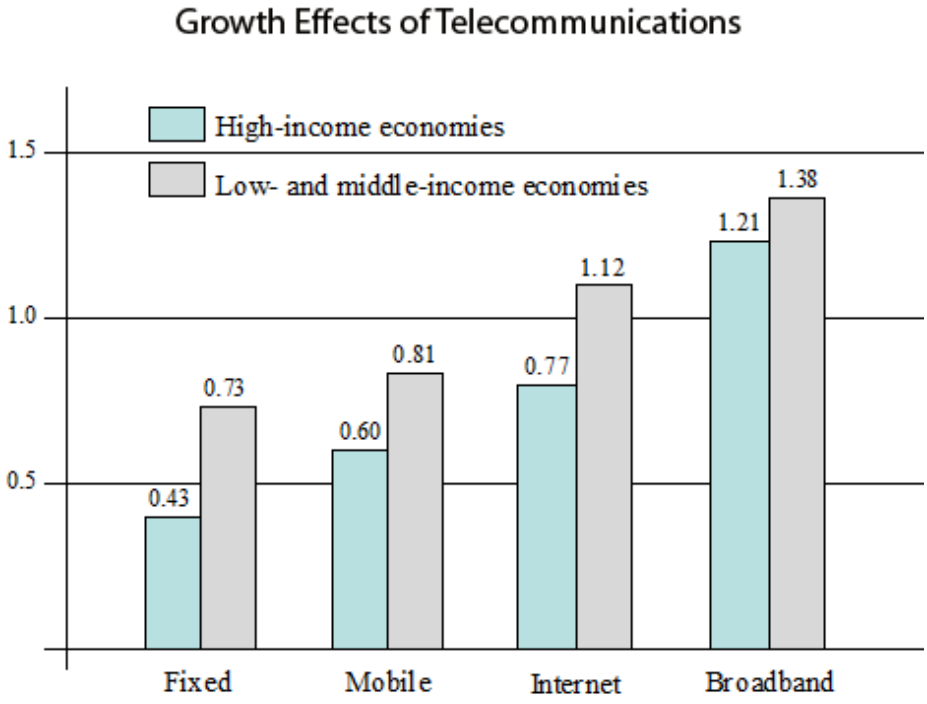


Figure 30: Growth effects of telecommunications. Source:[6]

In 2012, McKinsey[11] found that the Internet contributes an average of 1.9 percent of GDP in aspiring countries in 2010. The Internet in developed countries contributes an average 3.4 percent of GDP. In 2014, McKinsey[86] suggested that not all countries have utilized the Internet's benefits to the same extent. The report studied the progression of Internet adoption worldwide, the factors required to enable the growth of a dynamic Internet ecosystem, and the barriers faced by a large proportion of the world population from accessing the Internet.

Irawan (2003) summarized that there are at least three key points that can be learned from the previous findings regarding ICT and the country's economic performance. Firstly, countries that are more developed are expected to benefit greater when compared to less developed countries. Secondly, the impact of ICT is dependent on the intensity to which ICT is utilized. Thirdly, both the size and structure of the ICT sector in a country's economy affects its economic performance[14].

Regarding correlating ICT development to socioeconomic impact, the conducted researches can be summarized into six types according to their findings.

Type 1: ICT development has a positive impact on socioeconomic status. There is a clear, positive correlation between telecommunication and economic growth.

- *Jipp (1963), Hardy (1980), Moss (1981), Cronin et al. (1991), Norton (1992),* Lau and Tokutsu (1992), *Saunders et al. (1994), Dholakia and Harlam (1994), Lichtenberg (1995), Greenstein and Spiller (1996), Madden and Savage (1998), Jorgenson and Stiroh (2000),* Dewan and Kraemer (2000), *Dutta (2001), Röller and Waverman (1996, 2001), Colecchia and Schreyer (2002), Chakraborty and Nandi's (2003), Cieslik and Kaniewsk (2004), Yoo and Kwak (2004), Datta and Agarwal (2004, Waverman, Meschi and Fuss (2005), Orbicom-ITU (2005),* Mas and Quesada (2005), *Bakhshi and Larsen (2005), Campos (2006),* Walsham and Sahay (2006), *Wolde-Rufael (2007),* Kuppusamyand Shanmugam (2007)*, Jalava and Pohjola (2007), Welfens (2008),* Kuppusamy, Pahlavani and Saleh (2008), Hosman (2008), *Shiu and Lam (2008), Stenberg and Morehart (2008), Qiang and Rossotto (2009), Koutroumpis (2009), Venturini (2009),* Kottemann and Boyer-Wright (2009), *Seo et al. (2009),* Kuppusamy, Raman & Lee (2009), *Ericsson (2010), Czernicz et al. (2011), Gruber and Koutroumpis (2011), Cortez and Navarro (2011), Vu (2011), Ahmed and Ridzuan (2012), Booz & Co (2012), Ericsson (2010, 2013),* Katz and Koutroumpis (2012, 2013), *McKinsey (2012, 2014),* Katz & Berry (2014)

Type 2: ICT development has different levels of impact on different markets. The magnitude of the positive impact of ICT differs across countries.

- *Röller and Waverman (1996, 2001),* Dewan and Kraemer (2000), *OECD (2004), Gruber and Verboven (2001), Daveri (2002), Bassanini and Scarpetta (2002), Jalava and Pohjola (2002), Waverman, Meschi and Fuss (2005),* Kuppusamy, Pahlavani and Saleh (2008), *Shiu and Lam (2008), Seo et al. (2009), Booz & Co (2012), Ericsson (2010, 2013),* Katz and Koutroumpis (2012, 2013), *McKinsey (2012, 2014)*

Type 3: Economic growth accelerates ICT development for developed markets. For under-developed markets, BB penetration has no significant impact on socioeconomic growth. In this case, ICT development has less impact on developing countries and LDCs. In other words, the developing countries and LDCs benefited less than developed countries from the development of ICT. For developing and under-developed market, a relationship from GDP to ICT development is unidirectional.

- Dewan and Kraemer (2000), Kridel et al. (2001), Crandall et al. (2002), *OECD (2004), Chakraborty and Nandi's (2003),* Garcia-Murillo (2005), Choudrie and Dwivedi (2006), *Karner and Onyeji (2007), Cieslik and Kaniewsk's study (2004),* Prieger and Hu (2008), Goldfarb and Prince (2008), Richard Cadman and Chris Dineen (2008), Kuppusamy, Pahlavani and Saleh (2008), *Stenberg and Morehart (2008), Shiu and Lam (2008), Ericsson (2010, 2013), McKinsey (2012, 2014),* Katz & Berry (2014)

Type 4: ICT development accelerates socioeconomic growth. Tele-density (adoption) increased with wealth and productivity and welfare. The higher the BB and ICT adoption, the more important the economic and social benefits. [Note: The relationship from ICT development to GDP is unidirectional for a developed market.]

- *Jipp (1963),* Lau and Tokutsu (1992), *Röller and Waverman (1996, 2001), Kathuria, Madden and Savage (1998), Jorgenson and Stiroh (2000),* Dewan

and Kraemer (2000), *Dutta (2001), Röller and Waverman (2001),Colecchia and Schreyer (2002), Chakraborty and Nandi's (2003), Cieslik and Kaniewsk (2004),* Mas and Quesada (2005), *Orbicom-ITU (2005), Campos (2006),* Walsham and Sahay (2006), Kuppusamy, Pahlavani and Saleh (2008), Hosman (2008), *Shiu and Lam (2008),Welfens (2008),Qiang and Rossotto (2009), Koutroumpis (2009), Venturini (2009), Uppal and Mamta (2009),* Kuppusamy, Raman & Lee (2009), *Booz & Co (2012), Ericsson (2010, 2013), Czernicz et al. (2011), Gruber and Koutroumpis (2011),* Katz and Koutroumpis (2012, 2013), *McKinsey (2012, 2014),* Katz & Berry (2014)

Type 5: ICT development (under certain circumstances) has a more significant impact in developing countries than in developed countries. For developing markets, BB penetration accelerates socioeconomic growth faster than the developed market.

- *Waverman, Meschi, and Fuss (2005), Qiang and Rossotto (2009)*

Type 6: ICT development accelerates socioeconomic growth, and socioeconomic growth accelerates ICT development (but NOT in a reversed relationship). Only the developing and developed markets are commercially viable for private investments and not the under-developed markets. The underdeveloped market shall require government support in BB development. In this case, telecommunication investment enhances economic activity and growth while economic activity and growth stimulate demands for telecommunications infrastructure investment. [Note: Relationship from ICT development to GDP is bidirectional for a developed market.]

- *Cronin et al. (1991), Chakraborty and Nandi's (2003), Yoo and Kwak (2004), Wolde-Rufael (2007),* Kuppusamyand Shanmugam (2007)*, Shiu and Lam (2008), Ericsson (2010, 2013)*

In summary, BS will stimulate socioeconomic growth, and socioeconomic growth will accelerate further development of BS. This bidirectional relationship applies to developed

or developing areas. Past researches commonly used regression analysis with historical data to derive research outcomes. For decision making on future BB investment, the ROI model is suitable for network planning in the developed and developing areas.

In contrast, this thesis proposes to use a support vector machine to study the current geographical data of certain geographic areas, and classify the areas according to their socioeconomic potential for BB investment. If the current state of a geographic area is classified to be a non-potential area, the curve-fitting technique used in the framework of this machine learning technique can be used to suggest policies to influence the features of targeted geographic areas to increase their socioeconomic potential. Nevertheless, this research focuses its study on the intelligence of machine classification.

2.5 BROADBAND

Box 8: The Birth of Broadband

Many people associate BB with a particular speed of transmission or a certain set of services, such as digital subscriber loop (DSL) or wireless local area networks (WLANs). However, the definition of BB constantly evolves as the BB technologies continue changing. Nowadays, the term BB often describes recent Internet connections that range from 5 times to 2000 times faster than earlier Internet dial-up technologies. However, the term BB does not refer to either a certain speed or a specific service. BB combines connection capacity (bandwidth) and speed. Recommendation I.113 of the ITU Standardization Sector defines BB as a "transmission capacity that is faster than primary rate Integrated Services Digital Network (ISDN) at 1.5 or 2.0 Megabits per second". Source: ITU (2003)

Broadband services (BS) can be provided through a range of telecommunication technologies (e.g., cable, telephone wire, fiber, satellite, wireless) that gives users the capability to send and receive data at large volumes and speeds which are far greater than the traditional "dial-up" Internet access over telephone lines[87]**.** Different BB technologies deliver BS at different speeds for use in various applications.

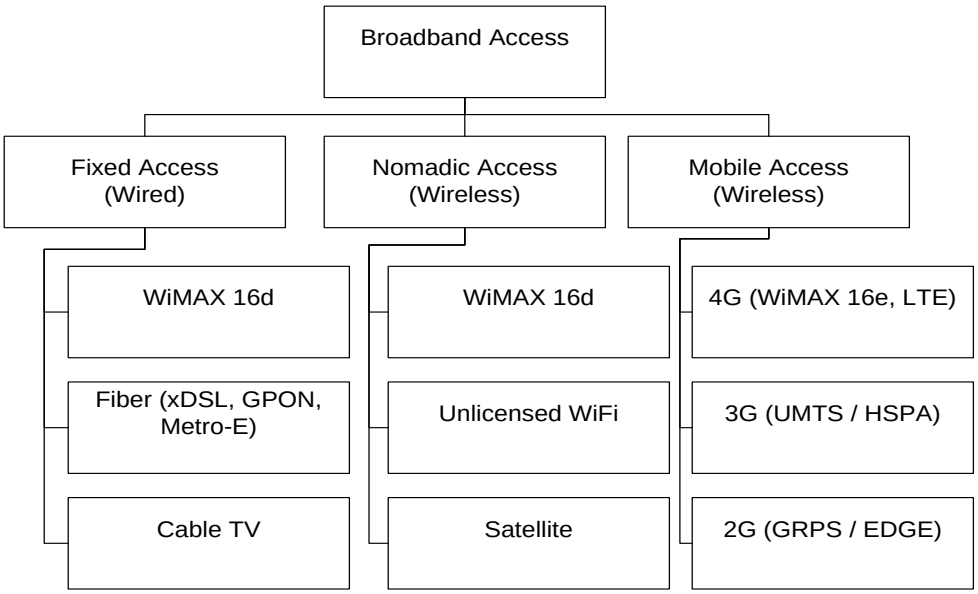


Figure 31: Telecommunication technologies for BS. Source: the author

There is no standard definition of BB that is agreeable internationally. The ITU defines the minimum speed for BB connection is 256 kb/s or higher, whereas OECD defines as "not being dial-up" (OECD 2013). The speed of BB connection has been found to have variable economic impacts[88], and the bandwidth rate for BB has been increasing ten folds every 5 to 7 years since the 1980s (Figure 32).

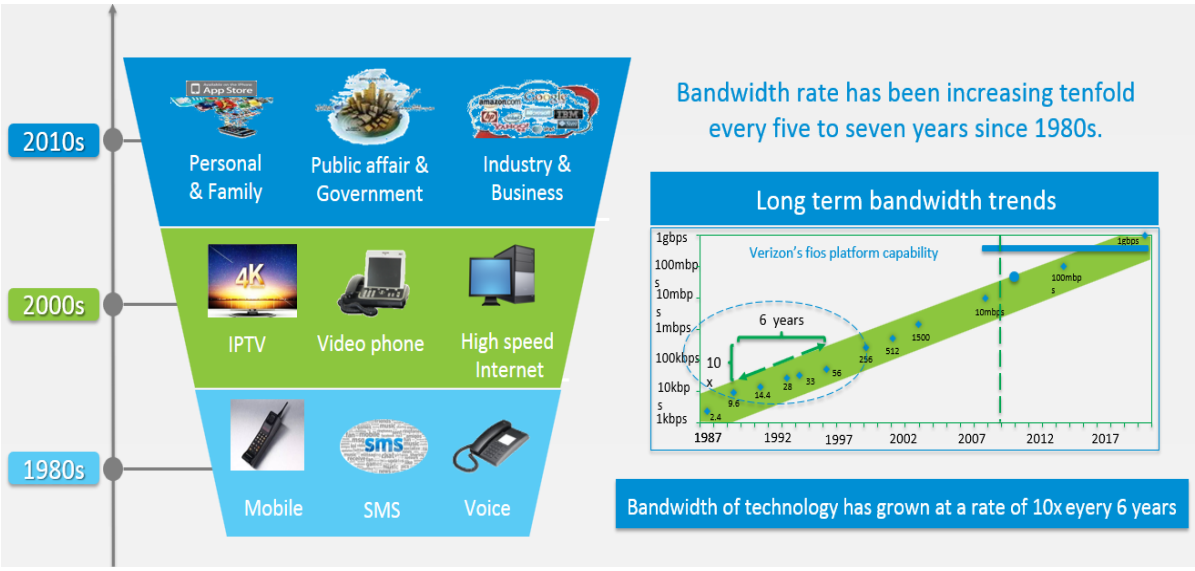


Figure 32: Bandwidth increments since the 1980s. Source: ZTE

With regards to mobile services, there are various wireless BB technologies as shown in Figure 33.

Technology	Network Theoretical Data Transfer Rates	User Typical Data Transfer Rates
2G- GSM (early 1990s)	9.6 – 115 kbps	About 10 kbps
2.5G-GPRS (2001)	9.6 - 171.2 kbps	Between 30-50 kbps
2.5G- EDGE (2003)	9.6 -384 kbps	Between 75-135 kbps
3G- UMTS (Release 99, 2001)	144 kbps - 2 Mbps	Between 200-300 kbps
3.5G-HSPA (Rel. 7, 2007) (HSDPA , Rel. 5, 2005) (HSUPA, Rel. 6, 2008)	DL: 14.4 Mbps UL: 5.76 Mbps	DL : 1-4 Mbps UL : 500Kbps -2Mbps
HSPA+ (Rel. 7, 2007)	DL: 21.6 Mbps UL: 11.5 Mbps	DL: ~2 - ~9 Mbps UL: 1-4 Mbps
HSPA+ (DL: 64 QAM, UL: 16 QAM, Dual Carrier, 10+5 MHz)	DL: 42 Mbps UL: 11.5 Mbps	DL: 3.8 – 17.6 Mbps UL: 1-4 Mbps
3.5G- Mobile WiMAX (IEEE 802.16e, 2005)	DL: 46 Mbps UL: 4 Mbps	DL: UL: 1 – 5 Mbps
3.9G LTE (Rel. 8, 2008)	DL: 300 Mbps (20MHz) UL: 71 (20MHz)	DL : 6.5 - 26 Mbps (10MHz) UL : 6.0 – 13.0 (10MHz)
4G-LTE-Advanced (Rel. 10, 2010)	DL: 1.2Gbps UL: 568 Mbps	TBD
4G- WirelessMAN-Advanced (IEEE 802.16m, 2010)	DL: >1Gbps UL: >100 Mbps	TBD
5G by 2020?	TBD	TBD

Figure 33: Wireless Technologies and Data Rates. Source:[89]

Both fixed and wireless BB technologies are fast evolving. Therefore, policymakers are constantly under pressure to embrace the technology evolution. *Key issues such as spectrum allocation, licensing, infrastructure sharing, technical standards, interconnection rules and termination rates all come into play*[90].

This research is neutral to all BB technologies without any preferences. Whenever broadband services are mentioned, the services could be provided with any telecommunication technologies applicable.

2.5.1 THE IMPORTANCE OF BROADBAND

Broadband services are crucial in the 21st century as people depend on it more and more in their daily lives. The use of BB today is probably as vital as electricity was in the 1930s[91].

Organizations at all levels - global, regional, national, communal and corporate, believe that broadband services are a key stimulant that accelerates socioeconomic growth. Subsequently, countries across the continents have implemented national BB initiatives to improve their socioeconomic statuses and competitiveness in the world market. Hence, the telecommunication service companies (TELCOs) in all countries, including Malaysia, have been depending on the development of telecommunications technologies to obtain their return on investment on different scales.

Figure 34 provides some examples of national BB strategies adopted by certain countries in different continents.

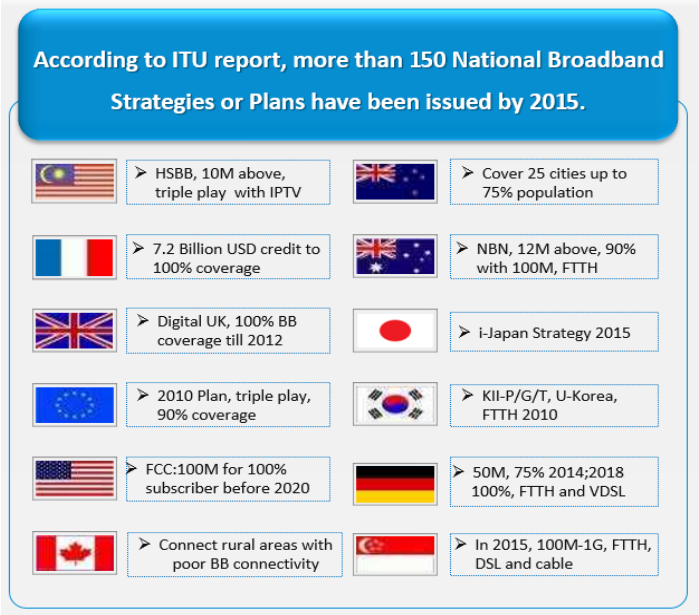


Figure 34: National BB Strategies in Selected Countries. Source: ZTE

In Malaysia, various national BB initiatives have been implemented with the aim to enable the efficient exploitation of BB technology to initiate economic activity in rural Malaysian

villages. Enhancing BB connectivity is one of the key initiatives in the10th Malaysia Plan. McKinsey (2012) reported that the Internet's impact on the Malaysian economy is among the highest of the countries studied[11].

The Malaysian government has formulated a range of policies and plans to guide the management of national development under the New Economic Policy (NEP) which has evolved to become the National Development Plan (NDP) and again evolved into the National Vision Plan (NVP). The NEP and NDP in the 20th. Century were the core national policies[92] based on a philosophy of growth with equitable distribution along with policies nurturing national unity as the goal of development. A two-pronged strategy is aimed at (1) the eradication of poverty and (2) the restructuring of society. Before the turning point to the 21st. Century, the government established the Communications and Multimedia Act 1998 to provide for and to regulate the converging communications and multimedia industries, and for incidental matters[93]. Some of those significant and impactful ICT initiatives are the National Broadband Initiatives and Malaysia National Broadband Implementation Strategy (Figure 35) and Digital Lifestyle Malaysia.

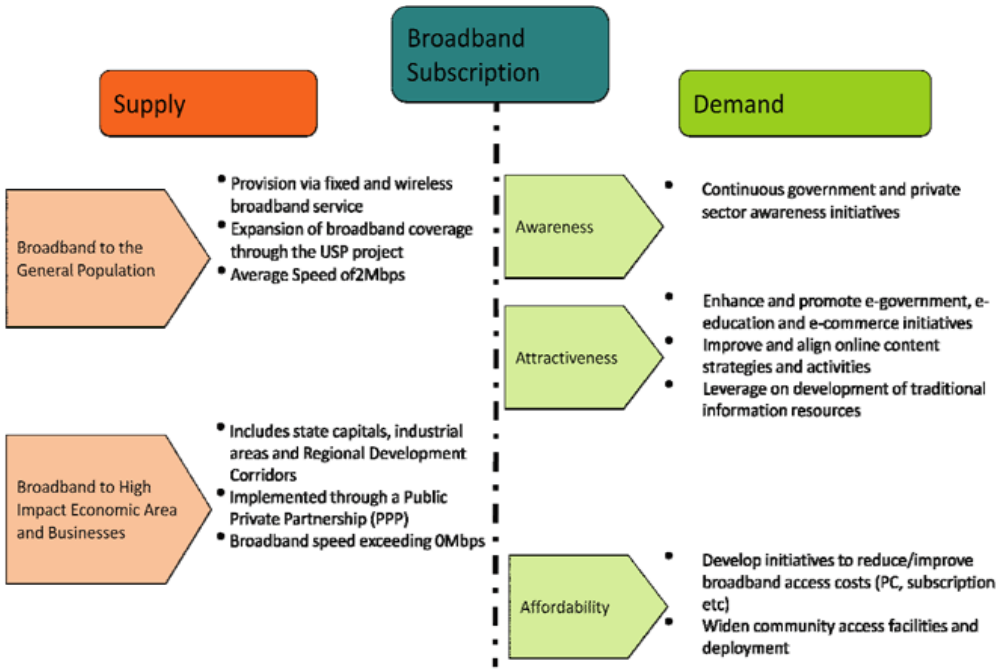


Figure 35: Malaysia National Broadband Implementation Strategy for NBI. Source: MCMC

In South East Asia and Oceania, a region that is closer to Malaysia, urbanization will continue to drive the ICT industry[94]. The 10-members ASEAN countries signed the first e-ASEAN Framework Agreement in the fourth ASEAN Informal Summit[95]. The e-ASEAN agreement has four main objectives, namely:

i. *to promote cooperation to develop, strengthen and enhance the competitiveness of the ICT sector in ASEAN;*
ii. *to promote cooperation to reduce the digital divide within the individual ASEAN Member States and amongst the ASEAN Member States;*
iii. *to promote cooperation between the public and private sectors in realizing e-ASEAN;*
iv. *to promote the liberalization of trade in ICT products, ICT services, and investments to support the ASEAN initiative.*

The European Union has “A Digital Agenda for Europe”[96] which aims to expand BB access for all by 2013 and access to much higher Internet speeds (at least 30 Mbps) by 2020. European Commission (2015) reported that basic BB is available to everyone in the EU, Next Generation Access (NGA) networks, offering speeds above 30Mbps, cover 68% of households. However, rural coverage remains significantly lower, especially in NGA[97].

In the United Nations Millennium Summit (2000), 189 world leaders collectively committed to adopt and agree to the Millennium Development Goals (MDGs)[98], comprising of 8 goals, 18 targets, and 48 indicators. The 8th goal is to develop a global partnership for development, and to make available the benefits of new technologies, especially regarding information and communication.

The motivations of having a national BB policy could vary from country to country. Nevertheless, the motivations could fall into one or more perspective as tabled below.

Table 8: Different Motivations of Having National BB Policy

Perspective	Possible Benefits

Political	• Promote BB availability through:
	  ○ funding or incentives to reduce TELCOs' CapEx
	  ○ spectrum farming to improve service quality
	  ○ development of infrastructures on sharing basis
	• Promote BB adoption through:
	  ○ improving the value of BB access with more contents and applications
	  ○ improving BB affordability by reducing the cost of BB access

Economic	• Create new market opportunities for business owners and consumers • Improve market competitiveness • Increase job opportunities
Social	• Reduce the digital divide between urban and rural areas • Extend universal access to unserved or under-served people • Improve the quality of public service
Technological	• Improve national building with advanced technologies • Promote creativity and innovation in the telecom industry
Environmental	• Promote green ICT technologies • Promote green environment with the use of BS
Legal	• Regulate policies to control BB implementation • Establish telecommunication laws to govern the development of BS

(*) Source: Florence School of Regulation (2011)[99], modified by the author.

The ITU Handbook[100] published 80 internationally agreed indicators to help ITU and countries to track ICT developments with standard indicators. A large proportion of the indicators are related to the Internet and BB development, again showing the significance of Internet/BB as global and national agenda.

Intel, a multinational corporation and technology company has deliberately explained the importance of BB in its World Ahead program, which summarizes that "*Broadband access and ICT networks enable delivery of information, goods, and services that stimulate economic growth and help domestic businesses compete. Without such access, remote communities risk becoming increasingly marginalized and lacking in essential educational, medical, government, e-commerce, and social services. Deploying broadband services provides numerous benefits to developing countries, particularly in rural and remote regions*."[101]

2.5.2 THE DIFFERENT STAGES OF BROADBAND DEVELOPMENT

Once a geographic area has been identified for BB deployment, the BS can be developed through four stages: BB availability, accessibility, adoption and sustainability (Figure 36).

The availability, accessibility, adoption, and sustainability (sometimes equivalent to affordability) are common words used in the telecom industry to explain the state of BB development. These words were commonly found in both commercial and academic literatures[102][103][104][105][106][83].

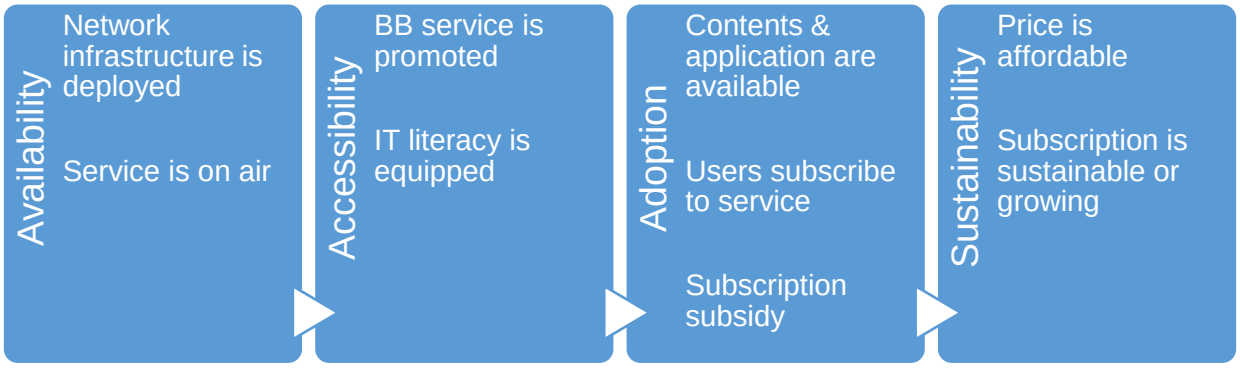


Figure 36: Stages of broadband development of a geographic area

Typically, BB availability and accessibility are the on the supply-side; whereas the adoption and sustainability/affordability are on the demand-side. According to the

Florence School of Regulation (FSR), policy to promote BB availability should come before the policy for BB adoption[99].

According to McKinsey (2014), the four categories of barriers facing the offline population around the world are incentives, low incomes and affordability, user capability, and infrastructure (Figure 37). The infrastructure category limits the BB availability issue. The category of user capability and incentives limits the BB accessibility issue whereas the category of low incomes and affordability are seen as a limitation to the BB adoption issue.

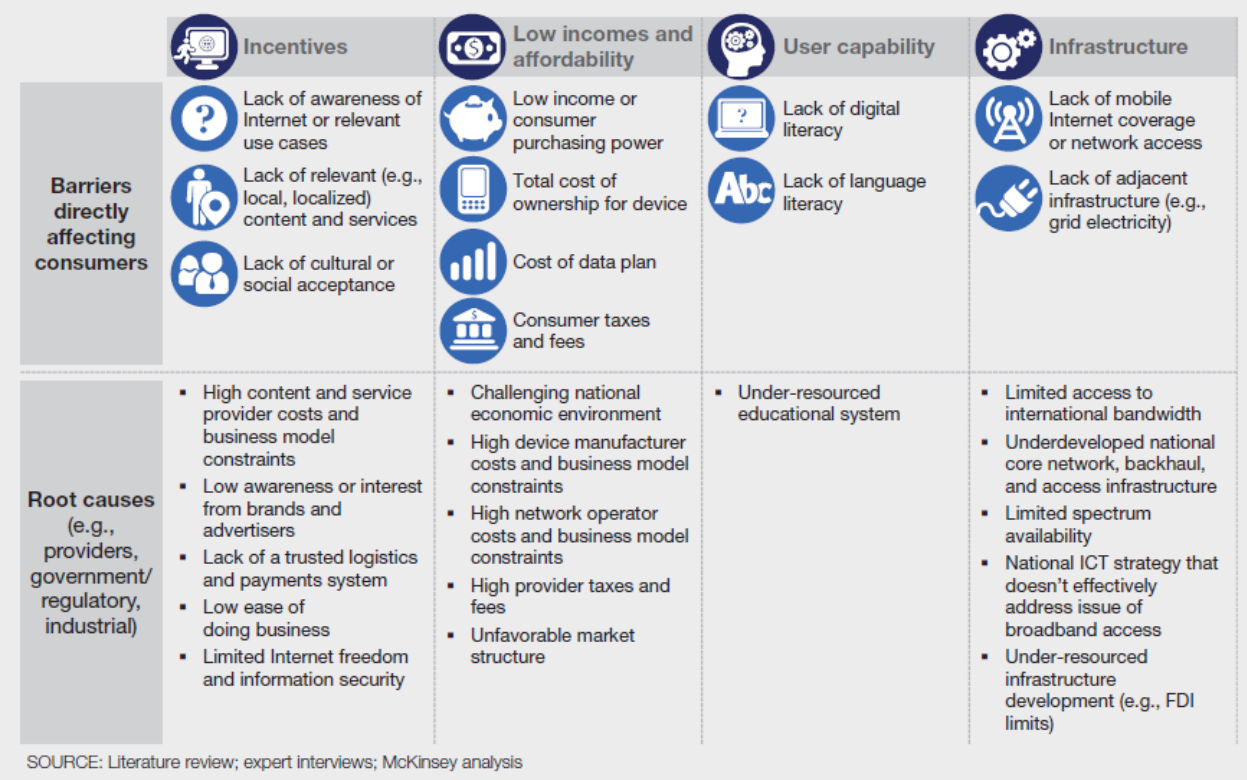


Figure 37: Non-Internet users face four categories of barriers. Source:[86]

According to the ITU (2016), the four key reasons for lack of connection to the Internet are lack of infrastructure, non-affordability, lack of skills and lack of digital content[107]. It is natural for a country's government to start the investment to build the infrastructure and subsidize the subscription while the residents in rural areas are benefiting from the BS, including improvement of skills. As long as the selected areas have a socioeconomic potential, the government initiative could be bridged with TELCO investments as the

affordability improves.

2.5.2.1 BROADBAND AVAILABILITY

BB availability exists through the deployment of network coverage to make BS available. BB availability is the initial BB investment to enable users to connect to the world of the Internet. The key elements of BB availability are the geographic areas and network service coverage.

In 2013, Ericsson, Arthur D. Little and Chalmers University of Technology found that both BB availability and speed drive growth in an economy[86]. It was also found that there is a bidirectional relationship between telecommunications development and economic growth for high-income countries in Europe. However, for countries at lower income levels, the relationship is in unidirectional where economic growth stimulates telecommunications development.

Federal grants may be the most effective when they stimulate private sector competition and are paired with BB accessibility programme such as community education efforts[108]. BB availability and speed alone will not drive growth in geographic areas with lower income. This thesis aims to prove that the socioeconomic potential of certain geographic areas can be classified according to their geographical features. Subsequently, the empirical model used in the research can suggest policies to influence the features of target geographic areas to increase their socioeconomic potential. Ultimately, the initial government funding to create BB availability will be bridged with BB investment by the TELCOs to sustain the BS.

2.5.2.2 BROADBAND ACCESSIBILITY

BB accessibility focuses on creating awareness and providing devices, applications and quality services for users to use the BS. BB access enables users to experience online applications and to realize the benefits to their socioeconomic growth. The key elements

of BB accessibility include, but are not limited to awareness of service, the presence of the school, number of children at home, computer accessibility, education level, IT skills, occupation, gender, age, ethnicity and so forth.

Stenberg and Morehart (2008) found that education programs drive household demand for Internet access[18]. B. Whitacre et al. (2015) recommended that rural households with lower education and income level need further attention with the support of policies that focus on driving demand rather than just BB availability[109].

2.5.2.3 BROADBAND ADOPTION

The BB adoption gap refers to the areas whereby BS are available and accessible, but face a lack of subscribers to the available services.

FSR (2011) concluded that the growth of BB penetration is the complex outcome of many factors such as income, location, education, family size, individuals' characteristics, market structure, technological endowments, regulatory actions, public policy interventions and so forth. The findings encompass PESTEL factors mentioned in Chapter 1. FSR (2011) argued that BB adoption is dependent upon both demographic and socioeconomic factors. The younger age, high-income and educated population have a strong correlation with Internet adoption. Social interactions such as peer pressure and peer influence also affect the probability of Internet adoption.

GAO (2006) highlighted that households in rural areas are more unlikely to subscribe to BS as compared with households in urban areas[110]. Regarding BB deployment in rural areas, GAO summarizes that the five influencing factors are market, technical, federal and state government efforts as well as access to local resources. GAO recognizes that the most important factor is the market whereby the demand in rural areas is low (due to population size) and also costlier for the private TELCO to build, operate and serve BS. GAO believes that government assistance will help BB implementation in rural areas. Once BB services are implemented, the factors that will improve subscriber adoption will

be higher income level, higher education level, lower price barrier and availability of contents and application.

Roller and Waverman (2001) found that the higher the BB and ICT adoption, the more important the economic and social benefits are[3].

Booz & Co (2013) reported that a 10 point increase in the digitization score leads to a 1.02 percent drop in the unemployment rate[8]. On the other hand, McKinsey (2014) reported that countries with low Internet penetration tend to have multidimensional bottlenecks and require all stakeholders in the Internet ecosystem to overcome those bottlenecks[86].

Rappoport et al. (2002) found that a 10% price reduction for BB would increase BB adoption by 14.91% in the United States. Lee et al. (2011) found that a 10% price reduction of BB results in a 15.80% increase in penetration[12]. Richard Cadman and Chris Dineen (2008) found that a 1% decrease in price would lead to a 0.43% increase in demand[111]. It looks like any price incentive offered in a PPP project would help to improve BB adoption in rural areas.

The factors that influence BB adoption vary at a different stage of BB development. Katz and Berry (2014) found that BB coverage is the key factor at the initial stage of BB development. As the BB availability expands, affordability becomes the most important to drive BB adoption. Limited affordability is critical to ensure BB adoption when BB penetration hits between 3 and 20%. The importance of affordability declines when BB penetration is high[23]. Table 9 shows the influencing factors in correlation with BB adoption at different stages of BB development.

Table 9: Stages of broadband adoption. Source:[45]

	Stage 1	Stage 2	Stage 3
Broadband population adoption	=< 3%	3-20%	>20%
Ownership of access devices (e.g., computers, smartphones)	Low adoption	Medium adoption	High adoption
Availability of web applications and services	Very low	Limited	High
Factors driving non-adoption	Service coverage	Affordability	Digital literacy Cultural relevance

2.5.2.4 BROADBAND AFFORDABILITY AND SUSTAINABILITY

BB sustainable areas are those areas whereby BB services are not just available and accessible, but the users also find the service to be affordable leading to sustainable subscription or even growth.

BB affordability is about providing services at a price which users can and are willing to pay to use the e-services via the Internet. The affordability is to increase BB adoption which will, in turn, accelerate socioeconomic growth. The key elements of BB sustainability are BB price, cost of device ownership and income level. Figure 38 shows the comparison of BB affordability in different regions.

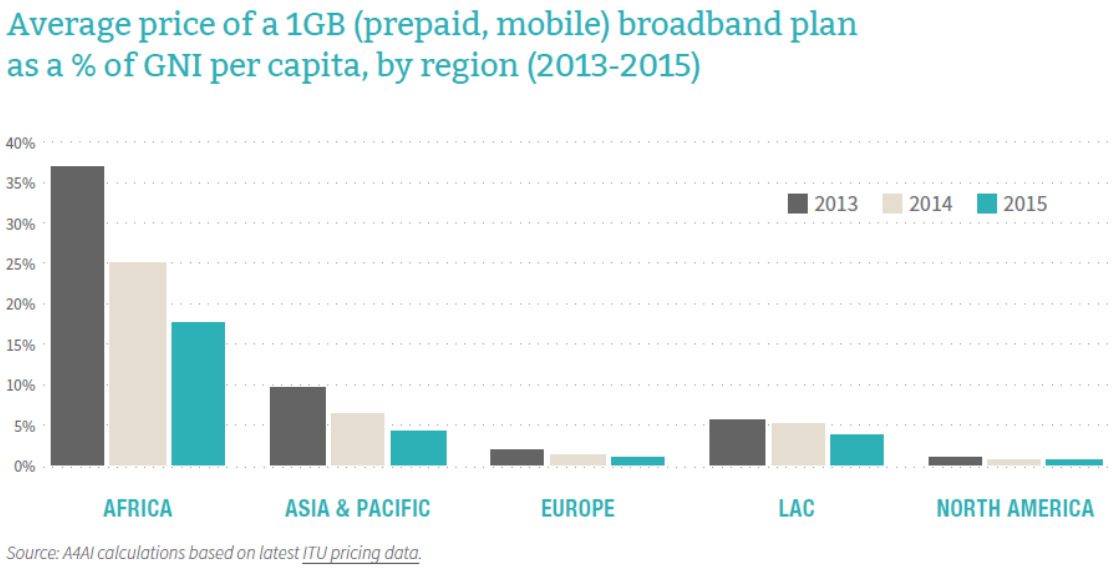


Figure 38: Comparison of BB Affordability by Region

In 2011, the Broadband Commission for Digital Development set a target that the entry-level BS should be made affordable in developing countries by 2015 through adequate regulation and market forces (amounting to less than 5% of average monthly income)[53].

Alliance for Affordable Internet (A4AI) has summarized six key actions to make Internet affordable[102]. They are:

1. *Employ Public Access Solutions to Close the Digital Divide*
2. *Foster Market Competition Through Smart Policy*
3. *Implement Innovative Uses of Spectrum through Transparent Policy*
4. *Take Urgent Action to Promote Infrastructure and Resource Sharing*
5. *Make Effective Use of Universal Service and Access Funds*
6. *Ensure Effective BB Planning Turns into Effective Implementation*

These actions could be effective for governments to provide universal service and to make BB available, accessible and affordable to the people. However, very little was mentioned on actions to develop the socioeconomic status in targeted areas as to sustain and grow the BB adoption once the government withdraws the financial support and allow the TELCO to take over business operations.

A typical household can afford to spend up to 5% of their income on ICT services. This perceived affordability has particularly spurred the growth of mobile communication[90]. The average cost of BB Internet is 1-2% of the monthly per capita income in developed countries. In some developing and emerging countries, the starting level of BB subscription costs over 27% of average earnings. The BB subscription costs per earnings may go up to 90% in certain under-developed countries like Zimbabwe. The goal of the A4AI is to achieve the UN Broadband Commission target of entry-level BS priced at less than 5% of the average monthly income[112].

In Malaysia, the average monthly household consumption expenditure in the year 2014 was RM3578, with 5.3% or RM189 spent on ICT per month. However, there is a gap in household expenditure in urban and rural areas. Taking RM189 as the average revenue

to service providers, that could mean the average household in urban areas was spending 4.8% monthly while the average household in rural areas was spending 7.8% of their household income on ICT per month. The statistics (DOSM 2014) shows that the bottom 40% of the household was spending only 4% of their expenditure on ICT, versus 5.8% for the top 20%. There are two methods to close the supply and demand gap – either to increase rural household income or to reduce the cost of service subscription.

BB adoption and sustainability are the primary concerns in BB investment, to both the public and private sectors.

Box 9: Relevance of the BB Development Stages in this Research

- This research deals with a machine learning technique to classify the socioeconomic potential of a geographic area for BB investment.
- If a geographic area is classified as having socioeconomic potential, BB investment is feasible to create BB availability that can evolve into BB accessibility and adoption.
- If a geographic area is classified as having low socioeconomic potential, BB investment is possibly limited to BB availability and accessibility stages only, whereas achieving both BB adoption and sustainability are unlikely.
- For these low socioeconomic potential areas, the empirical model applied in the machine learning technique can be used as a simulation tool to formulate a game theory for various stakeholders to improve the geographical features which have a high correlation coefficient to the socioeconomic response.

2.5.3 FEATURES THAT INFLUENCE BROADBAND DEVELOPMENT

The literature review on geographical features that influence BB development is based on four perspectives:

- Reference features from world organizations
- Reference features from academic research
- Reference features from Malaysian government agencies

- Reference features from the business sector

In the following sub-sections, words that appear in *Italic red* denote geographical features that are used in the research experiments or observations.

2.5.3.1 REFERENCE FEATURES FROM WORLD ORGANIZATIONS

The ITU Manual for Measuring ICT[83] is a useful guide to support countries in their efforts to measure and monitor their development towards becoming information societies. The manual has since become the statistical standard and measurement topics for ICT household statistics, not just for government agencies but also for private enterprises and researchers worldwide. The classificatory variables for ICT household statistics defined in the manual are one of the key references in selecting the right features for data collection to train the support vector machine in this research. The features extracted from the ITU Manual is shown in Table 10 below.

Table 10: Features that affect ICT development

Section 1: Household characteristics
Household size
Household composition (whether there are children under 15)
Household access to electricity
Household income
Household location (urban / rural)
Area populations
Section 2: Household access to ICT
Household with a radio
Household with a television
Household with a fixed telephone line
Household with a mobile telephone
Household with a computer
Household with Internet

BB availability and affordability
Cultural reasons
Section 3: Individual characteristics
Age
Sex
Ethnicity
Highest educational level attained
Labor force status (employee, self-employed, unemployed, etc.)
Occupation
Spoken language
Section 4: Individual use of ICT
Use of mobile phone
Use of computer
IT literacy
Experience with Internet
Others: Institutional or public policies
Availability of government agencies or local authority to develop and promote public policies

Source: Manual for Measuring ICT Access and Use by Households and Individuals, ITU 2014[83]

It is worth noting that these factors are not comprehensive. Individual countries may face other challenges that influence ICT development. For example, when studying the strategies and PPP options for supporting the ICT sector and BB connectivity in Somalia, the World Bank Group found that the development of a national fiber-optic backbone network must not only follow the population distribution in the country but also depend on the availability of *underlying infrastructure such as roads*[113].

World Bank (2013) reported that *roads* and the *provision of electricity* are needed to integrate rural areas into urban economy, and subsequently improve the markets and productivity of agricultural outputs[114].

2.5.3.2 REFERENCE FEATURES FROM ACADEMIC RESEARCH

Florence School of Regulation (2011) is another source of reference to select the right features for data collection to train the support vector machine in this research. The relevant research and the features that impact the BB deployment are shown in Table 11.

Table 11: Past Research on Determinants of BB Demand

Authors	**Research Findings**
Strover (2001)	Due to lack of *IT literacy* or understanding on the importance of digital information, rural users comparatively do not realize the value of BB.
Mills and Whitacre (2003)	The gap of Internet access in metropolitan and non-metropolitan areas are caused by differences in *education*, *income*, and other household attributes.
Franzen (2003)	Individuals with *high financial*, *human and social capital* tend to be earlier Internet adopters.
Hollinfield and Donnermeier (2003)	One way to maximize ICT demand in rural areas is to encourage locally owned rural businesses to adopt ICT and use online services.
Grubesic (2004)	The BB landscape is shaped by a mixture of various factors such as *geography*, *socioeconomic* status, *market forces*, and *policy*.
Chaudhuri, Flamm and Horrigan (2005)	*Income and education* are strong determinants of Internet adoption, while the influence of pricing or cost of access is moderate.
Goldfarb (2006)	*University students*, especially those who use email are more likely to be Internet users.
Preston, Cawley, Metykova (2007)	Initiatives to promote BB adoption in less developed areas need to be supported by *universal service policies*.
LaRose et al. (2008)	*Income* and *age* have a direct impact on BB diffusion.

Westlund and Bohlin (2008)	Uncertainties on the total cost of usage and low speed are barriers to *mobile BB* diffusion.
Orviska and Hudson (2009)	*GNI per capita* and Internet security are Internet barriers in certain countries in EU.
Drouard (2010)	Income, education, online experience, *socioeconomic factors,* and *computer skills* have an impact on BB adoption.
NTIA (2010)	In the US, the use of home BB is highest among Asians and Whites, married couples, younger people, urban residents, people with *higher incomes*, and people with *more education*.
LaRose (2011)	*Community education* has a positive correlation with BB adoption.
Lee Sa., Marcu, Lee Se. (2011)	*Income*, *population density*, *education*, and *price* have a positive correlation with *fixed BB adoption*.
Michaillidis (2011)	Value-add to people's live, awareness of benefits and local content are important to motivate BB adoption for users in rural Greece.

Source: Florence School of Regulation, 2011. Modified by the author.

Many past researches studied the socioeconomic impact of BB investment or the impact of BB on socioeconomic growth. Table 12 shows some examples of such research with its findings in correlation with ICT development and socioeconomic impact with the corresponding features that influence ICT development.

Table 12: Past Research on Correlation between ICT and Socioeconomic Impact

Authors (year)	Research areas/findings
Cronin et al. 1991, 1993a, b	*GDP*, *employment* and *economic activity* have a bidirectional relationship with telecommunication investment in the USA. For example, telecom investment enhances growth while economic activity and growth stimulate demands for telecom investment.
Dholakia and Harlam (1994)	*Telecom infrastructure investment* is one of the significant factors in the *economic growth* of a country.

Jill Windle and R.A. Cramb (1997)	*Road access* to large urban gives rural households wider market and opportunity for higher income.
Lee (2001)	*Secondary education* and computer access are important to enable human skills and capabilities to use new technology[115].
Daveri (2002)	The impact of ICT is different across countries due to different economic structure and *policy issues*.
Crandall, Lehr & Litan (2007)	BB penetration has a positive correlation with *employment* in both the manufacturing and service industries.
Hosman (2008)	In developing countries, *GDP per capita* is directly correlated with ICT investments.
Navarro & Sanchez (2011)	Females are 6% less likely to adopt BB.
Hilbert (2011)	When income and education are on par, gender is not a variable factor to BB adoption.
Girish J. Gulati and David J. Yates (2012)	*Income*, *BB penetration*, *education level*, *population density*, *market competition*, *national regulation*, *ICT investments*, *political institutions*, *socioeconomic control variables* are determinants of BB development.
James E. Prieger (2013)	*Population density*, *income*, *number of households*, race, and *education* affect BB adoption in rural areas.
Guldi and Herbst (2017)[116]	*Birth rate* decreases as BB access increases.

2.5.3.3 REFERENCE FEATURES FROM MALAYSIAN GOVERNMENT AGENCIES

The USP is a crucial vehicle in introducing BS to rural areas. However, there is a lack of evidence of success stories that the USP has improved the socioeconomic status of villages to a commercially viable stage that attracts private investment.

MCMC found that the top 10 challenges for USP projects are:

1. Rural areas are often remote and have lower income levels

2. Geographic dispersion of the population
3. Hilly terrain and dense trees areas
4. Availability of proven technology to effectively connect rural areas
5. Technology limitation - distance, coverage due to the LOS/NLOS system, CPE needs a power supply to operate, etc.
6. Unavailability/Unstable electricity supply
7. Difficulty in acquiring suitable sites for wireless solutions
8. High costs to rollout
9. Low take-up rates
10. Billing and payment issues

In this thesis, features mentioned in the literature review will guide the decision to choose real-life data which is available for use in the proposed machine learning process, including training and testing of the support vector machine.

2.5.3.4 REFERENCE FEATURES FROM THE BUSINESS SECTOR

A worldwide management consulting company, McKinsey (2014) reported that the four barriers that affect the offline population for not adopting Internet are: incentives, low incomes and affordability, user capability, and infrastructure. On the other hand, a multinational networking and telecommunications company, Ericsson (2013) reported that the factors that influence ICT development include: education, skills, and socioeconomic variables, e.g., age, gender, occupation, marital status, geographic area, and type of housing. The barriers and influencing factors reported by these multinational companies directly or indirectly will impact the ROI of BS deployed by TELCOs.

The ROI indicates how much economic benefit is derived from a project or program about its costs of investment. In the telecom industry, the economic benefit is usually seen as the number of years required to collect sufficient revenue to recover the cost of investment.

TELCO uses the techno-economic analysis for BB investment. The analysis involves processes of market analysis, services, and network modeling as well as technical and economic evaluation. Payback period, net present value (NPV) and internal rate of return (IRR) derived from the analysis are commonly used to determine the profitability of the business scenarios[117].

Elvidge and Martucci (2003) mentioned that a commercially viable project should consider net present value (NPV), payback, internal rate of return (IRR), cash flow and breakeven point. Table 13 shows the financial indicators commonly used in project management[118].

Table 13: Financial Performance Indicators for Project Management.

Indicator	Description
Cash flow	An indicator that shows the amount of money in hand to spend or the difference between cash collected and expenses incurred.
Breakeven	An indicator that shows a business is neither making money nor losing money.
Payback period	An indicator that shows the number of years required to recover the money spent or invested.
Net present value (NPV)	NPV is the present value of future revenues minus the present value of future costs. Negative NPV means the present value of the costs exceeds the present value of the revenues.
Internal rate of return (IRR)	IRR is a discount rate that makes the NPV of a project equal to zero.

Source: Elvidge and Martucci[118], modified by the author.

2.5.4 PUBLIC POLICIES THAT INFLUENCE BROADBAND DEVELOPMENT

Numerous researches, i.e., FSR (2011), Analysis Mason[119], Falch, M. & Henten A.[120], Blackman, C. & Srivastava, L.[121], Kelly, T. and Rossotto, Carlo Maria[122],

World Bank Group[113] and so forth have recommended various models for effective public interventions or PPP to promote BB deployment.

As mentioned in previous chapters, there is a discrepancy in the impact of BB penetration in technologically developed countries compared to less developed countries. The journey to worldwide accessibility and use of BB requires different strategies and policies depending on a nation's level of technological development[123].

F. Belloc et al. (2012) found that both the supply and demand policies (Figure 39) have a positive impact on BB adoption. However, the impact level varies at a different stage of BB development[124].

SUPPLY-SIDE POLICIES	DEMAND-SIDE POLICIES
• adoption of fiscal incentive programs and subsidies	• public demand of specific services
• implementation of long-term loans programs for broadband suppliers and national financing programs	• provision of incentives to business demand
• creation of Public-Private Partnerships with public ownership of the infrastructure network	• provision of incentives to private demand
• creation of Public-Private Partnerships with private ownership of the infrastructure network	• provision of demand subsidies in favour of individual consumers or particular categories of consumers
• implementation of territorial mapping programs	• adoption of demand aggregation policies
• administrative simplification initiatives	

Figure 39: Direct Public Policies to Stimulate BB Penetration. Source:[124]

The abovementioned discussion about the challenges faced by BB investors, particularly rural areas in developing countries, has led to the proposal of a machine learning technique. This thesis proposes an SVM that can classify geographic areas according to its socioeconomic potential. The output of the research will help policymakers in both private and public sectors to formulate corresponding BB funding or investment policies effectively.

2.5.5 BROADBAND INVESTMENT

Generally, BB investment comes from the TELCO in the private sector or government agency in the public sector. The private sector in Malaysia continues to invest in geographic areas that are commercially feasible whereas the government agencies continue shouldering investment in underserved or unserved areas. As mentioned in section 2.5.2, BB development evolves in four different stages. These stages are BB being available, accessible, adopted and sustainable. Every stage of the development requires financial investment to upkeep the service and to reach a sustainable stage.

In those geographic areas that are commercially feasible, the TELCO will invest in every stage with the expectation to achieve a sizeable adoption rate with a predictable return on investment. TELCOs also continue with their investment to upgrade the BS to sustain or grow the adoption rate.

However, in those geographic areas that are not commercially feasible, government agencies will make BB available to the people through PPP projects. Further financial commitment and a special program will be introduced by the government to help users access BS. Unless the adoption rate can be achieved within a predictable timeline and return on investment, government agencies will need to continue its funding to continue BB availability and accessibility to the people. If the adoption rate is growing and expected to reach a certain level within a specific timeline, the TELCO will be interested in turning the PPP into private investment.

Both private and public sectors continue investing in BB development as the BB is boosting ICT applications development (Figure 40) across urban, suburb and rural areas.

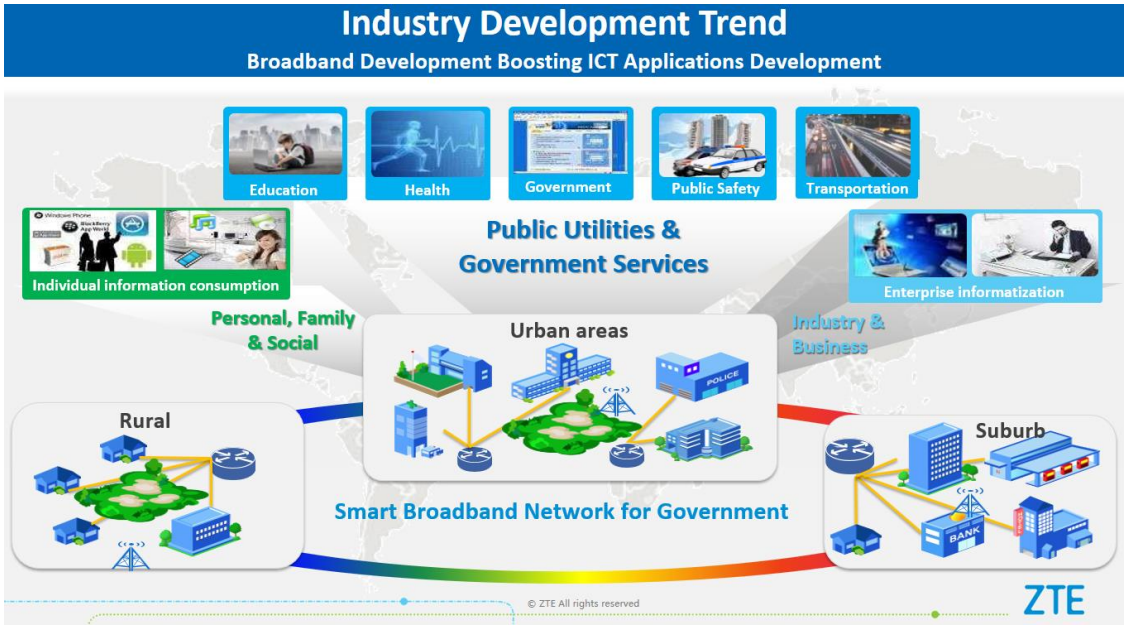


Figure 40: Broadband Boosting ICT Applications Development. Source: ZTE

To materialize the national BB strategies, it is common for public sector and private sector to partner in BB development. Figure 41 illustrates a BB construction model that is commonly adopted in a PPP in the telecom industry.

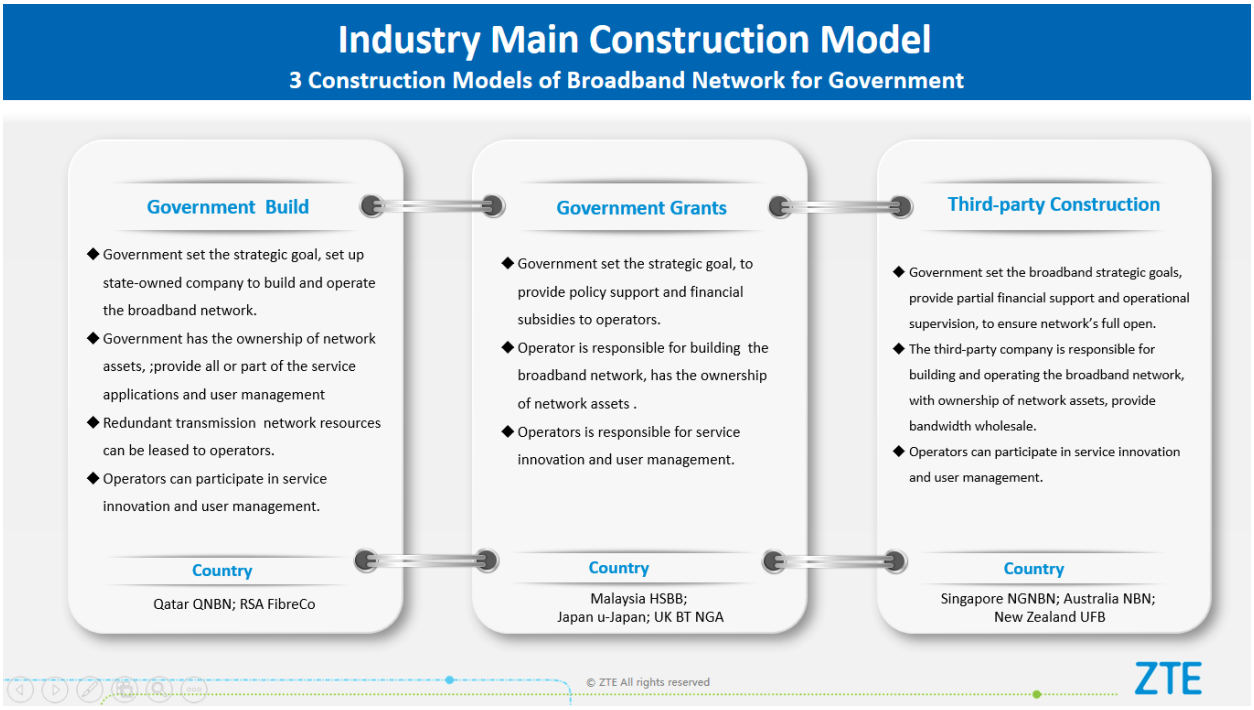


Figure 41: Construction Models of BB Network for Government. Source: ZTE

J. Navas-Sabater et al. (2002)[125] divided the total geographical reach into 4 zones (Figure 42). The first zone is the zone with the current network reach and access. This zone typically covers the geographic areas which have good commercial value for private investment. The second zone is the zone with a market efficiency gap which is a commercially feasible reach for private investment. The third zone is the smart subsidy zone which can become commercially feasible after a one-time subsidy from public funding. The fourth zone is geographic areas with a true access gap which will not be commercially feasible even with government subsidies.

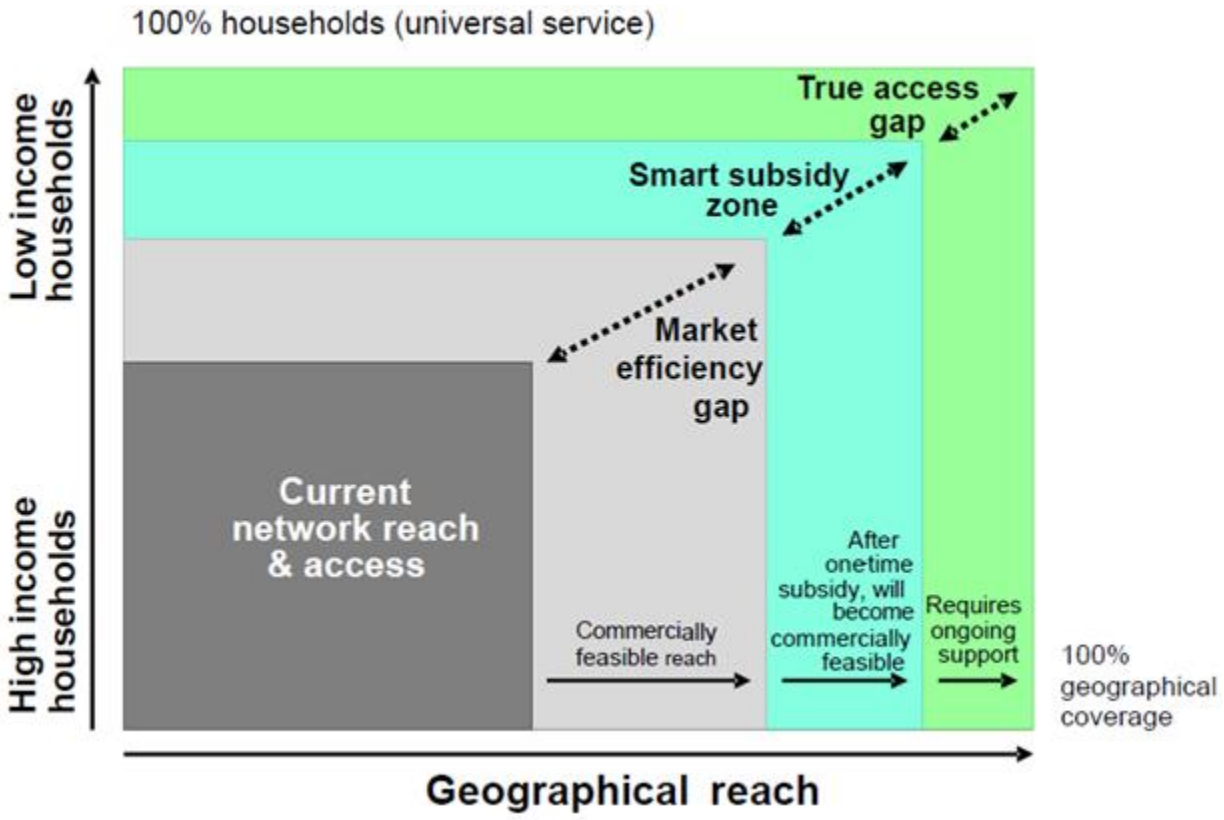


Figure 42: Four Geographic Zones for BB Investment. Source:[125]

This thesis intends to propose a support vector machine that can classify the geographic areas with socioeconomic potential for BB investment. Once, the SVM classifies a geographic area for BB investment, the classified area can be targeted as a smart subsidy zone for PPP. However, if a geographic area does not have socioeconomic potential,

then the empirical model can be used to correlate the geographical features to the socioeconomic response to simulate a game theory to develop the area further.

2.5.5.1 TELCO AS A BROADBAND INVESTMENT VEHICLE IN PRIVATE SECTOR

TELCOs in Malaysia continue to invest heavily in either expanding or upgrading their networks. For example:

- Maxis invested RM815 million in 2013 primarily to expand its coverage of 4G networks nationwide[126]. In June 2015, Maxis announced to invest another RM1.1 billion in improving its 4G networks and service quality[127].
- Celcom invested RM923 million in 2013 to expand its coverage of 4G networks nationwide[128]. Celcom has been consistently investing more than RM1 billion a year in CapEx and OpEx since 2006[129].
- Digi invested RM2.3 billion from 2012-2014 in CapEx. The TELCO committed to invest another RM900 million in 2015 to strengthen its network coverage[130].
- U Mobile spent RM1.5 billion between 2014 and 2015 to improve its network[131].
- Jointly TM and Green Packet to invest up to RM1.65 billion (over 1-3 years) to fund P1's operation[132].
- Altel will be investing RM1 billion over the next five years to deploy 4G mobile BS[133].
- YES invested more than RM2 billion between 2011 and 2014 to expand its network[134].

The abovementioned investments though, only reach the urban or sub-urban areas where ROI is secure. However, Malaysia still has rural areas where BS is either unserved or underserved.

Rural population (% of total population) in Malaysia peaked at 73.40% in 1960 and dropped to 24.63% as of 2016 (Figure 43). In spite of the decreasing % of rural areas, it is still encouraged to invest in those areas. On the other hand, BB penetration has crossed 70% nationwide but with a slower growth rate in the last five years (Figure 44). Therefore,

it is essential for the private TELCO to expand its BB investment in rural areas with socioeconomic potential. An investment decision based on ROI alone might not be sufficient because the ROI model cannot identify areas which have the socioeconomic potential to sustain BS in the long run. Hence, this thesis suggests the assistance of new intelligence or method to pin-point viable areas for BB investment.

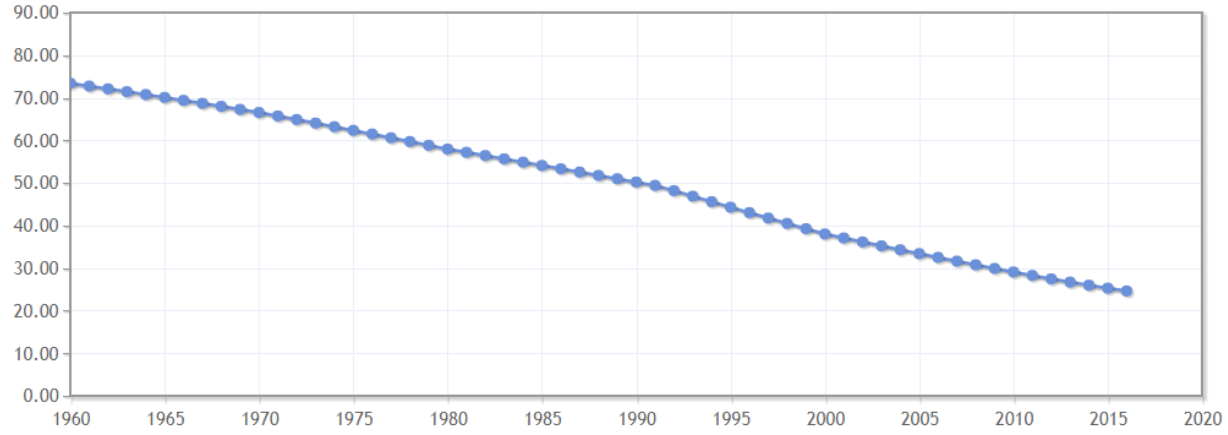


Figure 43: % Rural Populations in Malaysia. Source:[135]

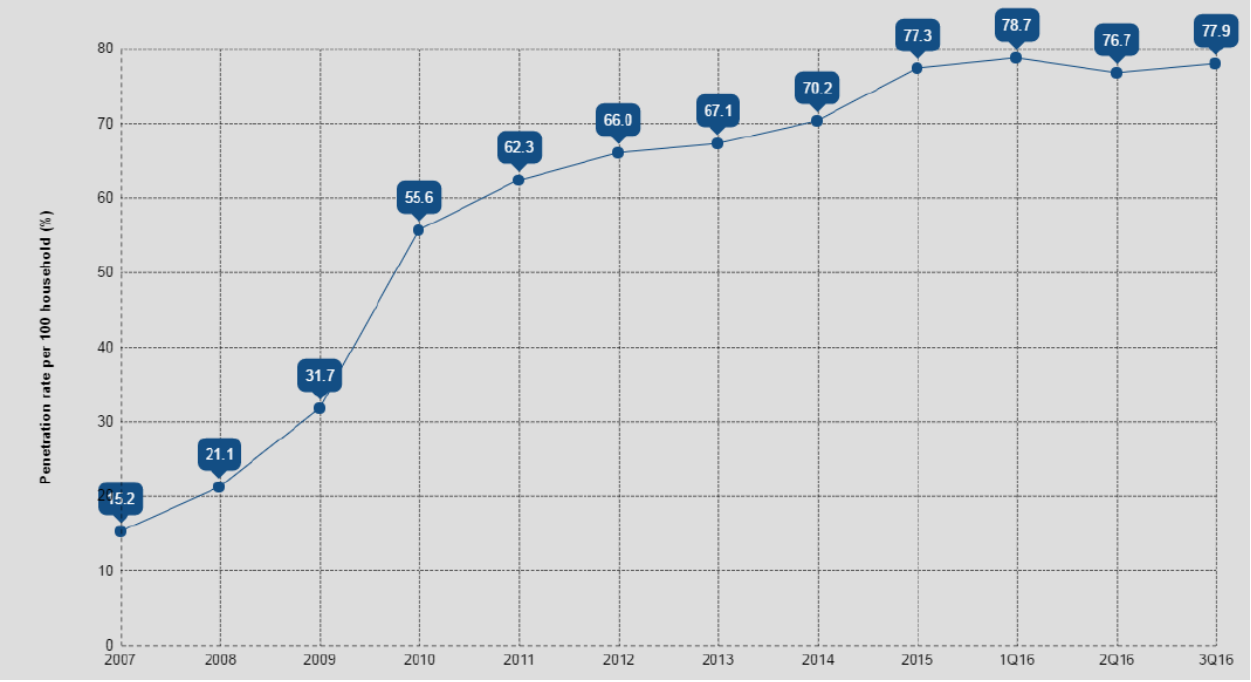


Figure 44: Household BB Penetration. Source: MCMC

Even though the Malaysian TELCO industry continues expanding its network coverage, statistics show that the TELCO's coverage with newer BB technology is in the range of 50% to 80% of the entire Malaysian population. The competition among the TELCOs for the urban and suburban areas will only become increasingly intensified especially among those LTE licensees. Furthermore, the United Nations (Department of Economic and Social Affairs, Population Division, 2014) forecasted that the urban population in Malaysia would only increase from 74.7% (2015) to 80.1% by 2025. BB subscription will remain extremely competitive among TELCOs. TELCOs that can find new geographic areas with socioeconomic potential will also find new opportunities for BB investments.

Figure 45 illustrates an overview of the architecture for BB infrastructure. TELCOs usually refrain from investing in rural areas due to high deployment cost and limited commercial feasibility. The cost of deploying a high-capacity BB access network outside in a rural area is on average 80% higher than the cost of deploying the network in the town area[136]. Upgrading the last mile of networks in rural areas is much more expensive than urban areas in all countries. The network investments in rural areas are delayed due to high deployment costs at low population density areas. Furthermore, the ROI in sparsely populated areas is too low to sustain commercial operations[121].

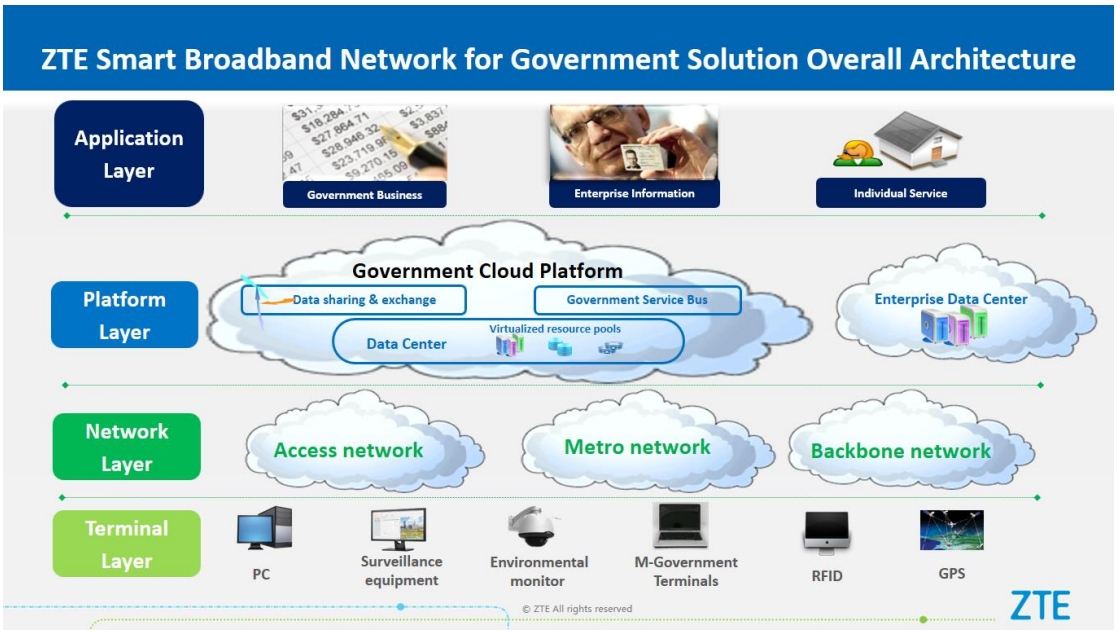


Figure 45: Overview of Architecture for BB Infrastructure. Source: ZTE

The Global System for Mobile Communications Association - GSMA (2016) reported that 3G network coverage has increased from 63% in 2014 to 80% in 2016 worldwide, driven by investment and network sharing. However, the uncovered areas tend to be rural, often remote localities where the economics of expansion need different models of investment decisions[137].

2.5.5.2 USP AS A BROADBAND INVESTMENT VEHICLE IN PUBLIC SECTOR

In Malaysia, Universal Service Provision (USP) is a common vehicle that the telecommunications regulators use for (a) initially delivering basic connectivity and communication services for underserved areas and (b) subsequently bridging the digital divide. The vehicle serves to enable service availability, affordability and accessibility of telecommunication services to all, particularly the underserved areas.

Usually, the USP funded projects exist to fill the “access gap” and to foster the closure of the adoption gap and to reach a point that is commercially attractive for private investments. The PPP can then fill up the value gap which is crucial for service sustainability. Figure 46 illustrates the path of gap fulfillment for ICT development.

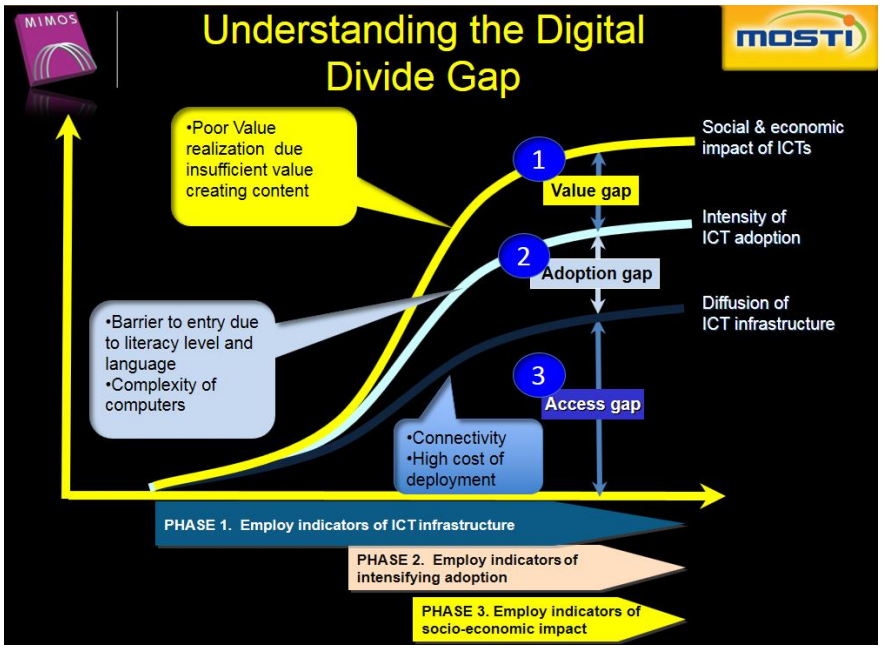


Figure 46: Understanding the Digital Divide Gap. Source:[138]

In the 11th. Malaysia Plan, the government of Malaysia targets to increase the populated areas covered by BB infrastructure to 95%. Nevertheless, the government recognizes the high cost of network deployment and low ROI in rural areas. Furthermore, it is uncertain on the level of logistics and utility support required to ensure economic growth in rural areas. Hence, there should be a strategic PPP to deploy telecom networks in rural areas at a reduced cost or predictable socioeconomic growth rate.

Muente-Kunigami and Navas-Sabater (2009) found that telecommunication investment in rural areas is either challenged by high investment cost or low demand rate [139]. Serving the rural areas with high CapEx and OpEx would be unprofitable to TELCOs unless provided with alternatives that could either reduce deployment cost or increase the economic growth at a predictable rate.

2.5.5.3 SUPPLY AND DEMAND FOR BROADBAND INVESTMENT

Katz & Berry (2014) defined the supply gap as the number of households whereby either fixed or mobile BB is not available. In those geographic areas where TELCOs cannot make profits and thus do not provide their services, government agencies are required to implement policies addressing the supply gap. The demand gap focuses on the potential users that can buy BS but do not[12].

<table>
<tr><td rowspan="3"># of households</td><td># of households where BB is not available</td><td colspan="2">Supply Gap</td></tr>
<tr><td rowspan="2"># of households where BB is available</td><td># of households that do not subscribe BB</td><td>Demand Gap</td></tr>
<tr><td colspan="2"># of households that subscribe BB</td></tr>
</table>

Figure 47: BB Supply Gap and Demand Gap Outlined by Katz and Berry. Source:[12]

The sum of the supply gap and demand gap makes up the digital gap or the digital divide gap.

2.6 TECHNO-ECONOMIC ANALYSIS FOR BROADBAND INVESTMENT

A techno-economic analysis is a process to evaluate how the economy has impacted the introduction of certain technologies. In the context of BB investment, the techno-economic analysis is a process to evaluate the economic impact of the introduction of BS to a geographic area.

There are several types of techno-economic modeling for BB investment. A different model could be addressing the techno-economic viewpoint according to technical constraints, regulatory constraints, physical constraints, network design and deployment process, service model, service quality and capacity, customer demand and desired business results such as total investment cost, revenue, rate of return, return on investment and so forth. This section highlights the various models available for techno-economic analysis that is found in the literature review and also based on the author's own experience as part of the management team of a TELCO.

Although there is a strong economic rationale for investment in telecom infrastructure, the Asian Development Bank (1997) emphasized that policy analysis is also necessary to:

1. *encourage migration from public sector domination to a commercial mode of operation;*
2. *encourage the development of a pricing and regulatory regime that facilitates efficiency and liberalization, where appropriate;*
3. *identify areas, such as in rural regions or in establishing new international links for transitional economies, where lower financial returns may mask potentially high economic benefits and where some form of public sector leadership, or special development funding or subsidy, may be justified until the network builds up to a commercially attractive level of operation*[140].

In other words, the techno-economic analysis is important to maximize revenue and economic benefits as a return of technology investments. However, policy analysis is also necessary to encourage migration from public sector domination to a commercial mode of operation in rural telecommunication projects. While the ROI model is commonly used by TELCOs in Malaysia, this thesis proposes a machine learning technique as an alternative solution.

Both ROI and machine learning technique can be used in the techno-economic analysis for BB investments. But the machine learning framework also encompasses the development of an empirical model which establishes the statistical correlations between the geographical features and the socioeconomic response.

ROI is a natural consideration in the business decision-making process. In the context of BB development, the telecom equipment cost and its operating cost are the primary investment whereas the revenue and profits will predominantly determine the rate of return on investment. The investment and its return form the basis of the techno-economy analysis. The financial investment is subject to the type of BB technology, network design, services to offer and the operations and maintenance of the technical network. The financial return is subject to the economic value of the services offered and willingness of people to subscribe and pay.

TELCOs in the private sector commonly apply techno-economic analysis before making an investment decision to deploy BB network and services in any targeted geographic area. Similarly, government agencies in the public sector apply techno-economic analysis before formulating a national BB initiative or creating a project with USP funding.

Typically, a full cycle of techno-economic analysis involves three main steps, namely:

1. Market analysis
2. Techno-economic calculation
3. Techno-economic evaluation

Figure 48 is an example of a 3-step techno-economic analysis (TEA) illustrated by M. Kantor et al. (2010) for telecommunication services[141].

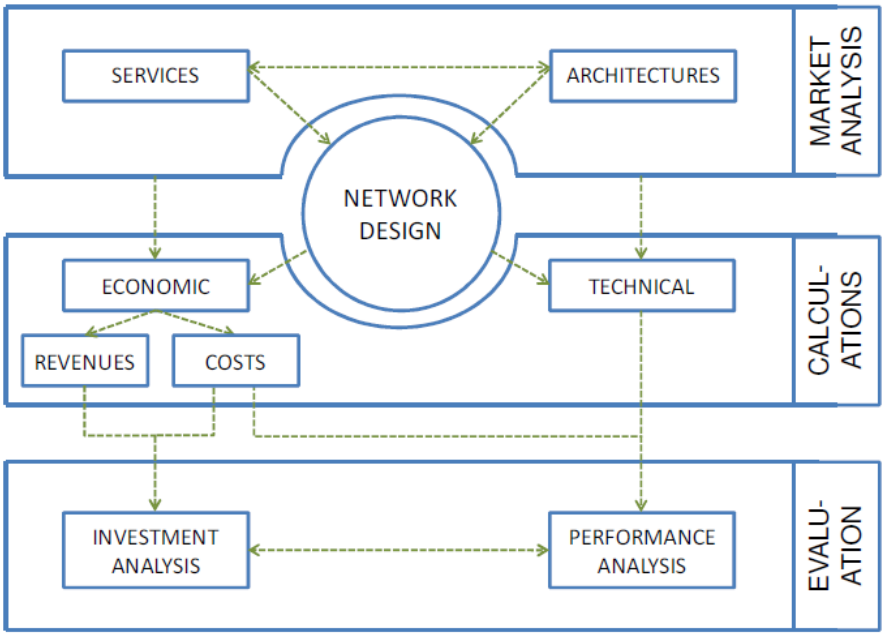


Figure 48: 3-Step TEA for Telecommunication Services. Source:[141]

Besides the 3-step TEA lifecycle illustrated by M. Kantor *et al*. (2010), there is another TEA model that is based on the PDCA lifecycle. PDCA stands for Plan-Do-Check-Act.

Verbrugge et al. (2009) summarized the techno-economic analysis for telecom network planning into four stages of scoping, modeling, evaluating and defining. The 4-stage techno-economic analysis (Table 14) is loosely based on the PDCA cycle.

Table 14: 4-Stage Techno-Economic Analysis for Telecommunication Service

Process	**Activities**
Scoping	• Breakdown the scopes to be studied, i.e., areas, users, type of services, technologies, costs, revenues and so forth • Collect required data such as technology, market, and target area

	• Create business scenarios such as user adoption, business models, and technical design
Modeling	• Establish cost and revenue models • Perform network dimensioning and establish parameters for use in cost/revenue models
Evaluating	• Perform techno-economic evaluation by importing the data from “scoping” into “modeling” • Analyze financial performance on payback time, ROI, NPV, and IRR
Defining	• Refine techno-economic evaluation with what-if analysis • Provide options for decision making.

Source: Verbrugge et al.[142], modified by the author.

The scoping and modeling steps in the PDCA cycle are similar to market analysis in the 3-steps TEA lifecycle. The evaluating step in the PDCA cycle contains the element of techno-economic calculation and part of the techno-economic evaluation as proposed by M. Kantor et al. The final step of defining in the PDCA cycle is equivalent to the techno-economic evaluation step that wraps up the whole techno-economic analysis cycle.

Deploying BB networks is capital intensive with a high operating cost. A detailed study on techno-economic analysis is crucial for (a) when TELCO decides to invest in deploying the BB network in any targeted geographic area, and (b) for government agencies to fund the provision of BS in underserved or unserved areas. The guidelines for the economic analysis of telecommunication projects provided by Asian Development Bank provide further details on the relationship between technology (i.e., network design) and economy (i.e., financial sustainability, poverty impact, affordability) in telecommunication investment projects[140].

Various techno-economic analysis (TEA) models have been developed out of the methodology mentioned above. Models that are commonly used in telecom industry include:

- TITAN – Tool for Introduction Scenario and Techno-Economic Evaluation of Access Network
- OPTIMUM – Optimized Architecture for Multimedia Network and Services
- TERA – Techno-Economic Results from ACTS (Advanced Communications Technologies and Services).
- TONIC – Techno-Economics of IP Optimized Networks and Services
- ECOSYS – Techno-Economics of Integrated Communication Systems Services

Typically, a TEA model includes market demand, cost modeling, revenue modeling and financial analysis for business decision making. Figure 49 illustrates the typical structure of a core model for TEA in telecom industry[143].

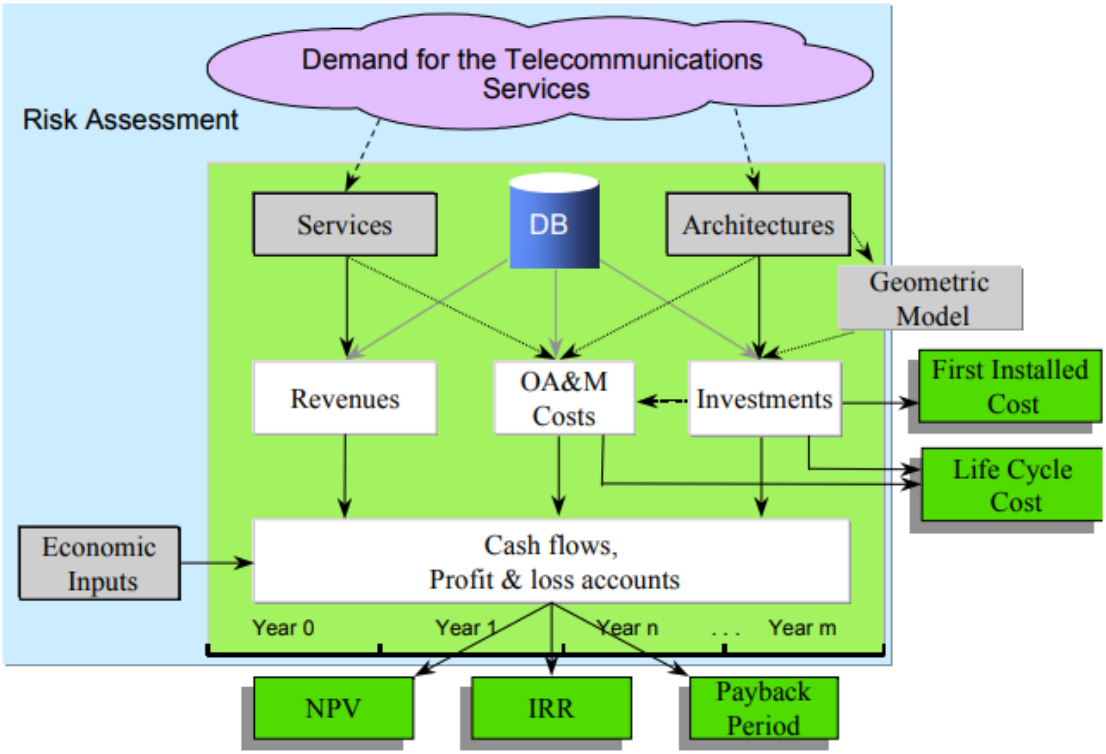


Figure 49: Techno-Economic Analysis Model for Telecom Industry. Source:[143]

Regardless of the TEA methods, all models will deal with the technological and economic aspects of the techno-economic analysis.

Total Cost of Ownership (TCO) is a cost model that is commonly used in the telecom industry. TCO includes CapEx, OpEx, regulatory requirement and so forth. TCO and ROI calculations are used by many organizations to justify the cost of investment in business systems (network and otherwise)[144]. TCO can be evaluated with a top-down approach or a bottom-up approach (Figure 50). The top-down approach starts with the existing business condition and available network. The bottom-up approach looks into the forecasted demand to map the network required and cost involved. In general, the bottom-up approach is preferred in the techno-economic analysis for decision making on investments.

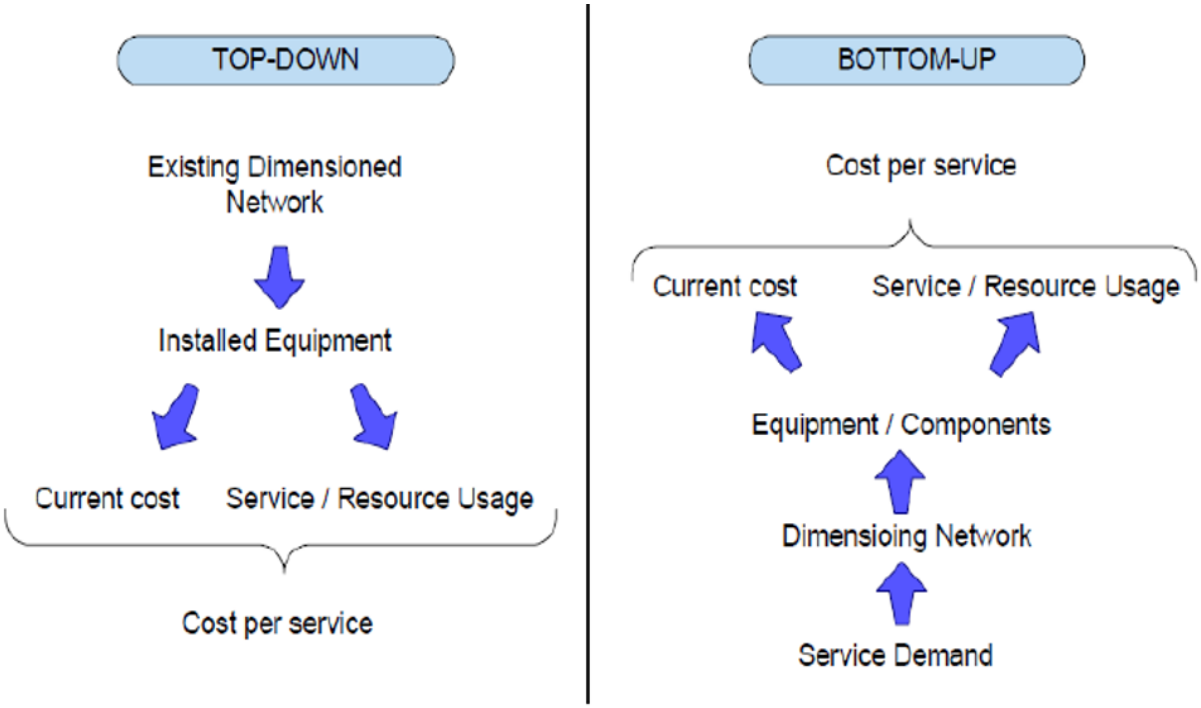


Figure 50: Top-Down vs. Bottom-Up Cost Modelling. Source:[145]

To cover all costs of ownership, TELCOs usually look through the whole lifecycle of the network operating model (refer to Appendix I) and perform a gradual cost-breakdown for the costs of the different phases.

2.6.1 ROI AS TECHNO-ECONOMIC ANALYSIS FOR BROADBAND INVESTMENT

ROI (Return on Investment) is commonly used in the telecommunication industry, including Malaysian TELCO, to plan for the long-term viability of their business and make decisions regarding how to allocate resources. Sometimes, return on investment is referred to as return on capital.

A 3-step process of techno-economic analysis is used to illustrate the return on investment as a techno-economic model in this section.

I. Market Analysis

The market analysis begins with a target market in mind by having a targeted geographic area to be covered with BS. The geographic area is usually measured by territorial, districts or more scientifically in square kilometers. The accurate network planning is possible when the terrain type (e.g., urban, suburban, rural) of the geographic area is known[146].

Whether a geographic area has the potential for economic growth, the primary factor for investment decisions is based on the potential revenue can be earned.

$$Revenue = ARPU \times number\ of\ subscribers$$

ARPU denotes average revenue on product pricing per user, whereas the number of subscribers representing the users who are willing to pay for the service.

Therefore, a commercial viability study is a primary concern to be analyzed. Examples of decision factors for commercial viability are:

- Type of products and services
- Selling price for products and services
- Population number/household number

- Cluster planning
- Adjacent area planning
- Area income level (at least state level per national statistics)
- BB penetration or market share

Network architecture and network design need to be developed to support the result of the commercial viability study. Figure 51 and Figure 52 show the example of a telecommunication network architecture based on Long Term Evolution (LTE) technology. The RAN, transmission, core and IP transit are key elements of the network that TELCO needs to rollout or deploy, which is the primary investment on CapEx and also OpEx once service is on air.

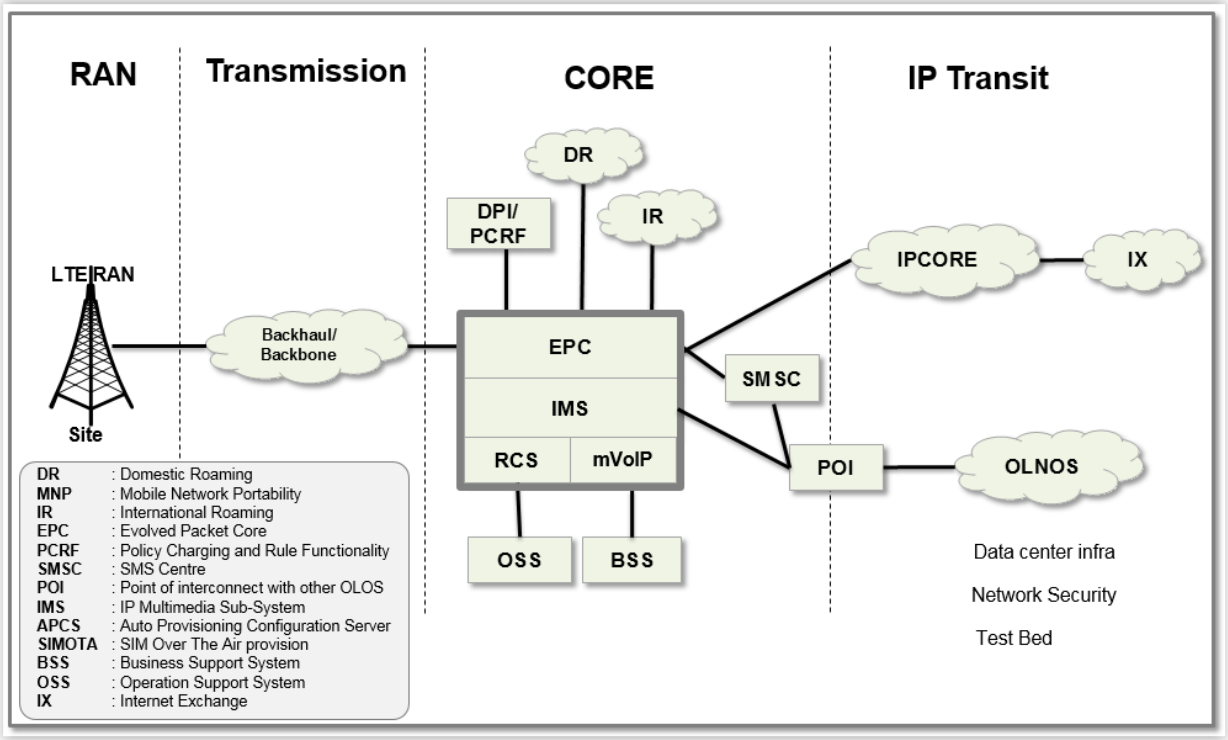


Figure 51: Example of LTE Network Architecture. Source: Packet One Networks

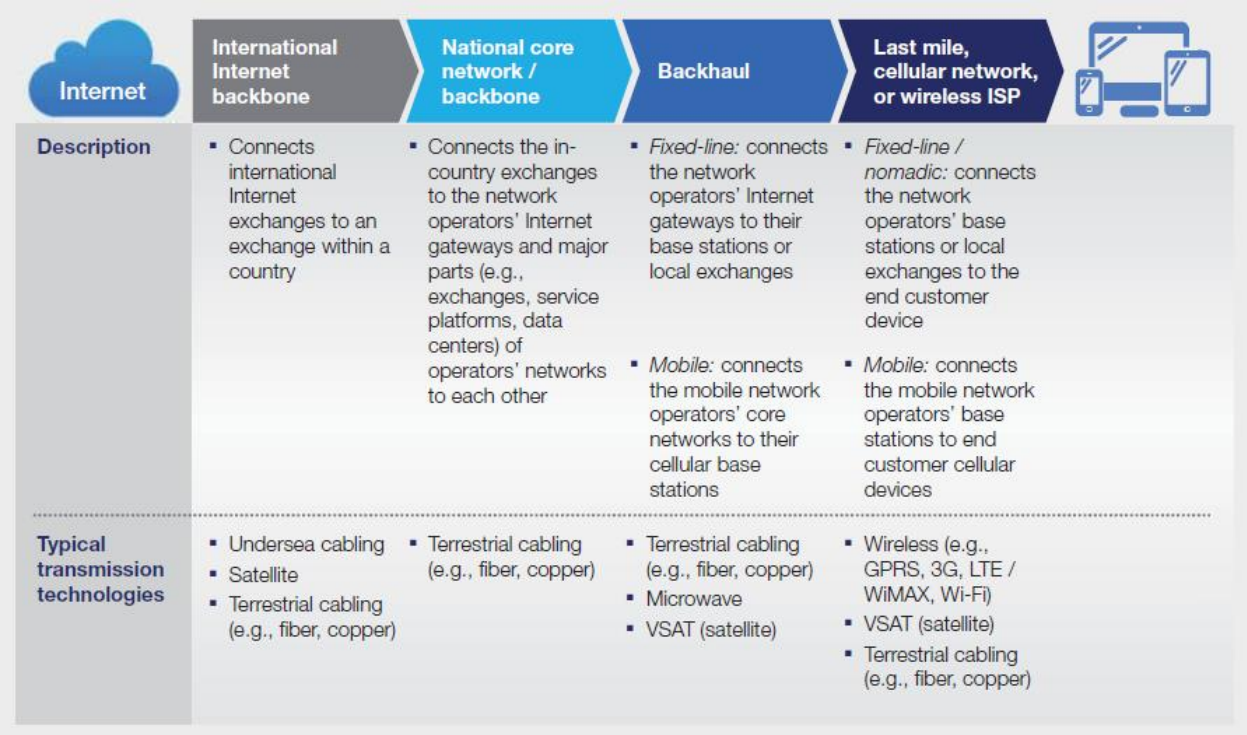


Figure 52: Data must travel through several networks to reach an end customer[86]

Another key component for cost is the operating expense (OpEx) for the TELCO to operate and maintain its BS.

The network rollout cost and operating cost constitute the capital investment and incremental operating cost to upkeep the telecommunication services for end users.
Many technical factors need to be considered when designing a BB network. Examples of decision factors for technical viability are:

- Spectrum
- Type of technology
- RF planning (radio characteristics, link budget, spectral efficiency, and antenna configuration)
- Backhaul planning
- Site acquisition
- Permitting
- Technical integration

- Network acceptance and optimization

II. Techno-Economic Calculation

For a targeted geographic area, the mathematical equation for ROI is written as:

$$\text{ROI} = \frac{payback - investment}{investment}$$

$$= \frac{revenue - expenses}{investment}$$

$$= \frac{(ARPU \times subscribers) - (operating\ expenses)}{capital\ expenses}$$

$$= \frac{(ARPU \times \%\ of\ pops) - (operating\ expenses)}{capital\ expenses}$$

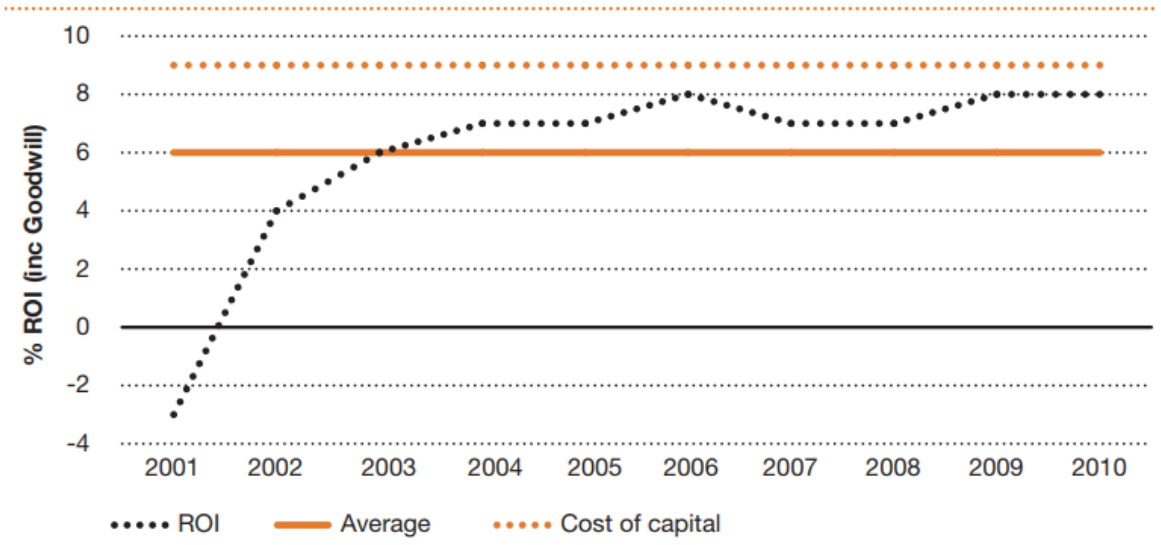
Where,

- Payback is the total amount of money earned from the investment.
- Investment relates to the number of resources put into generating the given payback.
- Revenue is the total amount of money collected from paying users (subscribers)
- ARPU is the average revenue per user according to the product mix and price matrix.
- % of pops is the percentage of the population in the geographic area who subscribe to the BS and pay for the service.
- Operating expenses (OpEx) is the running cost to maintain and operate the network. It includes the direct cost (e.g., backhaul cost and Internet transit cost) which will increase according to the increase (or decrease) of subscriber count.
- Capital expenses (CapEx) is the initial investment to build the network, and also the new investment to upgrade or expand the network.

The two most important cost factors for BB technology are the CapEx and OpEx.

- CapEx components include spectrum licenses, customer premises equipment, radio access network equipment, backhaul equipment, core network equipment, site acquisition and construction, network operating center setup and so forth.
- OpEx components include site rental, site utilities, backhaul cost, IP transit cost, operations maintenance and support, customer service, customer acquisition cost, general and administrative expenses and so forth.

In the past decade, the average long-term return on investment has been around 6% in the telecom industry (refer to Figure 53). High CapEx and drops in ARPU are causing the TELCOs to decrease the ROI in the telecommunication industry[147]. Nearly two-thirds of the executives working in the TELCOs argue that CapEx planning is driven by technology, not business objectives (PWC Analysis, 2013).



Source: Capital IQ, PwC analysis

Figure 53: Average returns on investment for 78 network operators

III. Techno-Economic Evaluation

The measurement unit for ROI can be either (a) the number of years to recover an investment or (b) the percentage increase or decrease of an investment over a period.

The thresholds of the definite output are subject to the appetite of the shareholders and investors over market shares and financial gain.

Another challenge in techno-economic evaluation is to address the BB affordability issues which are a high-impact factor to BB adoption leading to the revenue calculation.

Affordability = % of income spent on BB per month

Figure 54 shows a comparison of fixed BB prices between Malaysia and other ASEAN countries. In the 11th Malaysian Plan, the Malaysian government targets to reduce the fixed BB cost from 2.42% of GNI per capita in 2013 to 1% in 2020.

Country	Fixed-broadband prices as % of GNI p.c.
Brunei	1.9
Cambodia	34
Indonesia	9.1
Laos	N/A
Malaysia	3.1
Myanmar	N/A
Philippines	12.4
Singapore	0.8
Thailand	5.6
Vietnam	11.3

Source: ITU Measuring the Information Society 2013

Figure 54: Comparing BB prices between Malaysia and other ASEAN countries

The ROI model is commonly used by Malaysian TELCOs. Generally, the TELCOs provide BS in urban and suburban areas but not necessarily in rural and remote areas. The ROI in rural or remote areas can drop drastically due to unpredictable overall market conditions and lack of operating experience[148].

2.6.2 COMPARING ROI TO SVM

ROI is an econometric model which is based on a statistical model whereas SVM is a machine learning model which is based on statistical learning theory and computational learning theory.

Statistical models are mathematical equations that formulate the relationships between variables in the data. Machine learning models are algorithms that can learn from data without relying on rules-based programming. Statistics is about sample, population, and hypothesis whereas machine learning is about predictions and classifications.

The input and output of the SVM model are more scientific while the input and output of the ROI model are comparatively more subjective as it requires human dictation.

In machine learning, the input data is taken as it is for prediction. For example, the data of the geographical features for a geographic area will be taken as it is. The machine learning model will predict the outcome if that geographic area has the potential for BB investment according to the machine's learning experience.

In the ROI model, the user will need to decide the market size desired in a geographic area, and the level of investment with a targeted appetite on the return on investment such as rate of the payback period. In a geographic area where the market size does not meet the business appetite, capital investment is unlikely to take place. In the framework of machine learning technique proposed in this thesis, there is still a chance to develop a viable area for BB investment through improving certain geographical conditions to make the area become socioeconomic potential.

Table 15 summarizes the differences when comparing SVM and ROI models.

Table 15: Comparison of SVM vs. ROI

Model	SVM	ROI
Technique	Machine learning	Econometric
Theory	Statistical learning and pattern learning theory	Statistical calculation
Output Type	Classification	Regression
Output Results	Predictions and classifications (e.g., learn the pattern of the rural areas according to geographical features and predict the socioeconomic potential of rural areas)	Calculated results (e.g., calculate the rate of investment return based on a mathematical equation)
Execution	The manual calculation is unlikely	Can be manually calculated
Input Dictation	Objective: setting thresholds and boundaries as a learning environment based on real-life data and let machine predict the outcome	Subjective: dictating the magnitude of the thresholds according to business appetites such as market share to be captured and selling price
Market Analysis	PESTEL data	Market size, market shares, product mix and selling price, etc.
Technical Analysis	Technological and environmental features such as length of roads, distance from nearest town, existing BB penetration, etc.	BB technologies, network infrastructure, service quality and so on

Technical Calculation	Machine learning with training data	Cost Modeling
Economic Calculation	Machine learning with training data	Revenue modeling.
Investment Analysis	Classification of areas with socioeconomic potential	IRR, payback period, cash flow, etc.
Performance Analysis	What-if analysis on the correlation coefficient of the curve-fitting model and formulation of game theory among stakeholders involved	What-if analysis of investment analysis parameters and adjust according to business appetite

3 METHODOLOGY

The literature review has shown that BB investment is important to socioeconomic growth in a geographic area. However, the literature review has also shown that TELCO's practice of using ROI model only works well in currently commercially viable areas. Rural areas, on the other hand, may pose challenges to BB investment because of several factors such as population density, education level, economic activities, road system and so forth. Consequently, this thesis proposes another superior method of identification of geographic areas that are viable for BB investment, including both urban and rural areas. This method is based on a machine learning technique which will be implemented in 4 stages: data collection, machine learning, machine testing, and machine application. Figure 55 illustrates the complete flow of the 4-stage implementation. In this chapter, the research methodology will be discussed.

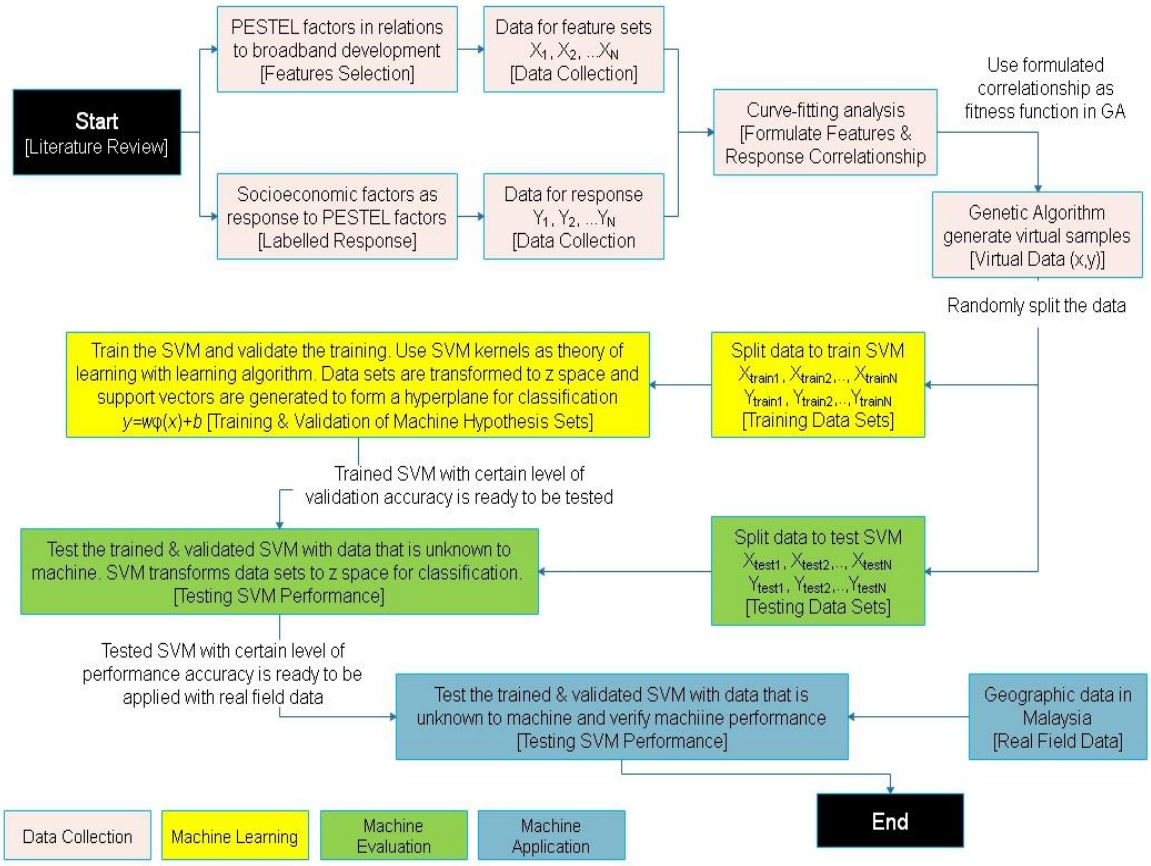


Figure 55: Research Methodology with 4-Stage Implementation. Source: the author.

3.1 DATA COLLECTION

There are four steps to collect the required data:

a. Defining data
b. Collecting data
c. Generating data co-relationship
d. Simulating large virtual data for machine training

Defining Data: At this stage, features of a certain geographic area are identified and labeled. For this thesis, geographical features are obtained from the World Bank's database and Department of Statistics Malaysia (DOSM). A total of 71 features is

selected in this research for data collection. Table 16 shows the selected geographical features. Once the features are decided, data of the features are collected.

Table 16: Geographical Feature Sets Defined for Data Collection

No.	Variable Factors (Characteristics)
1	Location (East Coast or West Coast)
2	Classification of areas (rural or remote)
3	Land Size (km2)
4	Agricultural Land (km2)
5	Distance from nearest town
6	Number of households (million)
7	Average dependency per household (HH) --> average HH size
8	% of HH with grid electricity (% of the population with access to electricity)
9	% of the house that is owned (not rent)
10	% of HH with household income of more than RM3000
11	% of HH with household income between RM2999 and RM1000
12	% of HH with household income below RM1000
13	Number of populations
14	The average age of head of household
15	Population (pops) density per sq. km
16	Birth-rate (per 1k people per year)
17	% of pops aged between 20 and 44
18	% of pops with secondary education
19	Labor force
20	Number of secondary school student enrolled
21	Number of local authority office (e.g., FELDA, JKKK) available
22	Number of post offices available
23	Number of years with community BB center available
24	% of household with computer access

25	% of fixed BB penetration by household --> Fixed BB per 100 people
26	% of wireless BB penetration by household (Per 100)
27	% of fixed telephony penetration by household --> telephone lines per 100 people.
28	GDP
29	GDP per capita
30	Agriculture - GDP contribution %
31	Livestock - GDP contribution %
32	Cottage Craft - GDP contribution %
33	Fishing - GDP contribution %
34	Homestay - GDP contribution %
35	Other Activity - GDP contribution % for International Tourism Receipts
36	Total income from agriculture
37	Distance from nearest wholesales agriculture-market
38	Farming experience of household (years)
39	A dummy variable for the occupation of head of household (full-time farmer/part-time)
40	A dummy variable for characteristics of household (full time/part-time)
41	Average of Monthly Rainfall per season (mm)
42	Average of Daily Temperature Rate per Season
43	Total material inputs such as seeds, pesticide, and fertilizer
44	Length of the tarred road (km/million person)
45	Availability of adjacent wireless BB network
46	% of household with pipe water
47	Average age or life expectancy
48	Gender ratio (man / woman)
49	% of pops aged 64 or younger
50	% of pops with post-secondary education
51	Labor force count in last three years
52	Entrepreneurs count in last three years

53	Household counts in last three years
54	Non-ICT investment in last three years
55	ICT investment in last three years
56	Government Deficit
57	Investments
58	Stock
59	Economic Freedom
60	Cultivated rice land value
61	Total family labor inputs
62	Damaged rice average (% to total planted acreage)
63	The average distance between house and farm
64	Average distance from the main road
65	Cultivated land value crop
66	Domestic material consumption
67	Output of Paddy (kg/ha)
68	GNI
69	GNI per capita (this indicator is selected as a response to all features)
70	Local Delicacy
71	Festivals

Note: Actual implementation is subject to data availability. The historical data for every selected variable factor needs to be available to ensure the accuracy of the target function. A function is a mathematical notation. Thus, the total number of variable factors for real implementation is most probably a smaller number.

Every one of these features fits into one of the PESTEL categories of political, economic, social, technological, environmental or legal.

Among the 71 features, GNI per capita is selected as the response because it is the indicator (Box 10) used by the World Bank to define the economic status of a country. In this research, the GNI per capita is the feature response (also known as the dependent variable) whereas the other 70 features are possible feature sets (also known as

independent variables).

Box 10: World Bank Country Classifications by Income Level 2015-2016

Each year on July 1, the World Bank revises the analytical classification of the world's economies based on estimates of gross national income (GNI) per capita for the previous year. The updated GNI per capita estimates are also used as input to the World Bank's operational classification of economies that determines lending eligibility. As of 1 July 2015, low-income economies are defined as those with a GNI per capita, calculated using the World Bank Atlas *method, of $1,045 or less in 2014; middle-income economies are those with a GNI per capita of more than $1,045 but less than $12,736; high-income economies are those with a GNI per capita of $12,736 or more. Lower-middle-income and upper-middle-income economies are separated at a GNI per capita of $4,125.* (Source: World Bank)

Collecting Data: Secondly, data is collected according to the pre-defined feature sets and feature response as shown in Table 16. The data for the corresponding feature is extracted from the World Bank's National Accounts data.

Generating Data Co-relationship: Thirdly, the collected data and labeled response are imported to a curve-fitting software to build an empirical model of a fitness function that correlates the selected features to the labeled response. Design-Expert (DEX), a curve-fitting software produced by Stat-Ease[149] is used to establish the statistical correlations between the geographical feature sets and the socioeconomic response.

A curve-fitting software is used to estimate the target function. Curve-fitting is a form of statistical modeling. The curve-fitting modeling in DEX is based on Analysis of Variance (ANOVA). During the curve-fitting modeling process, the collected data are generalized. It means the curve-fitting software screens for important factors and formulates an optimal product function. The DEX software has a graphical tool that helps identify the impact of each factor in correlation with the desired response.

Simulating Virtual Data: Lastly, the equation generated by DEX is used as a fitness function for the Genetic Algorithm (GA) to simulate more data that will be used to train the SVM. The simulated data serves as a large pool of samples to train the SVM.

3.2 MACHINE LEARNING

In this stage, the World Bank's data and GA-generated data are used to train the Support Vector Machine (SVM). SVM is selected as the machine learning technique because of its strong ability to generalize real-world problems. SVM is capable of responding to uncertainties, and it is recommended as the learning technique in this research.

This research proposes to use LIBSVM (an integrated software that is made available online[76]) as the SVM to be trained for generalization. The SVM will learn from its experience from the training samples collected from the World Bank databank and generated by GA.

There are two steps in this stage of the machine learning process:

a. Splitting data for training and test
b. Training and Validating SVM

Splitting Data: The GA-generated data is divided into different proportions for machine training and testing. This research proposes to split the data into different ratios to train and test the SVM. For results comparison and observations, the training-testing data sets are divided by using four different ratios such as 90:10, 80:20, 70:30, 60:40 and 50:50.

There are three data sets used for research in this thesis:

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

The purpose of using three data sets is to compare the training accuracy with the use of

different sample sizes and source of samples.

Training and Validating SVM: The training datasets are divided into three different scales of cross-validation of 10, 100 and 1000. Figure 56 below shows an example of how the 2000 sets of virtual data are divided 50:50 for training and testing purposes. Furthermore, the training data sets are divided by using three different scales for validation purposes.

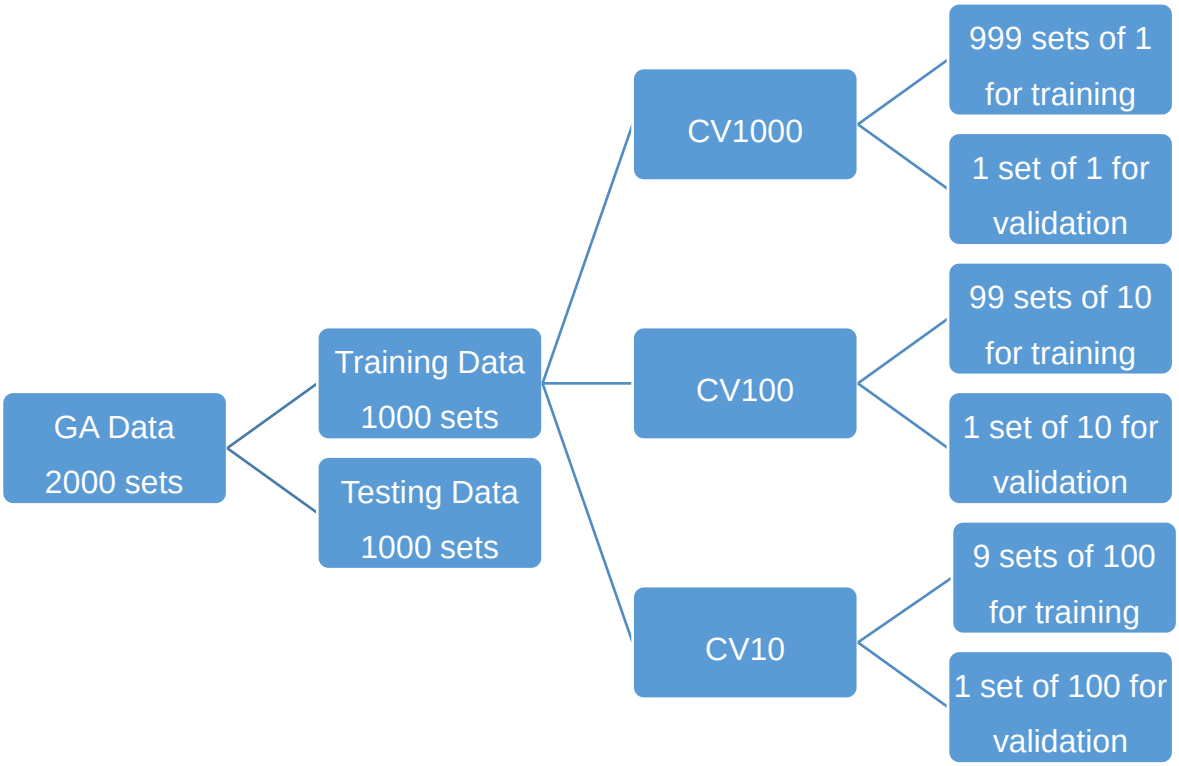


Figure 56: Using three different scales to cross-validate machine learning

During the process of SVM training and validation, a random permutation of the sample set is partitioned into K subsets ("folds") of equal size. A single subset is retained as the validation data for validating the machine, and the remaining K-1 subsets are used for training the machine. The process of training and validation is repeated K times, and the cross-validation accuracy is observed. Higher accuracy of cross-validation represents a better training result for the SVM to perform as a classifier.

For example, when the training data is used in scales of CV10 (meaning ten-fold cross-validation), the training data is divided into ten subsets, of which each subset has 100 data sets. Nine subsets are 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 training data sets of different scales are applied to the hypothesis sets in SVM for learning. Linear, Polynomial, and RBF are the hypothesis sets (or kernels) in SVM that are used as learning algorithm to find a target function which acts as an optimal classifier.

3.3 MACHINE TESTING

The testing data sets are provided to the SVMs that have been trained using different kernels such as linear, polynomial, and RBF. The testing data sets are unseen to the SVMs. The SVMs with different kernels will classify the data sets, and the accuracy of the classification results will be observed. The testing accuracy represents the SVM's capabilities of recognizing the pattern of unseen data and classifying the data into either class +1 (having socioeconomic potential) or -1 (not having socioeconomic potential).

If an SVM can perform accurately in the samples given (i.e., the training and testing data), it is assumed that the SVM can also perform accurately outside the samples (i.e., real-life data that is unknown to the SVM).

3.4 MACHINE APPLICATION

Finally, real-life field data sets of geographic areas in Malaysia are provided to the SVMs that have been tested. The SVMs will classify the real-field data sets, and the results will be observed.

4 RESULTS AND OBSERVATIONS

4.1 DATA COLLECTION

Defining and Collecting Data: Out of the 71 features defined in the research methodology, there are 20 features (including 1 response) that are included for use in this study. There are 174 countries in the World Bank databanks of World Development Indicators online[150] having complete data for these features. Countries without complete data are omitted. The list of 19 features and 1 response is shown in Table 17.

Table 17: Feature Sets with Data Available from World Bank

No	Features	Label
1	Land Size (km^2)	A
2	Agricultural Land (km^2)	B
3	% of the population with electricity access	C
4	Population size	D
5	Population density per sq. km	E
6	Birth-rate (per 1k people per year)	F
7	Labor force	G
8	Number of secondary school students enrolled	H
9	Fixed BB per 100 people	J
10	Wireless BB per 100 people	K
11	Telephone lines per 100 people	L
12	GDP	M
13	GDP per capita	N
14	Economic Activity – tourism	O
15	Average monthly rainfall	P
16	Average daily temperature	Q
17	Length of the tarred road (km/million person)	R
18	Life Expectancy	S

19	GNI	T
R	GNI per capita	RESPONSE

The data for these 19 features and 1 response are extracted from the databank for use as prototyping data sets. Table 18 shows the real-life field data for GNI per capita which is used as a response to the 19 features selected. Table 19 shows how the partial real-life data of the GNI per capita for each country are tied to the corresponding geographical features.

Countries of high income and upper middle income are labeled as having socioeconomic potential (+1); countries of low income and lower middle income are labeled as not having socioeconomic potential (-1).

Table 18: World Bank GNI per capita as Socioeconomic Response

Operational Guidelines	
Date:	1-Jul-2015
Bank's fiscal year:	FY16
Data for the calendar year:	2014
RESPONSE	*LABEL*
Low income	≤ 1,045 USD
Lower middle income	1,046-4,125 USD
Upper middle income	4,126-12,735 USD
High income	> 12,735 USD

Source: World Bank

Table 19: Partial real-life field data for the 19 features and 1 response by country.

Countries	GNI per capita	Land size	Agro land	Electric	Pop size	Pop dense	Birth	Labor	School	Fixed BB
Afghanistan	-1	652225	379100	55.17	30550000	46.83966	35.25	8334400	2508900	0
Albania	1	28748	11873	98.72	2774000	96.49367	12.88	1292900	322077	6.57
Algeria	1	2382000	414316.4	100	39210000	16.46096	24.74	12355000	4165500	4.01
Angola	1	1247000	591900	99.63	21470000	17.21732	45.99	8844200	484717	0.41
Argentina	1	2780000	1492000	99.92	41450000	14.91007	17.72	19540500	3750900	15.07
Armenia	-1	29743	16821	99.47	2977000	100.0908	13.31	1560000	228070	9.13
Australia	1	7692000	3966200	99.89	23130000	3.00702	13.2	12431000	1566100	25.76
Austria	1	83855	31544.7	99.97	8474000	101.0554	9.4	4462700	446948.5	27.54
Azerbaijan	1	86600	47698	99.08	9417000	108.7413	18.3	4952400	802580	19.83
Bahamas, The	1	13940	140	98.93	377374	27.07131	15.34	223955	34406	4.11
Bahrain	1	765.3	86	100	1332000	1740.494	15.04	750065	81896	21.39
Bangladesh	-1	147570	91080	100	1.57E+08	1061.191	20.14	78976800	11758700	1.19
Barbados	1	431	140	99.89	284644	660.4269	12.19	162318	19696	26.97
Belarus	1	207560	87260	99.83	9466000	45.60609	12.5	4482900	572580	28.84
Belgium	1	30528	13365	99.85	11200000	366.8763	11.2	4972600	652569	35.99
Belize	1	22970	1600	98.98	331900	14.44928	23.09	154602	35574	2.98
Benin	-1	112620	37500	80.53	10320000	91.63559	36.44	4445000	846328	0.4
Bhutan	-1	38394	5196	98.91	753947	19.6371	18.13	404127	68068	3.26
Bolivia	-1	1099000	376700	99.33	10670000	9.708826	24.24	5145900	1112700	1.59

When providing the prototyping data for curve-fitting to find the optimal mathematical equation, the upper and lower boundaries of each feature were capped according to the maximum and minimum data points found in the data. A single point of response was also fixed to reflect the socioeconomic potential of the areas (represented by its datasets). The curve fitting software successfully performed regression analysis by using the quadratic model to find the "best fit" line or curve for a series of data points.

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

Note: R is the socioeconomic response to the geographical features for each geographic area.

In the statistical curve fitting process, a series of experimental design is conducted with ANOVA, which simultaneously tests the relationship between a geographical feature and multiple independent features. This variance analysis is a statistical approach which extracts the contribution of each feature on the measured response. The original full equation of the empirical model is shown in Appendix 2.

Besides the formation of an empirical model, the curve-fitting process also produces a correlation coefficient of each feature in response to the socioeconomic response. The correlation coefficient is a statistical measurement of a linear relationship between paired data. Figure 57 shows an example of the correlation coefficient between GDP per capita and GNI per capita.

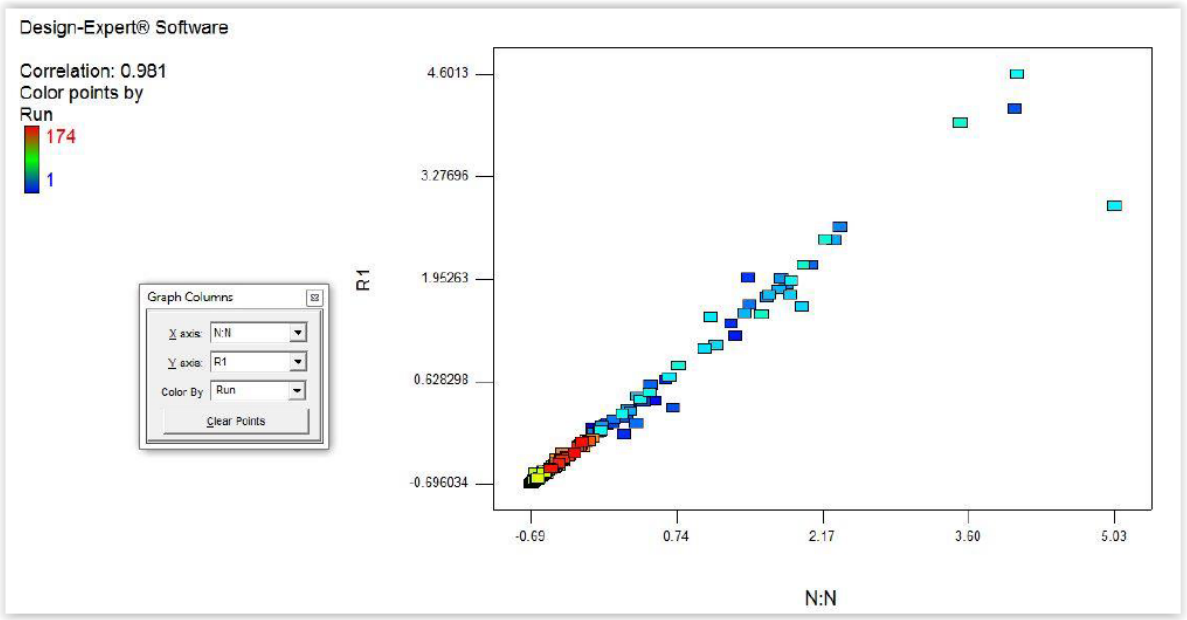


Figure 57: Correlation between a single variable factor and response

A correlation coefficient with a positive value denotes a positive linear correlation between a geographical feature and the socioeconomic response; whereas a correlation coefficient with a negative value denotes negative linear correlation. In other words, a linear correlation reflects a directional relationship between a geographical feature and the socioeconomic response, whereas a negative linear correlation reflects a reversed directional relationship.

The magnitude of the correlation coefficient determines the strength of the correlation. The closer the value is to 1 or –1, the stronger the linear correlation between a geographical feature and the socioeconomic response. A value of 0 denotes no linear

correlation. The range of the correlation coefficient tabled below shows the significance of a correlation coefficient as suggested by Evans (1996) [151].

Table 20: Magnitude of Correlation Coefficient and Its Significance of Linear Correlation

The range of Correlation Coefficient	The significance of the linear correlation
0.00 to 0.19	Very weak
0.20 to 0.39	Weak
0.40 to 0.59	Moderate
0.60 to 0.79	Strong
0.80 to 1.00	Very strong

For example, a correlation value of 0.45 would be a moderate positive correlation or moderate directional relationship; and a correlation value of -0.25 would be a weak negative correlation or reversed directional relationship.

The overall correlation coefficient between each geographical feature and the socioeconomic response is tabled in Table 21.

Table 21: Correlation Coefficient of Geographical Features to Socioeconomic Response

Local Characteristics	**Label**	**Coefficient**
GDP per capita	N	0.981
Fixed BB per 100 persons	J	0.746
Wireless BB per 100 persons	K	0.712
Telephone lines per 100 persons	L	0.665
Life expectancy	S	0.616
Length of the tarred road (km per million people)	R	0.454
Economic Activity – tourism	O	0.373
GDP	M	0.275
GNI	T	0.273
% of the population with electricity access	C	0.271

Population density ($per\ km^2$)	E	0.197
Land size (km^2)	A	0.105
Agricultural land (km^2)	B	0.083
Average monthly rainfall	P	-0.015
Labor 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

Appendix 6 shows how one geographical feature correlates with other interdepending features; and together how the 19 features are correlated to the socioeconomic response. Appendix 8 shows the correlation coefficient between geographical features and GNI per capita with a graphical presentation on the distribution of the real-life field data.

It is observed that GDP per capita has the highest-impact (0.981) of directional relationship with the socioeconomic response, followed by BB penetration and telephony services. Fixed BB (0.746), wireless BB (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 strong directional relationships with GNI per capita, and the results are in line with the World Bank and ITU's reports which indicate that BB penetration will accelerate economic GDP growth. The GDP per capita is an economic factor whereas BB and telephony services are technological factors that impact GNI per capita.

Life expectancy (0.616) and birth rate (-0.532) are two social factors that impactful to GNI per capita. Life expectancy has a strong correlation with GNI per capita, while the correlation between birth rate and GNI per capita is moderate in the reversed direction.

It looks like the higher the birth rate, the less the likelihood of BB investment. Nargund (2009) concluded that birth rates are declining in both the developing and developed

world, and fertility rates tend to be higher in poorly resourced countries[152]. Concerning the world fertility patterns reported by the United Nations, the fertility rate is the lowest in high-income countries and highest in low-income countries[153]. That explains why the birth rate has a moderate reversed directional relationship with socioeconomic status. On the other hand, life expectancy is the longest in high-income countries and shortest in low-income countries[154]. Another studied shows that the life expectancy of people in developed countries is higher than the people of developing countries[155]. That explains why life expectancy has a strong correlation with high socioeconomic response.

Length of the tarred road (0.454) has a moderate directional relationship with a socioeconomic response. The features of tourism activity (0.373), GDP (0.275), GNI (0.273) and % of the population with electricity access (0.271) have a weak correlation with GNI per capita.

The features of population density, land size, agricultural land size, labor force, population size, number of secondary school students, average monthly rainfall and average daily temperature have very week correlation (below 0.200) with GNI per capita.

When changing the response R from GNI per capita to fixed BB penetration, it is observed that each geographical feature has different magnitudes of impact towards the BB penetration response as shown in Table 22.

Table 22: Correlation between geographical features and fixed BB penetration

Local Characteristics	**Label**	**Coefficient**
Telephone lines per 100 persons	L	0.881
GNI per capita	J	0.746
GDP per capita	N	0.738
Life expectancy	S	0.735
Wireless BB per 100 persons	K	0.678
Length of the tarred road (km per million)	R	0.499
Economic Activity – tourism	O	0.395

% of the population with electricity access	C	0.347
GDP	M	0.276
GNI	T	0.274
Population density ($perkm2$)	E	0.206
Land size ($km2$)	A	0.089
Agricultural land($km2$)	B	0.042
Labor 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 6.

It is observed that telephony penetration has the strongest (0.881) linear directional relationship in correlation with fixed BB penetration. That is because dial-up connections through the telephone network are the original form of residential access to the Internet[156]. Today, copper and fiber-optic glass are still the two main substances used by TELCOs to deliver telephony services to each household and serve as backhaul for wireless BS.

GNI per capita (0.746), GDP per capita (0.738), life expectancy (0.735) and wireless BB penetration (0.678) have a strong correlation with BB penetration. On the other hand, birth rate (-0.741) has a strong correlation with fixed BB penetration in the reversed direction, i.e., a geographic area with higher birth rate indicates less BB penetration.

Length of the tarred road (0.499) has a moderate directional relationship with a socioeconomic response. Deploying wired networks involve activities such as digging up streets, gaining access to telephone poles or conduit space, and gaining physical access to homes[156]. Availability of tarred road affects the implementation of these deployment activities.

The features of tourism activity (0.395), % of the population with electricity access (0.347), GDP (0.276), GNI (0.274) and population density (0.206) have a weak correlation with GNI per capita.

Those geographical features with relatively very weak correlation with fixed BB penetration are land size, agricultural land size, labor force, average monthly rainfall, population size and number of secondary school students.

The pattern of fixed BB penetration responses to various geographical features is similar to the pattern of GNI per capita. But it is unique to observe that daily temperature has some impact on fixed BB penetration and GNI per capita, even though the correlation coefficients are weak and in a reversed direction. Some research[157][158] has found that daily temperature does have a reversed directional relationship with economic growth, i.e., a geographic area with higher temperature indicates slower economic growth.

At this point the first three sub-steps in data collection has been successfully completed, which covers the execution of defining data, collecting data and generating data correlations. The correlations can serve as a guideline when applying the empirical model to study game theory in PPP. The data for geographical conditions with medium to high impact, if missing, will affect the distortion of the empirical model. In other words, if the collaborative efforts are put in to improve the data points of those features with high impact, together the PPP will also improve the socioeconomic response in favor of BB investment.

The function of the correlations is then translated to a MATLAB format (Appendix 3) for GA to simulate virtual samples for machine training. MATLAB (matrix laboratory) *is a multi-paradigm numerical computing environment and proprietary programming language developed by MathWorks. MATLAB integrates computation, visualization, and programming in an easy-to-use environment where problems and solutions are*

expressed in familiar mathematical notation[159].

Through the research experiment, GA has successfully generated 64,200 samples. Each sample is labelled with a fitness value which virtually represents the socioeconomic status. The fitness breakpoint is the fitness value that separates the label for the generalized datasets shown in Figure 58.

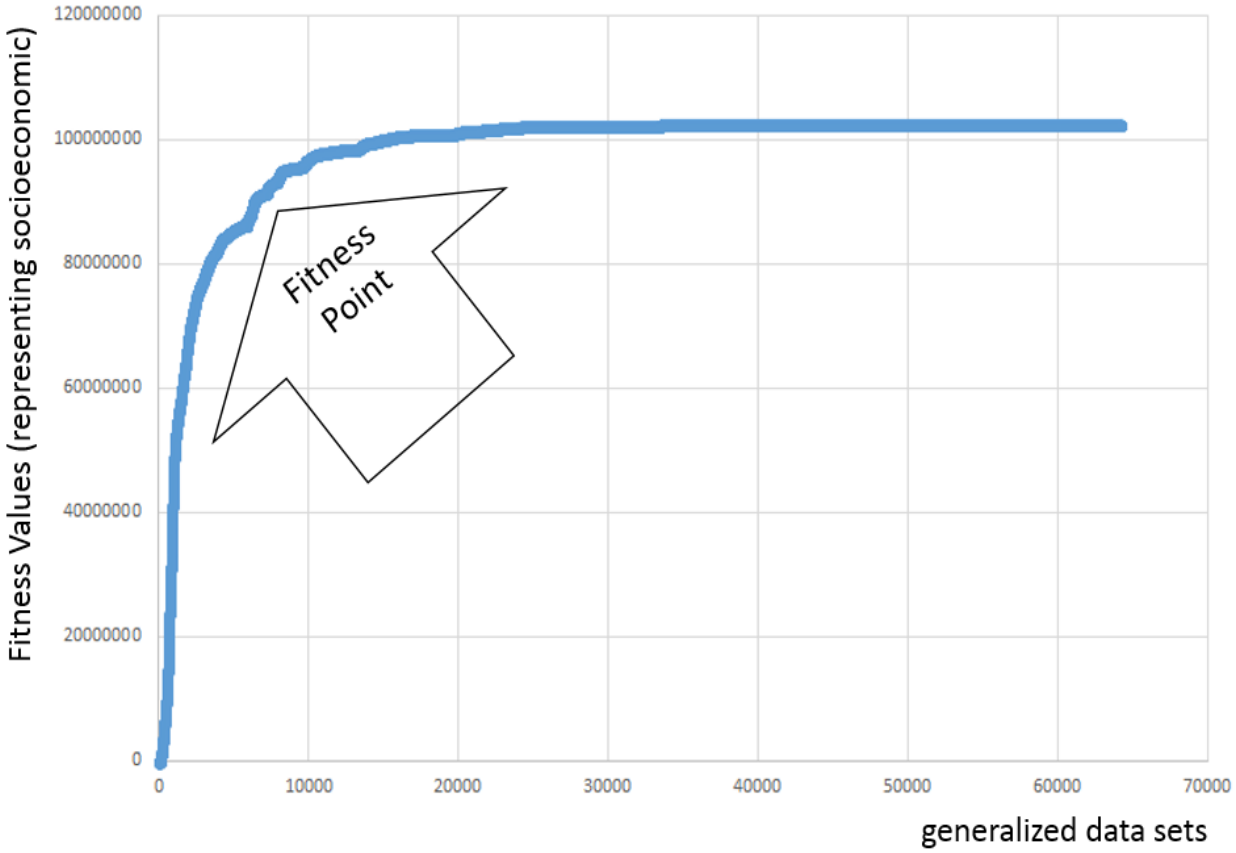


Figure 58: Fitness Values vs. Generalization

All available data are arranged into three groups (Table 23) to train the SVM to observe if the different data groups cause any variances to the accuracy of machine training. Data in each data group is further split into different ratios for machine training and testing respectively. For example, 80% of the data in the 80:20 ratio is used for machine training whereas the balance 20% is used for SVM testing. Each sub-set of training data is further divided into different folds for cross-validating the machine training.

Table 23: World Bank and GA Data are Arranged into 3 Groups for Machine Learning

Data Group A - DGA					Data Group B - DWB					Data Group C - DGAWB				
This group contains 64,200 virtual samples generated by GA					This group contains 64,374 samples combining both samples from GA and the World Bank					This group contains 174 real-life samples obtained from the World Bank databank				
50:50	60:40	70:30	80:20	90:10	50:50	60:40	70:30	80:20	90:10	50:50	60:40	70:30	80:20	90:10
CV10					CV10					CV10				
CV100					CV100					CV100				
CV1000					CV1000					CV1000				

4.2 MACHINE LEARNING AND TESTING

In this research, LIBSVM with linear, polynomial and RBF kernels is used for 2-class classification process where the artificially generated data, as well as the World Bank data for geographical features, are classified as either potentially successful or potentially unsuccessful. All three groups of data are used for machine learning: training and cross-validation. The SVM training has been validated according to the pre-defined cross-validation scales of 10, 100 and 1000.

Table 24 shows the results of training and cross-validation accuracy (CVA) by using different kernels for a different group of data which are split into a different ratio and further divided into different scales for cross-validation.

Table 24: Training and Cross-Validation Results

	Group A Data			Group B Data			Group C Data			
SVM	A1-GA 50:50			B1-GAWB 50:50			C1-WB 50:50			
Kernal	CV10	CV100	CV1000	CV10	CV100	CV1000	CV10	CV100	CV1000	Ratio 50:50
Linear	100.0%	100.0%	100.0%	98.8%	98.4%	98.4%	90.8%	90.8%	90.8%	
Polynomial	100.0%	100.0%	100.0%	95.9%	96.0%	96.0%	71.3%	71.3%	71.3%	
RBF	100.0%	100.0%	100.0%	99.0%	99.0%	99.0%	89.7%	89.7%	89.7%	
SVM	A2-GA 60:40			B2-GAWB 60:40			C2-WB 60:40			
Kernal	CV10	CV100	CV1000	CV10	CV100	CV1000	CV10	CV100	CV1000	Ratio 60:40
Linear	98.6%	98.6%	98.6%	96.4%	95.9%	95.7%	94.2%	94.2%	94.2%	
Polynomial	98.8%	98.8%	98.8%	95.6%	95.4%	95.4%	70.2%	70.2%	70.2%	
RBF	98.8%	99.0%	99.0%	97.7%	97.8%	97.6%	89.4%	87.5%	87.5%	
SVM	A3-GA 70:30			B3-GAWB 70:30			C3-WB 70:30			
Kernal	CV10	CV100	CV1000	CV10	CV100	CV1000	CV10	CV100	CV1000	Ratio 70:30
Linear	99.1%	98.9%	98.8%	96.3%	96.1%	96.0%	91.0%	93.4%	93.4%	
Polynomial	98.9%	99.0%	99.0%	95.5%	95.7%	95.7%	70.5%	69.7%	69.7%	
RBF	98.9%	98.9%	98.9%	97.8%	98.0%	98.0%	86.9%	87.7%	87.7%	
SVM	A4-GA 80:20			B4-GAWB 80:20			C4-WB 80:20			
Kernal	CV10	CV100	CV1000	CV10	CV100	CV1000	CV10	CV100	CV1000	Ratio 80:20
Linear	99.1%	98.9%	98.9%	96.4%	96.4%	96.3%	90.6%	89.9%	89.9%	
Polynomial	99.1%	99.1%	99.1%	95.7%	95.8%	95.8%	74.1%	74.1%	73.4%	
RBF	98.7%	98.8%	98.7%	97.5%	97.7%	97.8%	86.3%	86.3%	86.3%	
SVM	A5-GA 90:10			B5-GAWB 90:10			C5-WB 90:10			
Kernal	CV10	CV100	CV1000	CV10	CV100	CV1000	CV10	CV100	CV1000	Ratio 90:10
Linear	97.8%	97.8%	97.8%	95.2%	95.2%	95.2%	89.2%	89.8%	89.2%	
Polynomial	97.7%	98.2%	98.2%	94.8%	94.7%	94.8%	78.3%	78.3%	78.3%	
RBF	97.9%	98.1%	98.0%	97.0%	97.1%	97.0%	86.6%	86.0%	86.0%	

Based on the results shown in Table 24, it is observed that the SVM has been trained with high accuracy as reflected with the accuracy of cross-validation. The cross-validation accuracy (CVA) is better when the machine is trained with a larger sample size (e.g., Group A and Group B). The result is in line with the behavior of machine learning where large training samples will improve the machine's performance. However, it is also observed that the CVA for Group A and Group B are best when their data are split 50:50. It is also observed that the CVA drops as the ratio of data split increases from 50:50 to 60:40 and towards 90:10. The accuracy drop is probably caused by the presence of outliers causing noises in the machine training process.

The CVA for Group C (with 174 sets of data from the World Bank) is lowest as compared to Group A and B. The small sample size could be the reason of not having sufficient training data which affects the machine's performance. As mentioned in section 2.2, As mentioned earlier, the performance of a machine to do certain tasks improves with experience. A machine can gain more experience when it is trained with large samples.

The testing data sets from the same three groups are next supplied to the trained SVM for the machine to classify the testing data which are unseen to the SVM. Table 25 shows the results of machine test accuracy (MTA) which reflects the machine's performance in generalizing the unseen data.

Table 25: Machine Testing Results

	Group A Data	Group B Data	Group C Data	
SVM Kernal	A1-GA Test	B1-GAWB Test	C1-WB Test	Ratio 50:50
	Machine Test Accuracy (MTA)			
Linear	90.8%	77.2%	85.1%	
Polynomial	93.6%	91.0%	60.9%	
RBF	92.3%	86.5%	81.6%	
SVM Kernal	A2-GA Test	B2-GAWB Test	C2-WB Test	Ratio 60:40
	Machine Test Accuracy (MTA)			
Linear	95.3%	93.8%	87.1%	
Polynomial	92.6%	91.5%	67.1%	
RBF	97.4%	95.6%	82.9%	
SVM Kernal	A3-GA Test	B3-GAWB Test	C3-WB Test	Ratio 70:30
	Machine Test Accuracy (MTA)			
Linear	93.7%	92.9%	86.5%	
Polynomial	90.2%	89.4%	75.0%	
RBF	96.7%	95.6%	82.7%	
SVM Kernal	A4-GA Test	B4-GAWB Test	C4-WB Test	Ratio 80:20
	Machine Test Accuracy (MTA)			
Linear	96.4%	95.8%	91.4%	
Polynomial	96.4%	94.6%	82.9%	
RBF	96.3%	96.0%	82.9%	
SVM Kernal	A5-GA Test	B5-GAWB Test	C5-WB Test	Ratio 90:10
	GA	GAWB	WB	
Linear	100.0%	99.5%	94.1%	
Polynomial	100.0%	97.2%	88.2%	
RBF	100.0%	99.1%	88.2%	

Similar to the cross-validation results, the machine test accuracy is better when the machine is tested with a larger sample size (e.g., Group A and Group B).

MTA (machine testing accuracy) for Group A and Group B is best when data is split in the ratio of 90:10. The MTA drops as the ratio of data split is reduced from 90:10 to 80:20 and towards 50:50. The MTA contradicts the results of cross-validation. With a split data ratio of 90:10, the unseen data to SVM is less than the unseen data coming from a data split ratio of 50:50. The less data SVM needs to classify, then the fewer errors the SVM will probably make.

In general, the MTA for Group C (with 174 sets of data from the World Bank) is lowest as compared to Group A and B. Again; the small sample size could be the reason of not having sufficient training data which affects the machine's performance.

Based on the test results shown in Table 25, it is observed that the SVM can perform with high generalization accuracy.

4.3 MACHINE APPLICATION

4.3.1 APPLICATION OF REAL-LIFE FIELD DATA FROM 13 STATES IN MALAYSIA

Malaysia is a constitutional monarchy with a federal government located in Putrajaya. Furthermore, under this federal government, Malaysia has 13 state governments and 3 federal territories. This section of this thesis looks into the viability of the 13 states and 3 federal territories for BB investment.

With the framework of the machine learning technique being established and proven, the trained and tested SVM can be applied to classify real-life field data in the context of geographic areas in Malaysia.

Through the literature review and physical meetings with the Department of Statistics Malaysia, it is found that only ten geographical features are available for states in Malaysia as compared to 19 features obtained from the World Bank databank.

The % of agricultural land is taken as a new feature to replace the features of land size and agricultural land size.

Table 26: Global Countries Feature List vs. the Malaysia States Feature List

No	Features	World Bank Database	Malaysia State Database
1	Land Size (km2)	yes	yes (% of agricultural land)
2	Agricultural Land (km2)	yes	
3	% of the population with electricity access	yes	yes
4	Population size	yes	yes
5	Populations density per sq. km	yes	yes
6	Birth rate (per 1k people)	yes	yes
7	Labor force	yes	yes
8	Number of secondary school students enrolled	yes	yes
9	Fixed BB per 100 people	yes	yes
10	Wireless BB per 100 people	yes	
11	Telephone lines per 100 people	yes	yes
12	GDP	yes	
13	GDP per capita	yes	yes
14	Economic Activity – tourism	yes	
15	Average monthly rainfall	yes	
16	Average daily temperature	yes	

17	Length of the tarred road (km/million person)	yes	
18	Life Expectancy	yes	
19	GNI	yes	
R	GNI per capita	yes	

It is unknown if the use of real-life data with ten features will cause bias or errors in a machine that has been trained with 19 features. Thus, a new curve-fitting model is formulated with only ten features from the World Bank database.

The correlation coefficient for the features in the new empirical model is found to be identical as compared to the original experiment with 19 features. This observation confirms that the impact of the features on socioeconomic conditions remain consistent (refer to Table 27).

Table 27: Correlation Coefficient for 19 Features vs. 10 Features

No	Features	Correlation Coefficient for 19 features	Correlation Coefficient for ten features
1	Land Size (km2)	0.105	-0.236
2	Agricultural Land (km2)	0.083	
3	% of population with electricity access	0.271	0.271
4	Population size	-0.041	-0.041
5	Populations density per sq. km	0.197	0.197
6	Birth-rate (per 1k people)	-0.532	-0.532
7	Labor force	-0.028	-0.028
8	Number of secondary school students enrolled	-0.041	-0.041
9	Fixed BB per 100 people	0.746	0.746
10	Wireless BB per 100 people	0.712	
11	Telephone lines per 100 people	0.665	0.665
12	GDP	0.275	

13	GDP per capita	0.981	0.981
14	Economic Activity – tourism	0.373	
15	Average monthly rainfall	-0.015	
16	Average daily temperature	-0.122	
17	Length of tarred road (km/million person)	0.454	
18	Life Expectancy	0.616	
19	GNI	0.273	
Total number of features		19	10

Note: Appendix 7 shows how one geographical feature correlates with other interdepending features; and together how the 10 features are correlated to the socioeconomic response. Appendix 9 shows the correlation coefficient between geographical features and GNI per capita with a graphical presentation on the distribution of the real-life field data.

In this experiment, the % of agricultural land is used as a geographical feature, replacing land size and agricultural land size that are used in previous experiments. It is observed that the % of agricultural land (-0.236) has a stronger impact on socioeconomic potential as compared to land size (0.105) and agricultural land size (0.083). It looks like the relative importance of agriculture declines as a geographic area develops economically. Ernst Engel found that the proportion of income spent on food declines as incomes increase[160].

Similar to the previous experiment which used the curve-fitting technique, a new empirical model is formulated for use in GA. The empirical model formulated by curve-fitting is presented in Appendix 4.

The empirical model is provided to GA to generate new sets of virtual samples to train, validate and test the SVM. This time, GA generated 45800 virtual samples.

With the same execution as explained in section 4.1 and 4.2, the new machine is trained and tested again. It is observed that the new SVM can still perform with high accuracy (Table 28) using a cross-validation scale of 10. This experiment demonstrated that the training with CV10 delivered the best accuracy.

Table 28: SVM's Performance

	CV Accuracy	Test Accuracy
Linear	99.97%	85.99%
Polynomial	99.98%	93.81%
RBF	99.96%	82.95%

When provided with the real-life field data from 13 states and three federal territories in Malaysia, all the 3 SVMs of linear, polynomial and RBF predicted that all the geographic areas have socioeconomic potential as shown in Table 29.

Table 29: SVM's Prediction on Socioeconomic Potential for States in Malaysia

State	Prediction Results		
	Linear	Polynomial	RBF
Johor	+1	+1	+1
Kedah	+1	+1	+1
Kelantan	+1	+1	+1
Melaka	+1	+1	+1
NS	+1	+1	+1
Pahang	+1	+1	+1
Penang	+1	+1	+1
Perak	+1	+1	+1
Perlis	+1	+1	+1
Selangor	+1	+1	+1
Terengganu	+1	+1	+1
Sabah	+1	+1	+1

Sarawak	+1	+1	+1
KL	+1	+1	+1
Labuan	+1	+1	+1
Putrajaya	+1	+1	+1

According to the literature review, a geographic area is considered as having high BB adoption once its BB penetration has reached 20% or higher, and the socioeconomic status has also become more commercially feasible for BB investment or expansion. Among the 13 states and three federal territories, Negeri Sembilan is the area with the lowest BB penetration with 53.2%, way above the threshold of 20% for BB adoption. The state with the lowest GDP per capita (approximately USD3,380, the year 2014) has 57.6% BB penetration and is predicted by SVM as an area with socioeconomic potential for BB investment.

ROI for the 13 states and 3 federal territories is simulated and compared with the classified results as predicted by the machine learning technique, as shown in Table 30.

Table 30: ROI Simulation Results for 13 States and 3 Federal Territories in Malaysia

State	Predicted Results			
	ROI Payback Period (years)	Linear	Polynomial	RBF
Johor	4	+1	+1	+1
Kedah	6	+1	+1	+1
Kelantan	6	+1	+1	+1
Melaka	5	+1	+1	+1
NS	5	+1	+1	+1
Pahang	5	+1	+1	+1
Penang	4	+1	+1	+1

Perak	5	+1	+1	+1
Perlis	5	+1	+1	+1
Selangor	4	+1	+1	+1
Terengganu	5	+1	+1	+1
Sabah	7	+1	+1	+1
Sarawak	7	+1	+1	+1
KL	4	+1	+1	+1
Labuan	7	+1	+1	+1
Putrajaya	4	+1	+1	+1

Note: ROI simulation model is presented in Appendix 11. Detailed calculation on ROI simulation is not shown in this thesis as the data is the trade secret that involved a third-party company. The data and detailed calculation can be viewed offline upon reasonable request.

As mentioned in the previous chapter (refer to Figure 53), the average long-term return on investment has been around 6% in the telecom industry. TELCOs in Malaysia generally perceived a 5-year payback period as excellent ROI with 6 to 7 years as being acceptable. The maximum acceptable payback period is usually set to be equal to the duration of the CapEx amortization. A 10-year amortization period is commonly used by the TELCOs.

The ROI-simulated results show the payback period ranges between 4 to 7 years, which are within the acceptable range of TELCOs' appetite. Nevertheless, the actual ROI-simulated results might vary from one TELCO to another due to the assumptions used in the ROI model. Some of the key variances could be due to:

- CapEx which depends on vendor selection strategy, number of site count and network dimensions.
- OpEx which depends on the operating model and control policy.
- number of subscribers which depends on the appetite of market share.
- selling price which depends on the competitive strategy and so forth.

4.3.2 APPLICATION OF REAL-LIFE FIELD DATA FOR DISTRICT AREAS IN MALAYSIA

To further zoom in to the district level in every state, real-life field data in different districts are collected from DOSM. Nevertheless, out of the 19 features used to train and test the SVM, only seven features are available. Unfortunately, the available district data are for features of less dominant variables such as land size, agricultural land size, population access to electricity, labor force, secondary school students enrolled and length of tarred roads. The features of stronger dominant variables such as birth rate, BB penetration, telephony penetration, GDP per capita and life expectancy are not available for further research. Hence, due to insufficient data, the machine application is stopped at the geographic area of the state level.

Table 31: Difference of Data Availability from Different Database

No	Features	World Bank Database for Countries	DOSM Database for States	DOSM Database for Districts
1	Land Size (km2)	yes	yes	yes
2	Agricultural Land (km2)	yes	yes	yes
3	% of the population with electricity access	yes	yes	yes
4	Population size	yes	yes	yes
5	Populations density per sq. km	yes	yes	
6	Birth-rate (per 1k people)	yes	yes	
7	Labor force	yes	yes	yes
8	Number of secondary school students enrolled	yes	yes	yes
9	Fixed BB per 100 people	yes	yes	
10	Wireless BB per 100 people	yes		
11	Telephone lines per 100 people	yes	yes	
12	GDP	yes		

13	GDP per capita	yes	yes	
14	Economic Activity – tourism	yes		
15	Average monthly rainfall	yes		
16	Average daily temperature	yes		
17	Length of the tarred road	yes		yes
18	Life Expectancy	yes		
19	GNI	yes		
Total number of features		19	10	7

Note: Appendix 5 shows the similar table with highlights on features with a strong coefficient.

4.4 OBSERVATION OF SVM'S BEHAVIORS

4.4.1 MACHINE LEARNING APPEARED IN THIS RESEARCH

The SVM machine has found a pattern in the data used in this research. SVM does not memorize the data given. Instead, it learns the pattern from the training datasets and classifies the response for the testing datasets. The two obvious evidence of machine learning are:

- Consistency and accuracy of the cross-validation and testing with World Bank and GA-generated data
- Consistency and accuracy of the testing results with real-life data for states in Malaysia.

The flowchart below explains how machine learning is established during the experiment.

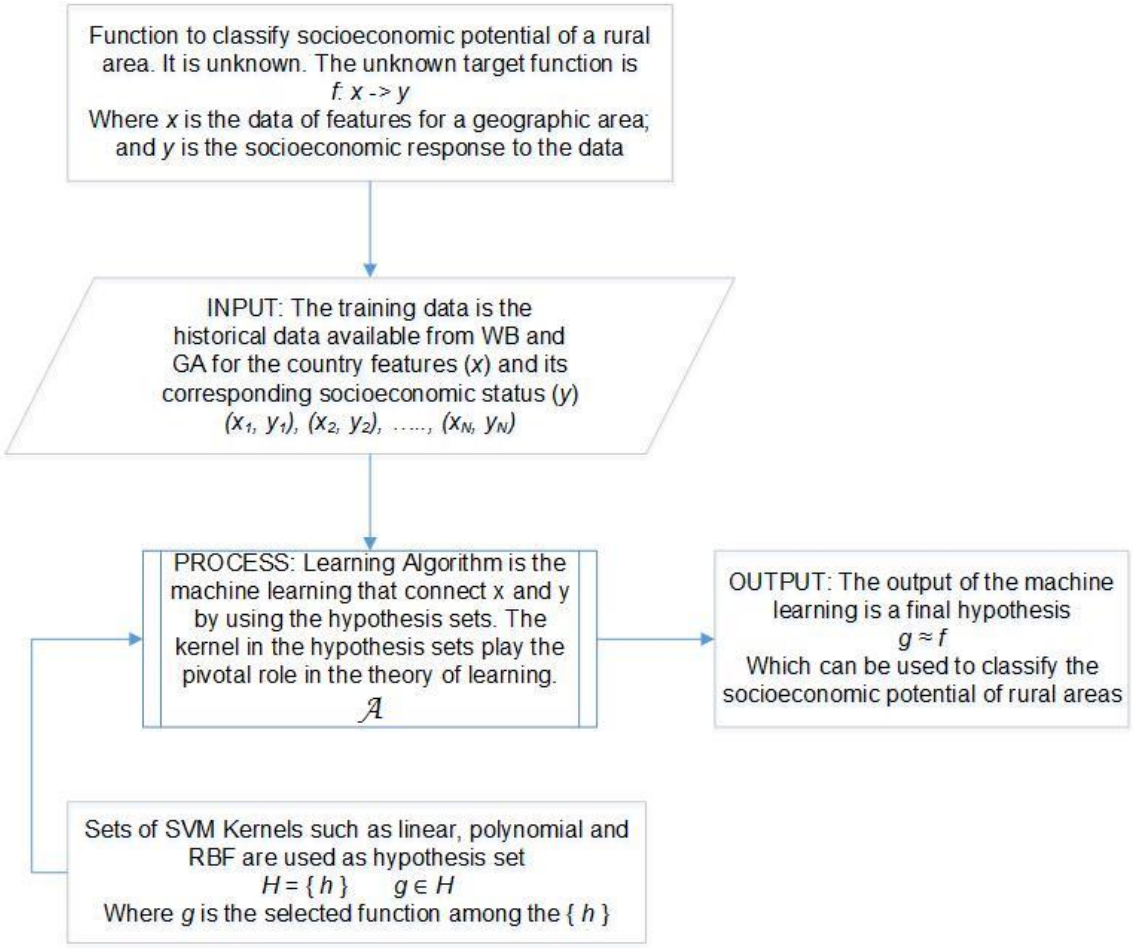


Figure 59: Machine Learning Diagram

Both learning algorithm and hypothesis sets are the solution tools (or components) in the machine learning process. And the solution tools worked successfully in this research. Together, the learning algorithm and the hypothesis sets are named as the learning model. The research results showed that a learning model had been successfully established. The SVM is the hypothesis *H*, whereas the kernels (linear, polynomial and RBF) are the sub-sets {*h*} of the hypothesis which is used for pattern recognition. The quadratic programming is the learning algorithm used in the research.

$$h = X \rightarrow \{+1, -1\}$$

$$h = \{x_1, x_2, , , x_N\} \rightarrow \{+1, -1\}$$

where,

- h denotes the hyperplane that generalizes the data of geographical features in the feature space
- X denotes data that are transformed from input space into features space with higher dimensions.
- $x_1, x_2, , , x_N$ are the specific data points of X.
- $\{+1, -1\}$ is the binary class, where +1 denotes areas having socioeconomic potential; and -1 denotes areas without socioeconomic potential.

4.4.2 DATA INFLUENCE AND FEATURE SELECTION

Through the experiment, it is found that larger training samples deliver higher training accuracy. This result is in line with the machine learning theory of "as sample sizes increase the impact of variance can be expected to decrease."[161]

An overall view of the training and testing results is shown in Appendix 10. It is observed that the delta between the cross-validation accuracy and test accuracy is small for Group A and Group B as compared to Group C which has only 174 data sets. This observation is also in line with another machine learning theory which says, when the sample size is big, both training and testing results will have about the same error (refer to Figure 60). Further details on this theory have been addressed in section 2.3.8.

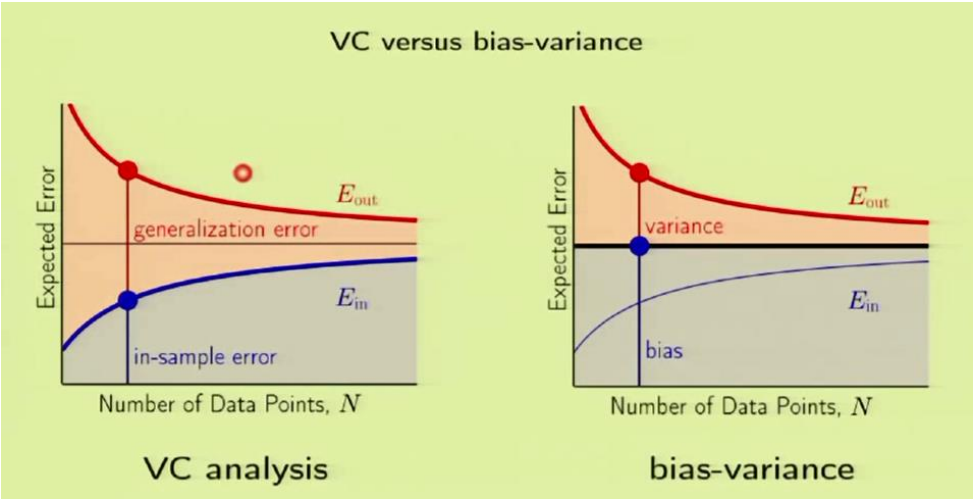


Figure 60: Delta of Bias and Variance are getting smaller as N increases. Source:[66]

The research experiment shows that SVM can learn and perform if adequate prototyping data are provided for learning. The label of the threshold for classification is important too as it affects the machines' performance. In the case of this thesis, we can say with confidence that the SVM is consistent within the data sets provided, and the framework of the machine learning technique works.

4.4.3 SVM KERNELS

The Radial Basis Function kernel also called the RBF kernel or Gaussian kernel, is observed to be the best kernel function in this research. The results are shown in Appendix 10 (SVM training and testing results). RBF outperformed linear and polynomial functions in most of the cases, especially in measuring the delta between cross-validation accuracy and testing accuracy.

The best model is found to be an RBF that is trained with GA data using a 60:40 data split on a cross-validation scale of 10. The CVA is 98.8%, and MTA is 97.4%, giving a delta of 1.4% which is the lowest among all the experiments.

Although the polynomial kernel has more hyperparameters than the RBF kernel, RBF is the optimal kernel that handled the number of hyperparameters in the data used in this research. The RBF kernel has fewer hyperparameters as compared to the polynomial kernel, hence reducing the complexity of model selection. RBF is also commonly used when the number of training data is much larger than the number of features. One key point is $0 < K_{ij} \leq 1$ in contrast to polynomial kernels of which kernel values may go to infinity $\left(\gamma x_i^T x_j + r\right) > 1$ or zero $\left(\gamma x_i^T x_j + r\right) < 1$, when the degree is large.

It is recommended to start a machine learning with a nonlinear kernel (e.g., RBF or polynomial) if the data sets are unknown to be separable or not separable. A linearly separable data set can still be treated as one kind of nonlinearly separable data set, and a nonlinear kernel can still handle the classification with high accuracy. In short, both

linear and RBF kernels are seen to be applicable with good accuracy in this research, but the RBF works better.

4.4.4 CORRELATION COEFFICIENCY IS CONSISTENT WITH PAST RESEARCH

Of the 19 features selected in this research for experiments, 12 features demonstrated some impact to high impact to BB development. These features are GDP per capita, fixed BB penetration, wireless BB penetration, telephony penetration, life expectancy, length of roads, economic activity, GDP, GNI, electricity access, population density, and birth rate. This observation is in line with findings as mentioned in the literature review chapter.

However, the impact of the labor force and secondary education are found to be in contrary to those findings reported in the literature review.

- As BB technology is evolving, the tertiary education might be a better indicator than secondary education. Goldfarb (2006) found that university education improved Internet diffusion.
- The low correlation coefficient for labor force might open up an area of new research or further literature review if the occupation is a better feature to replace the labor force.

Land size, agricultural land size, population size, rainfall, and average temperature are found to have minimum impact on BB development. The literature reviewed does not reveal the correlation between these five features against BB development. Nevertheless, the research experiment shows that % of agricultural land has a high impact as compared to land size or agricultural land by itself (refer to section 4.3.1).

Table 32: Consistency of Use of Features in Relevance to Literature Review

Geographical Features	Evidence in Literature Review	Coefficient
GDP per capita	Yes	0.981
Fixed BB per 100 persons	Yes	0.746

Wireless BB per 100 persons	Yes	0.712
Telephone lines per 100 persons	Yes	0.665
Life expectancy	Yes	0.616
Length of the tarred road (km per million people)	Yes	0.454
Economic Activity – tourism	Yes	0.373
GDP	Yes	0.275
GNI	Yes	0.273
% of the population with electricity access	Yes	0.271
Population density ($per\ km^2$)	Yes	0.197
Land size (km^2)	No	0.105
Agricultural land (km^2)	No	0.083
Average monthly rainfall	No	-0.015
Labor force	yes	-0.028
Population size	No	-0.041
# of students enrolled in secondary schools	Yes	-0.041
Average daily temperature	No	-0.122
Birth rate	Yes	-0.532

5 CONCLUSION

The traditional method to decide whether an area is viable for BB investment is based on the calculation of ROI. ROI is a handy tool as it is relatively easy to use and the calculation can be done manually or with the help of a simple computer. BB investors, either TELCOs or government agencies, can dictate the magnitude of the thresholds according to business appetites such as market share to be captured, price to sell, desired payback period and so forth.

This thesis, however, proposes another, novel method for deciding whether to implement BB, especially in rural and less developed areas. The literature review showed that statistical methods could be used in a machine that in its turn, can learn from data that is imported into the machine. Hence, this thesis proposes a machine learning technique as a basis for decisions on BB investments.

The tests with available data from the World Bank and virtual data from GA as well as tests with real-life data collected from Department of Statistics Malaysia showed similar results, i.e., the pattern of fixed BB penetration responses to various geographical features is similar to the pattern of GNI per capita. The correlation coefficients of the geographical features are also consistent in the tests with various data used in experiments. Besides, the training and testing accuracy are high as both training and testing results show about the same error rate. Therefore, it is proven that the machine can indeed learn to analyze real-life data.

In summary, we can confidently argue that such a machine learning technique in actuality is working. This research result is a valuable contribution to the research of machine learning techniques, as it is proven with real-life data in conjunction with BB investment.

5.1 SIGNIFICANCE OF RESEARCH RESULTS

Through the process of literature review, research methodology formulation and

execution of the research methodology, it is proven that the machine learning technique is a feasible model for use in the telecom industry to classify geographic areas according to their socioeconomic potential. Training data are available from the World Bank databank and the Department of Statistics Malaysia to initiate the process of machine learning. Even though there are shortcomings in the data sets regarding feature sets and sample size, the existing data are good enough to be used as prototyping data to be put through statistical modeling which results to the formulation of interdependencies (correlation coefficient) among the features and targeted response. The statistical modeling has been successful in generalizing the data and screening for important factors to establish the optimal product formulation, which is an equation that correlates the geographical features corresponding to the socioeconomic response. By applying the equation to a genetic algorithm, virtual samples in a large size have been generated for SVM training and testing. The high accuracy achieved in cross-validation and testing prove that the SVM has been properly trained. Finally, when real-life field data for states in Malaysia are provided to the SVM, the machine can successfully classify the states according to their socioeconomic potential.

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 geographic areas. Besides GDP per capita, the BB penetration is found to have the highest co-relationship with GNI per capita.

The research results show that the land size and population size have a very weak correlation with GNI per capita and fixed BB penetration. That means this machine learning model is applicable regardless the size of a geographic area, be it a country or state or district.

Using a machine learning technique to classify the socioeconomic potential of a geographic area according to its geographical features is a novelty of this research. The

research method and results prove that the machine learning technique applies to 3 areas of contribution as mentioned in section 1.3.2:

1. Application as National Development Planning Tool in the Public Sector
2. Application as a Project Management Tool
3. Application as a Business Planning Tool in the Private Sector

Machine learning technique can perform independently, as well as compliment the ROI model for business decision making, either helping the TELCO to expand its BB investment in new geographic areas or helping the policymakers to increase the efficiency of BB policy and the use of USP funds.

5.2 RELEVANCE OF RESEARCH RESULTS TO THE STAGES OF BROADBAND

The curve fitting theory establishes the correlations of the geographical features with an indication of the impact magnitude for each feature. If necessary, the principal component analysis in machine learning can be used to further demarcate the features with high influence from features with low influence. By combining the application of the curve-fitting theory and machine learning technique, game theory can be developed.

The result of the literature review shows there are many policy tools available to governments to promote BB adoption. The econometric methodology is also available to analyze the effectiveness of public policies about the supply-side and demand-side. Thus, TELCO and policymakers may develop a game theory with a 2-prong approach:

- Work across government agencies to set goals to improve the features of the rural areas, especially on those features with a high correlation efficient to the growth of economic or BB diffusion.
- Use the econometric methodology to measure the effect of public policies on BB adoption.

5.3 RECOMMENDATION FOR FUTURE RESEARCH

In extension to this thesis, future research is recommended in three perspectives:

- Research with local data
- Research with different SVM kernels
- Research with different SVM models: multi-classification and regression

Data Collection: In this research, the global country data secured from the World Bank databank is used as prototyping data whereas the Malaysian state data obtained from the Department of Statistics Malaysia is used as real-life data. It will be better if the prototyping data and real-life data are in proximity to the geographic areas targeted. There are approximately 150 districts and 24,000 villages in Malaysia. It will be ideal to have real-life PESTEL data collected from some of these districts and villages for future research. The data collection shall focus on those features with high correlation impact against socioeconomic response. Those features with low impact or coefficient could probably be eliminated by using Principal Component Analysis (PCA). Principal Component Analysis (PCA) is a multivariate technique that transforms some related variables into a smaller set of uncorrelated variables[162]. For example, four geographic areas might all be labeled as having socioeconomic potential, all with 19 features. But after PCA, probably one or more of the areas might be put in different principal components and treated as outliers. Nevertheless, PCA should not be used to prevent overfitting.

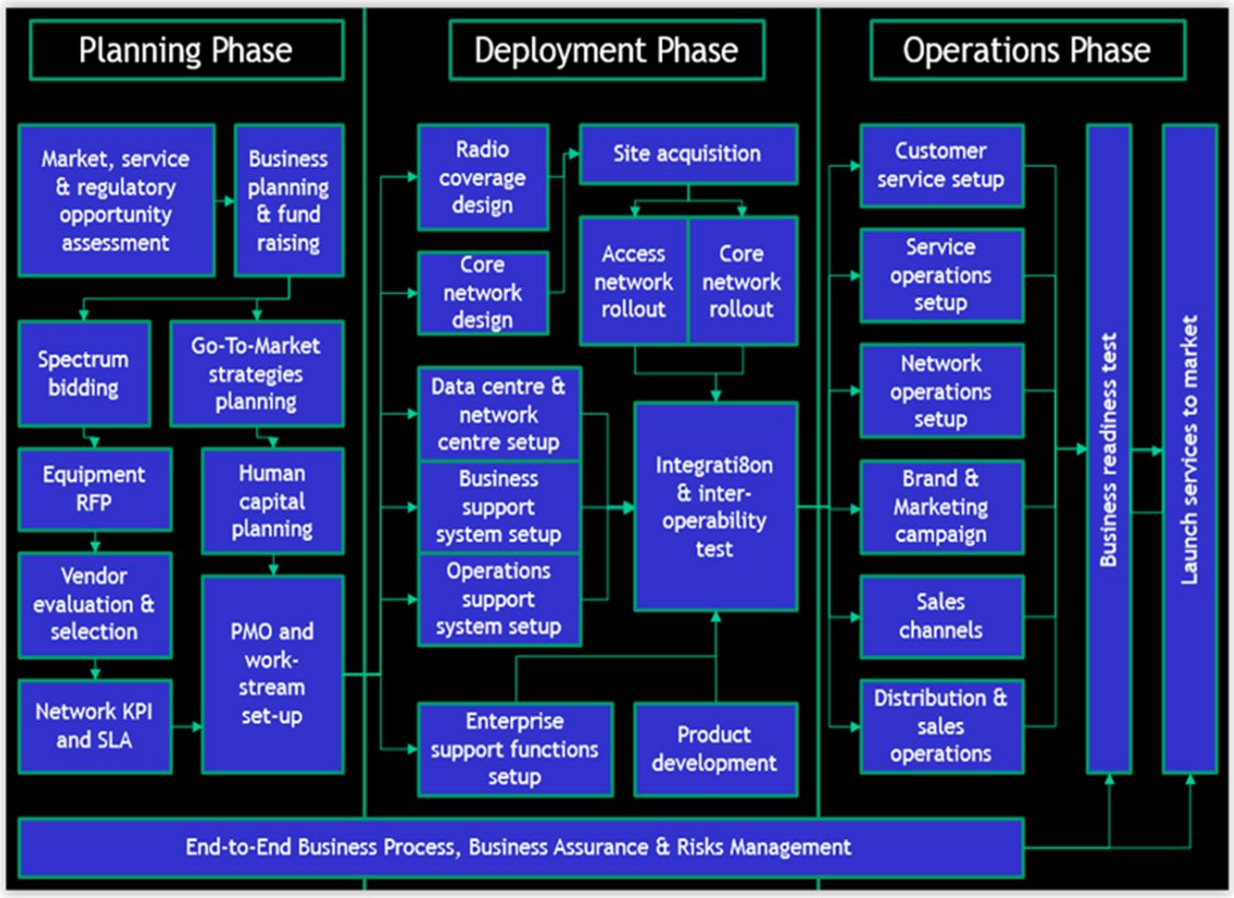
Machine Learning: Three common SVM kernels have been used in this research as a machine learning model: linear, polynomial and RBF. Other kernels that could be used in future research are such as Sigmoid, Fourier, B-spline, Cosine, Multiquadric, Wave, Log, Cauchy, T-Student[163][164] and so on. An extensive list of kernels is available on Biochemistry & Pharmacology: Open Access[165]. Each kernel could also be modified for further research as long as the modified kernel meets Mercer's Theorem. For example, a Gaussian RBF kernel [$k(x_i, x_j) = \exp\left(-\frac{1}{2\sigma^2}\left|\left|x_i - x_j\right|\right|^2\right)$] when modified [e.g.

$k(x_i,,,x_j) = \exp\left(-\frac{1}{2\sigma^2}\left||x_i - x_j|\right|^{2d}\right)$], could possibly improve the accuracy of machine learning and generalization. Furthermore, a hybrid kernel (refer to Appendix 12) of two or more SVM kernels would still meet Mercer's Theorem requirement.

Machine Modelling: A Support Vector Machine is a supervised machine learning algorithm which can be used for both classification or regression challenges. In this thesis, SVM is used for classification. This thesis has proven that the curve-fitting technique used in the machine learning framework is a reliable method for correlating geographical features with a targeted outcome. Future research may expand the study by using other kinds of geographical features against a target response.

[END]

APPENDIX 1 – TELCO OPERATING MODEL



Source: the author

APPENDIX 2 – EMPIRICAL EQUATION COMPUTED BY CURVE-FITTING SOFTWARE (QUADRATIC MODEL WITH WORLD BANK DATA)

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 – EMPIRICAL EQUATION COMPUTED BY CURVE-FITTING SOFTWARE (QUADRATIC MODEL IN MATLAB FORMAT WITH WORLD BANK DATA)

```
R =
- 53003.12 + 19482.23*X(1) + 1208.34*X(2) + 45242.04*X(3) - 55108.44*X(4) + 12152.05*X(5) + 3949.75*X(6)
+ 39626.27*X(7) - 13441.39*X(8) + 3126.98*X(9) + 1750.12*X(10) - 3815.07*X(11) + 74879.88*X(12)
- 38351.13*X(13) + 7755.48*X(14) + 262.06*X(15) + 2777.10*X(16) - 15230.85*X(17) + 1393.57*X(18)
- 1.060E+005*X(19) - 99.46*X(1)*X(2) - 513.78*X(1)*X(3) - 98315.78*X(1)*X(4) + 1240.30*X(1)*X(6)
+ 86901.40*X(1)*X(7) + 37148.31*X(1)*X(8) - 890.18*X(1)*X(9) - 604.35*X(1)*X(10) + 1689.63*X(1)*X(11)
+ 1.020E+005*X(1)*X(12) + 975.96*X(1)*X(13) + 850.02*X(1)*X(14) - 417.25*X(1)*X(15) - 367.68*X(1)*X(16)
- 282.38*X(1)*X(17) + 8.51*X(1)*X(18) - 1.105E+005*X(1)*X(19) + 174.22*X(2)*X(3) + 75682.74*X(2)*X(4)
+ 22975.17*X(2)*X(5) - 1413.09*X(2)*X(6) - 63971.80*X(2)*X(7) - 34206.30*X(2)*X(8) +1233.66*X(2)*X(9)
+ 349.68*X(2)*X(10) - 143.14*X(2)*X(11) - 1.436E+005*X(2)*X(12) - 275.10*X(2)*X(13) -4360.30*X(2)*X(14)
+ 253.71*X(2)*X(15) + 171.00*X(2)*X(16) - 573.14*X(2) *X(17) - 786.52*X(2)*X(18) + 1.485E+005*X(2)*X(19)
- 5153.16*X(3)*X(4) - 776.99*X(3)*X(5) + 57.80*X(3)*X(6) + 4920.95*X(3)*X(7) - 75.27*X(3)*X(8)
- 123.31*X(3)*X(9) + 34.59*X(3)*X(10) + 284.86*X(3)*X(11) - 2.324E+005*X(3)*X(12) - 725.62*X(3)*X(13)
- 1558.03*X(3)*X(14) + 9.03*X(3)*X(15) - 41.31*X(3) *X(16) + 242.00*X(3)*X(17) - 44.23*X(3)*X(18)
+ 2.810E+005*X(3)*X(19) + 35180.49*X(4)*X(5) + 717.81*X(4)*X(6) - 1529.94*X(4) *X(7) + 16947.67*X(4)*X(8)
- 21093.04*X(4)*X(9) + 9622.49*X(4)*X(10) + 9422.54*X(4)*X(11) + 7.653E+005*X(4)*X(12)
- 35976.62*X(4)*X(13) - 4818.83*X(4)*X(14) + 651.88*X(4)*X(15) - 139.97*X(4)*X(16) + 45906.81*X(4)*X(17)
+ 7668.78*X(4)*X(18) - 8.509E+005*X(4)*X(19) + 314.88*X(5)*X(6) - 13259.68*X(5)*X(7) - 25836.41*X(5)*X(8)
- 1195.76*X(5)*X(9) + 331.20*X(5)*X(10) + 623.94*X(5)*X(11) + 33117.50*X(5)*X(12) + 414.35*X(5)*X(13)
- 908.40*X(5)*X(14) - 91.08*X(5)*X(15) + 241.31*X(5)*X(16) + 1846.76*X(5)*X(17) + 736.70*X(5)*X(18)
- 41681.14*X(5)*X(19) - 1673.17*X(6)*X(7) + 1690.65*X(6)*X(8) + 1.55*X(6)*X(9) - 88.10*X(6)*X(10)
- 21.11*X(6)*X(11) - 599.39*X(6)*X(12) - 288.33*X(6)*X(13) + 1787.45*X(6)*X(14) - 18.27*X(6)*X(15)
- 15.55*X(6)*X(16) + 144.20*X(6)*X(17) - 34.46*X(6) *X(18) + 2204.01*X(6) *X(19) - 22327.62*X(7)*X(8)
+ 11790.30*X(7)*X(9) - 6366.37*X(7)*X(10) - 13596.52*X(7)*X(11) - 5.010E+005*X(7)*X(12)
+ 17029.84*X(7)*X(13) + 17720.81*X(7)*X(14) - 1346.58*X(7)*X(15) - 1142.60*X(7)*X(16) - 52342.23*X(7)*X(17)
- 5466.25*X(7)*X(18) + 5.805E+005*X(7)*X(19) + 8598.65*X(8)*X(9) - 2690.13*X(8)*X(10) + 1806.51*X(8)*X(11)
- 3.975E+005*X(8)*X(12) - 19138.76*X(8)*X(13) + 1483.43*X(8)*X(14) + 2265.16*X(8)*X(15)
+ 1397.58*X(8)*X(16) - 4327.59*X(8)*X(17) - 1994.54*X(8)*X(18) + 4.241E+005*X(8)*X(19) - 62.50*X(9)*X(10)
- 6.44*X(9)*X(11) + 1557.38*X(9)*X(12) -121.85*X(9)*X(13) + 1731.15*X(9)*X(14) - 13.88*X(9)*X(15) -
17.04*X(9)*X(16) - 45.81*X(9)*X(17) - 19.30  *X(9) *X(18) + 1441.49*X(9)*X(19) + 20.49*X(10)*X(11)
- 13562.16*X(10)*X(12) - 44.02*X(10)*X(13) - 599.96*X(10)*X(14) - 11.91*X(10)*X(15) - 4.97*X(10)*X(16)
+ 203.80*X(10)*X(17) + 9.57*X(10)*X(18) + 15279.57*X(10)*X(19) + 23107.61*X(11)*X(12) + 0.35*X(11)*X(13)
+ 331.33*X(11)*X(14) + 13.65*X(11)*X(15)  - 20.20*X(11)*X(16) - 156.75*X(11)*X(17) - 18.07*X(11)*X(18)
- 26613.10*X(11)*X(19) - 2482.64*X(12)*X(13) - 5666.71*X(12)*X(14) - 1041.78*X(12)*X(15)
+ 2698.99*X(12)*X(16) - 5696.09*X(12)*X(17) + 1273.24*X(12)*X(18) + 6044.75*X(12) *X(19)
+ 239.31*X(13) *X(14) + 17.85*X(13)*X(15) - 73.18*X(13)*X(16) + 414.67*X(13)*X(17)
- 43.23*X(13) *X(18)
```

APPENDIX 4 – EMPIRICAL EQUATION COMPUTED BY CURVE-FITTING SOFTWARE (QUADRATIC MODEL WITH MALAYSIAN STATE DATA)

$$
\begin{aligned}
R = {} & \\
& + 27.06 + 1.21*A + 11.93*B - 279.72*C + 22.11*D - 1.92*E + 313.27*F + 1.60*G - 13.37*H \\
& + 3.34*J + 15.15*K - 0.20*A*B - 9.86*A*C - 0.22*A*D - 0.11*A*E + 3.76*A*F + 7.92*A*G \\
& + 0.39*A*H - 0.28*A*J - 0.55*A*K + 2.21*B*C + 1.74*B*D - 0.016*B*E + 0.95*B*F \\
& + 2.13*B*G - 2.28*B*H - 0.080*B*J + 7.50*B*K - 164.77*C*D + 9.90*C*E - 382.10*C*F \\
& + 5.99*C*G + 67.06*C*H - 18.70*C*J - 73.24*C*K + 0.96*D*E + 179.28*D*F + 9.03*D*G \\
& + 0.84*D*H - 0.38*D*J - 3.16*D*K - 8.56*E*F - 4.27*E*G + 0.49*E*H - 0.51*E*J + 0.13*E*K \\
& - 0.19*F*G - 72.66*F*H + 18.47*F*J + 106.09*F*K - 11.90*G*H + 4.83*G*J - 9.39*G*K \\
& + 0.26*H*J + 0.35*H*K - 0.69*J*K - 0.22*A^2 - 0.058*B^2 + 143.59*C^2 + 1.34*D^2 - 0.17*E^2 \\
& + 233.51*F^2 + 1.74*G^2 + 0.28*H^2 - 0.24*J^2 - 0.35*K^2
\end{aligned}
$$

where,

A = % of agricultural land

B = % of population with electricity access

C = Population size

D = Populations density per sq. km

E = Birth-rate (per 1k people)

F = Labor force

G = Number of secondary school students enrolled

H = Fixed broadband per 100 people

J = Wireless broadband per 100 people

K = GDP per capita

APPENDIX 5 – MATRIX OF FEATURES AVAILABLE FOR MACHINE LEARNING

No	Features	Feature Label for WB data	WB data Correlation Coefficient to Response	World Bank Database	Malaysia State Database	Malaysia District Database
1	Land Size (km2)	A	0.105	yes	yes (% of agricultural land)	yes
2	Agricultural Land (km2)	B	0.083	yes		yes
3	% of the population with electricity access	C	0.271	yes	yes	yes
4	Population size	D	-0.041	yes	yes	yes
5	Populations density per sq. km	E	0.197	yes	yes	
6	Birth-rate (per 1k people)	F	-0.532	yes	yes	
7	Labor force	G	-0.028	yes	yes	yes
8	Number of secondary school students enrolled	H	-0.041	yes	yes	yes
9	Fixed BB per 100 people	J	0.746	yes	yes	
10	Wireless BB per 100 people	K	0.712	yes	yes	
11	Telephone lines per 100 people	L	0.665	yes		
12	GDP	M	0.275	yes		
13	GDP per capita	N	0.981	yes	yes	
14	Economic Activity – tourism	O	0.373	yes		
15	Average monthly rainfall	P	-0.015	yes		
16	Average daily temperature	Q	-0.122	yes		
17	Length of tarred road (km/million person)	R	0.454	yes		yes
18	Life Expectancy	S	0.616	yes		
19	GNI	T	0.273	yes		
R	GNI per capita	RESPONSE		yes		
Total number of features				19	10	7

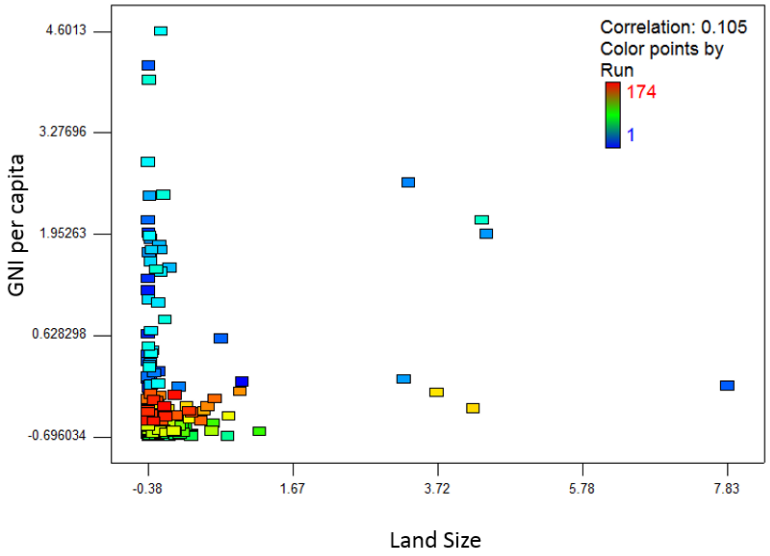
APPENDIX 6 – CORRELATION COEFFICIENT OF 19 GEOGRAPHICAL FEATURES TO GNI PER CAPITA

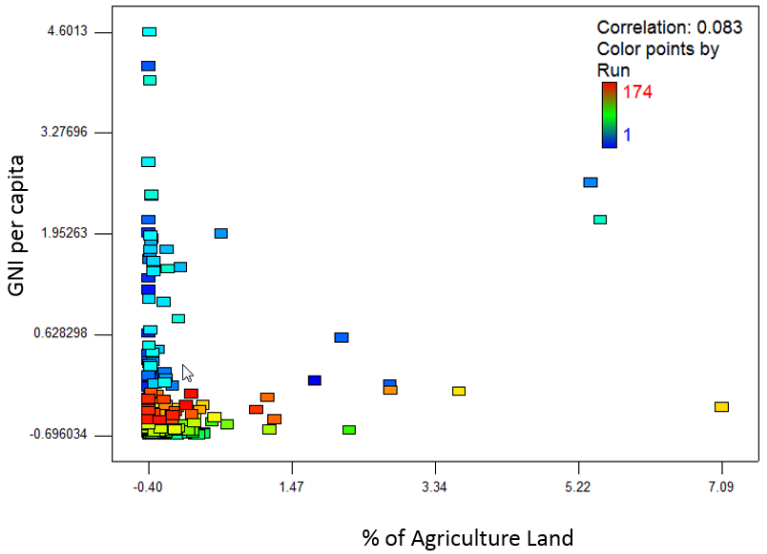
Features		Land Size	Agro Land	Electric Access	Pop Size	Pops Dens	Birth Rate	Labor	Second School	Fixed BB	Wireless BB	Phone	GDP	GDP /cap	Tourism	Rain	Temp	Road	Life	GNI	GNI/cap
	Label	A	B	C	D	E	F	G	H	J	K	L	M	N	O	P	Q	R	S	T	Respond
Land size ($km2$)	A		0.772	0.068	0.433	-0.084	-0.086	0.459	0.415	0.089	0.094	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 the 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
Labor 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 BB 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 BB 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	M													0.25	0.877	0.022	0.045	0.123	0.214	0.999	0.275
GDP per capita	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 the tarred road (km per million)	R																		0.359	0.122	0.454
Life expectancy	S																			0.21	0.616
GNI	T																				0.273

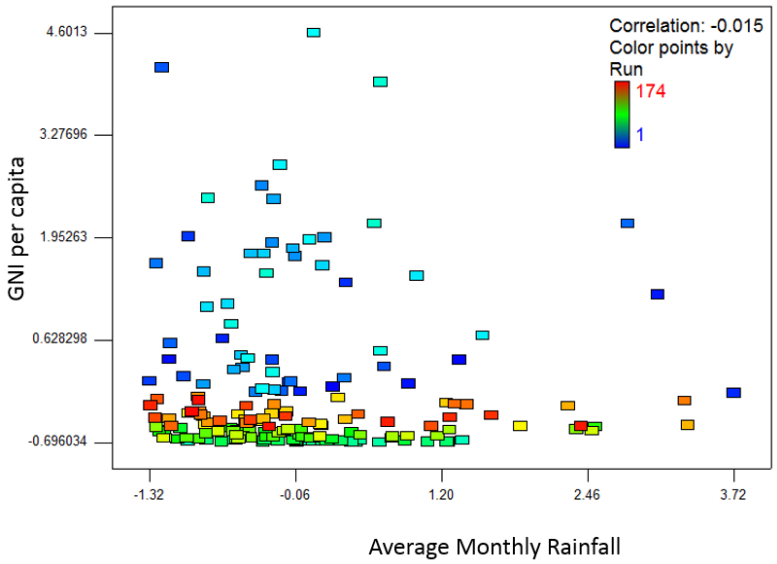
APPENDIX 7 – CORRELATION COEFFICIENT OF 10 GEOGRAPHICAL FEATURES TO GNI PER CAPITA

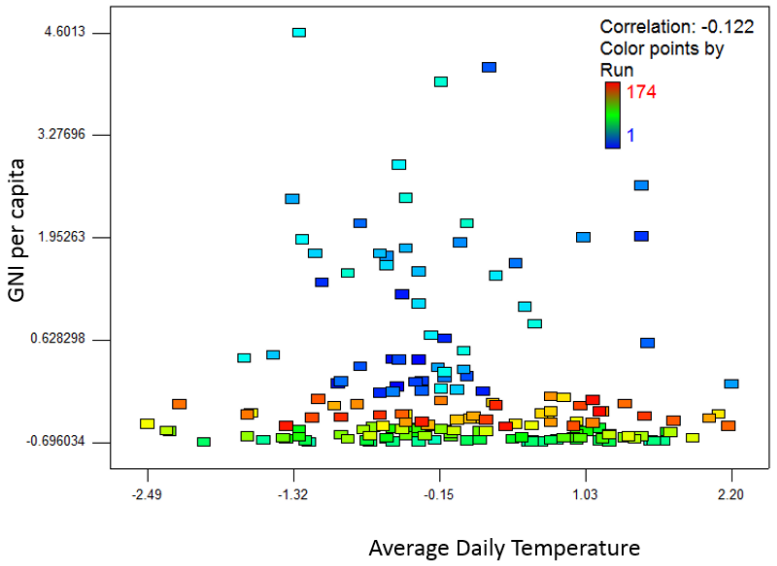
Geographical Features	% Agro-Land	Electric Access	Pop Size	Pop Dense	Birth Rate	Labor Force	Second School	Fixed BB	Wireless BB	GDP per capita
% Agro-Land										
Electric Access	-0.233									
Pop Size	0.095	0.077								
Pop Dense	-0.154	0.068	-0.006							
Birth Rate	0.163	-0.573	-0.061	-0.173						
Labor Force	0.085	0.077	0.981	-0.007	-0.082					
Secondary School	0.093	0.086	0.975	-0.006	-0.071	0.923				
Fixed BB	-0.097	0.347	-0.016	0.206	-0.741	0.004	-0.022			
Wireless BB	-0.122	0.374	-0.023	0.242	-0.748	-0.004	-0.026	0.881		
GDP per capita	-0.22	0.266	-0.048	0.192	-0.524	-0.035	-0.047	0.738	0.664	
GNI per capital	-0.236	0.271	-0.041	0.197	-0.532	-0.028	-0.041	0.746	0.665	0.981

APPENDIX 8 – CORRELATION COEFFICIENT OF 19 SELECTED FEATURES TO GNI PER CAPITA (1/5)

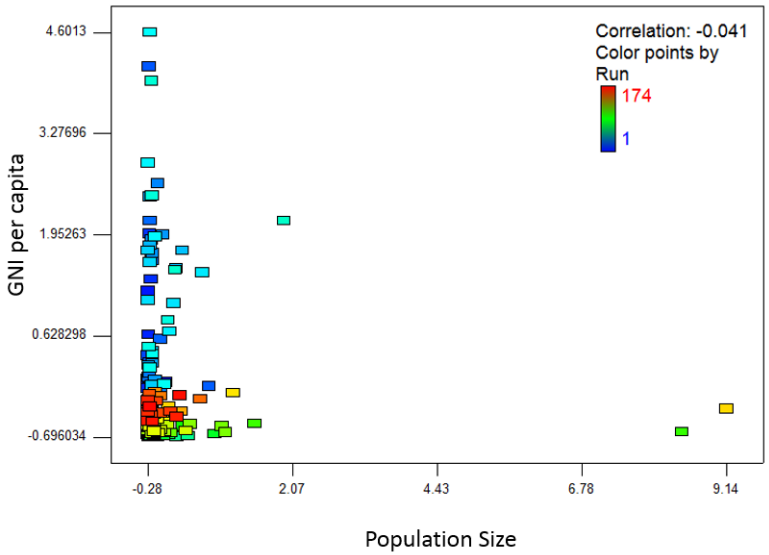


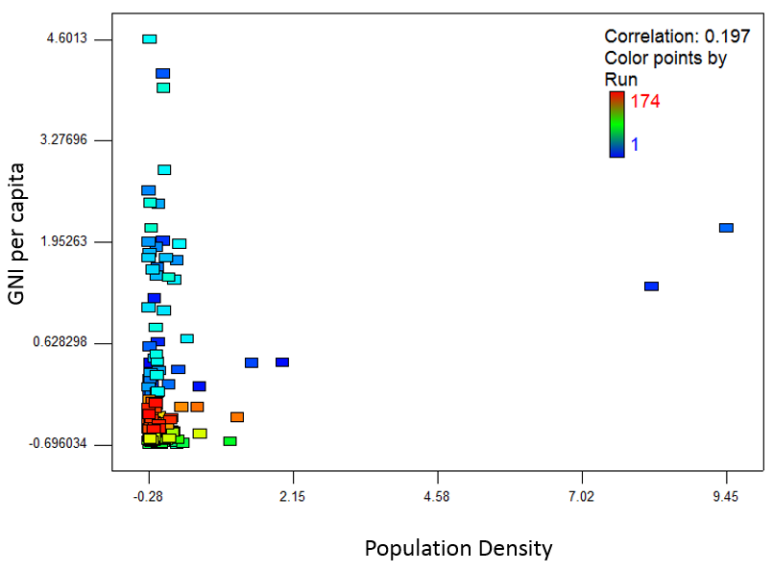


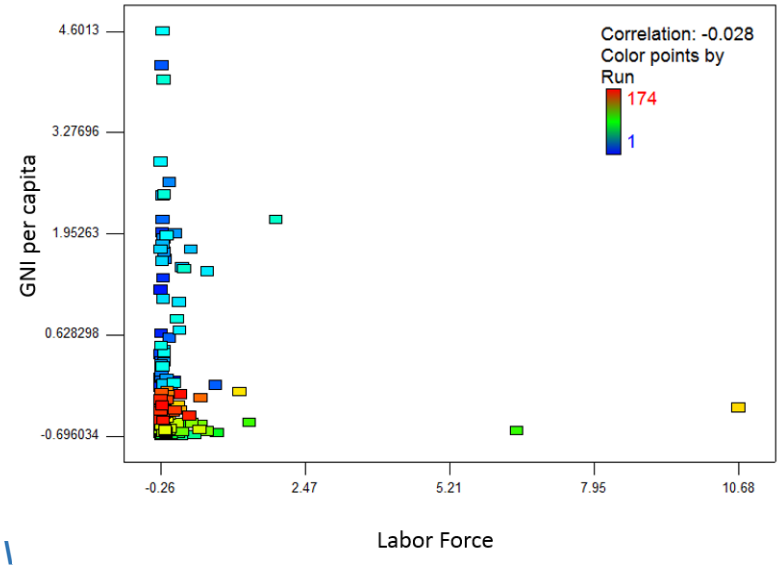


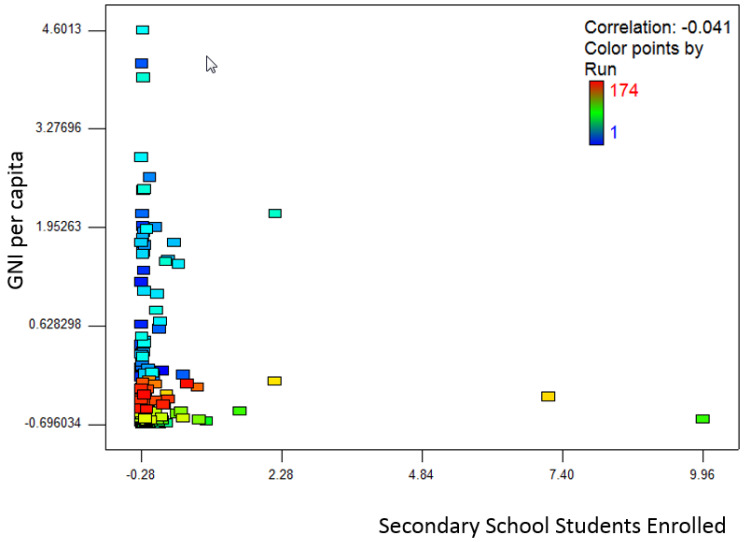


APPENDIX 8 – CORRELATION COEFFICIENT OF 19 SELECTED FEATURES TO GNI PER CAPITA (2/5)

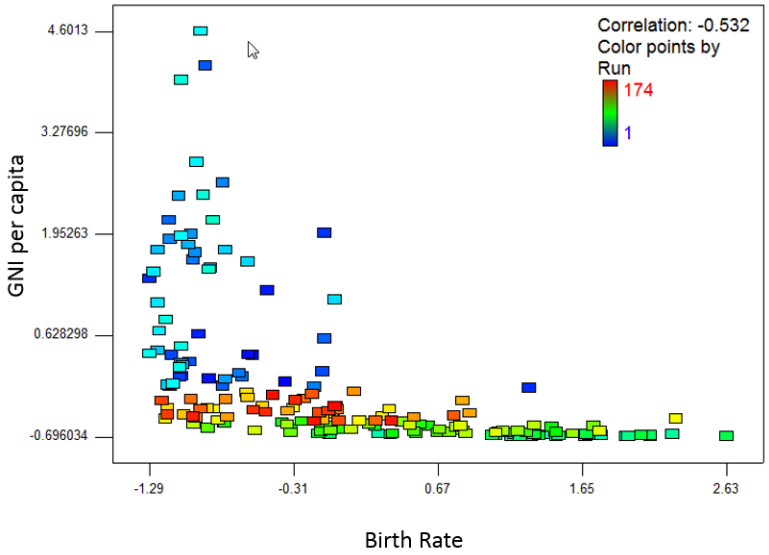


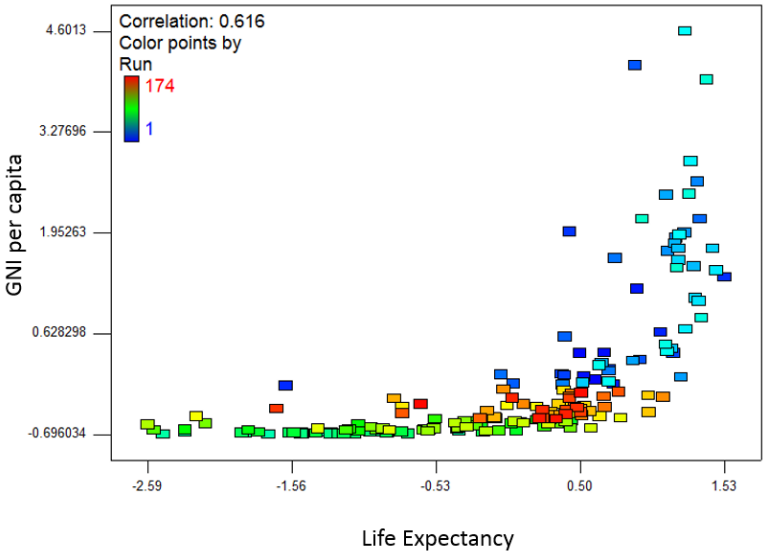


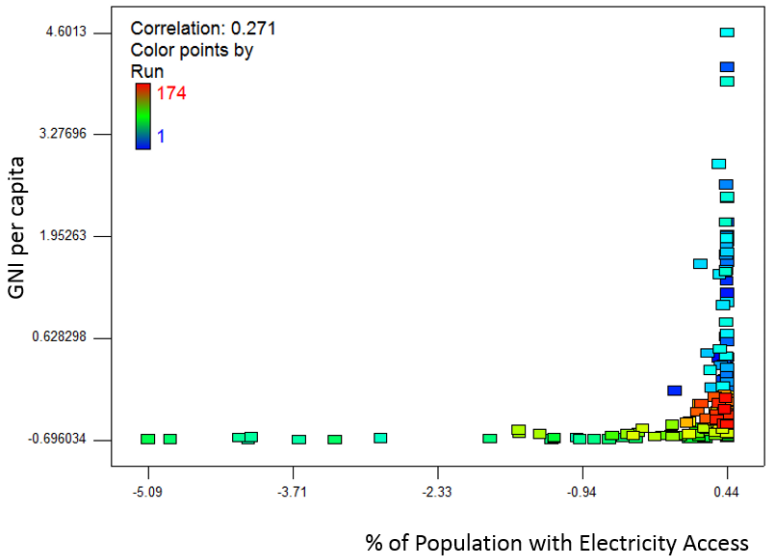


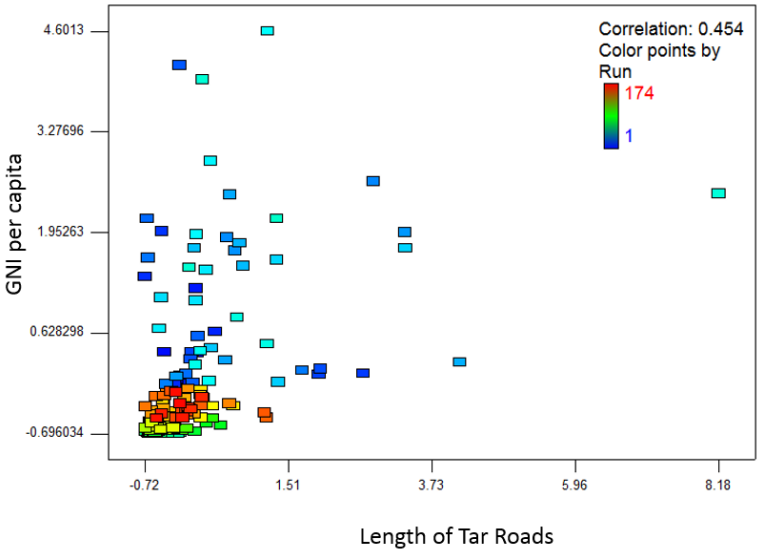


APPENDIX 8 – CORRELATION COEFFICIENT OF 19 SELECTED FEATURES TO GNI PER CAPITA (3/5)

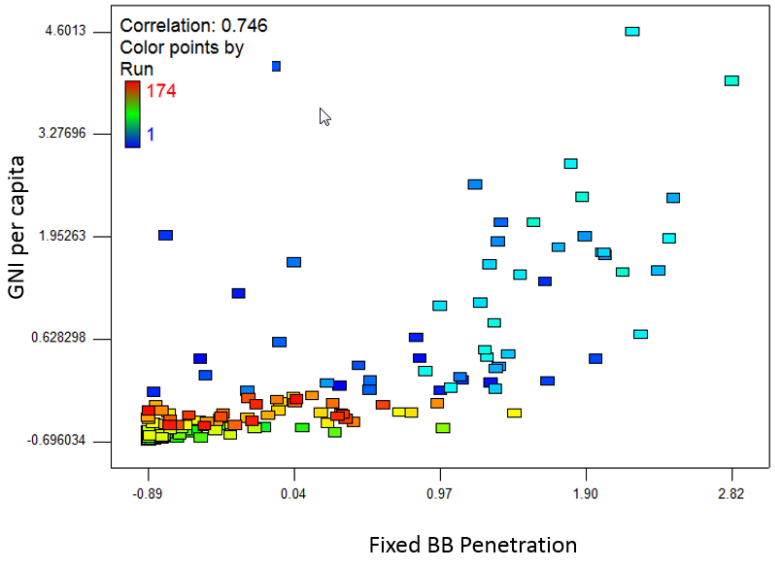


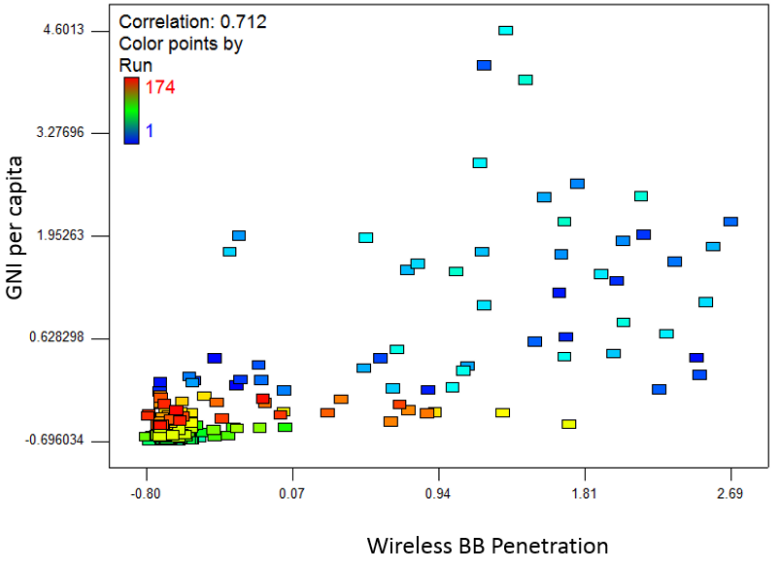


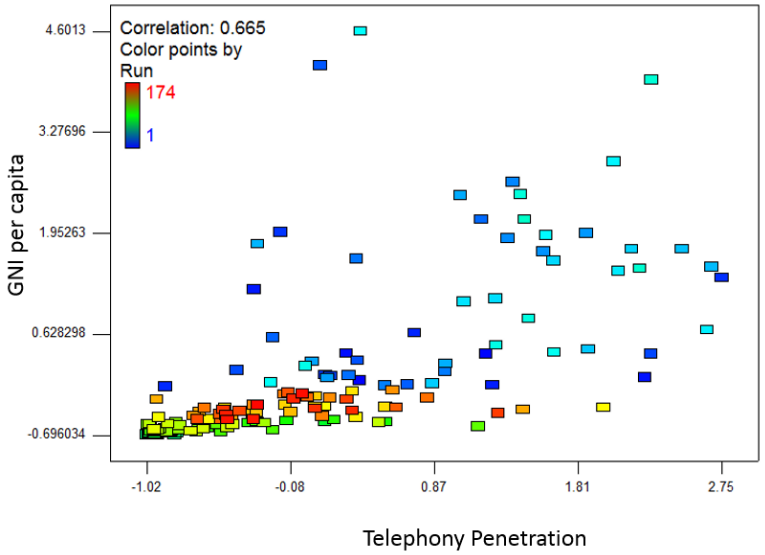


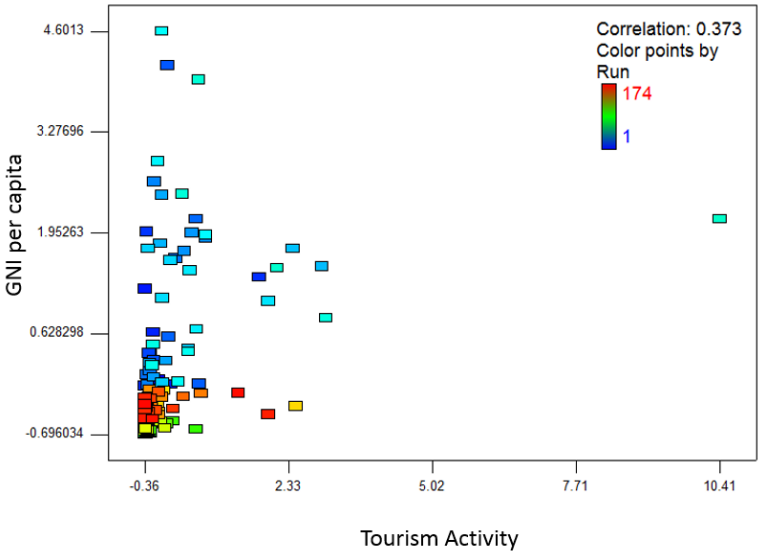


APPENDIX 8 – CORRELATION COEFFICIENT OF 19 SELECTED FEATURES TO GNI PER CAPITA (4/5)

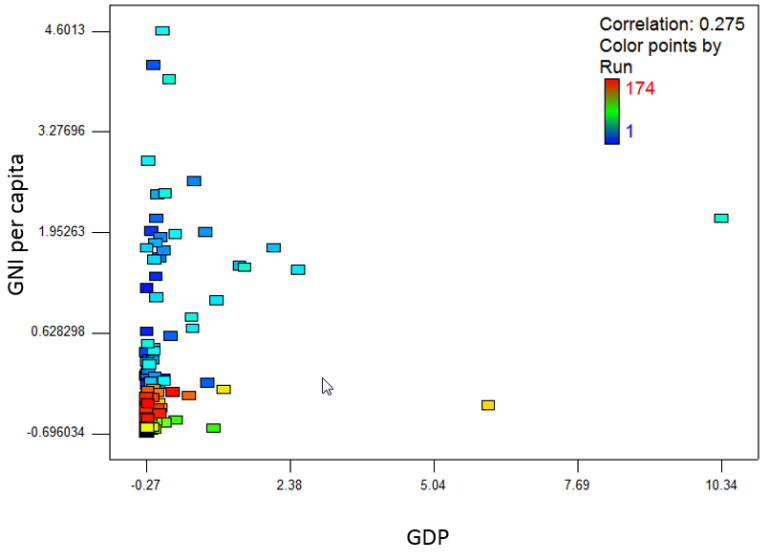


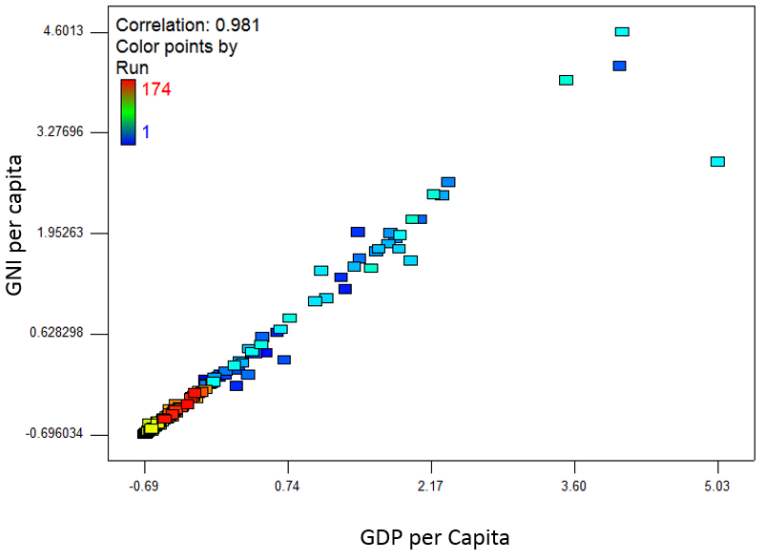


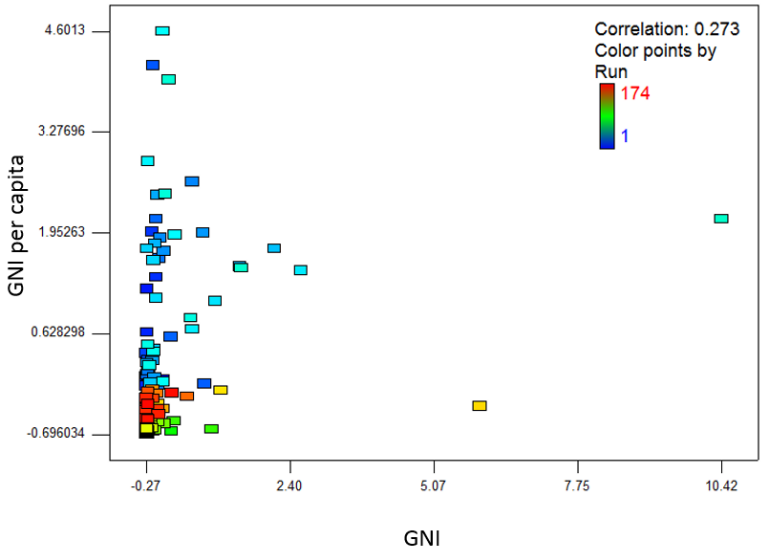




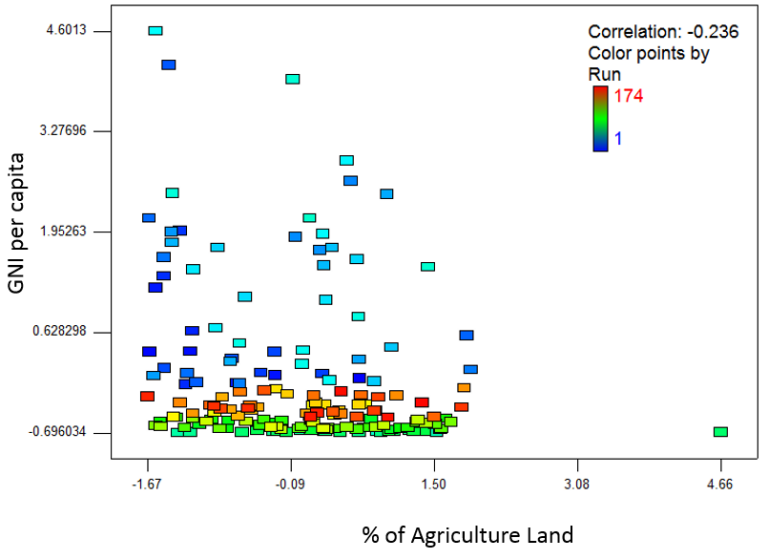
APPENDIX 8 – CORRELATION COEFFICIENT OF 19 SELECTED FEATURES TO GNI PER CAPITA (5/5)

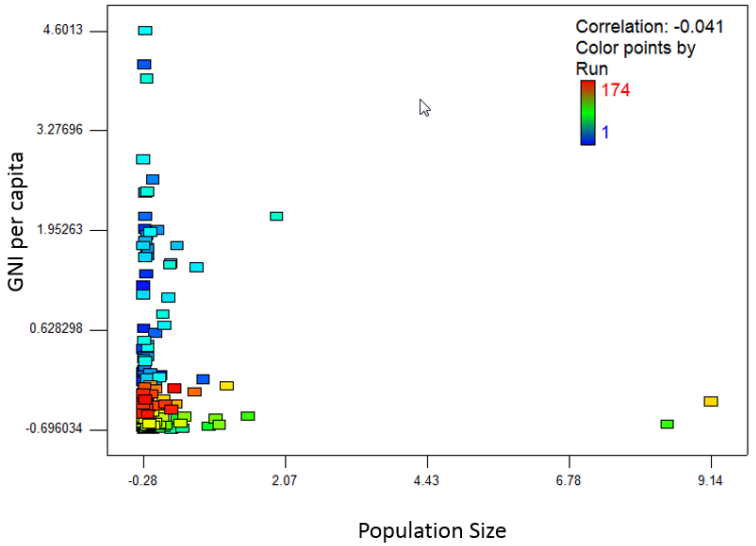


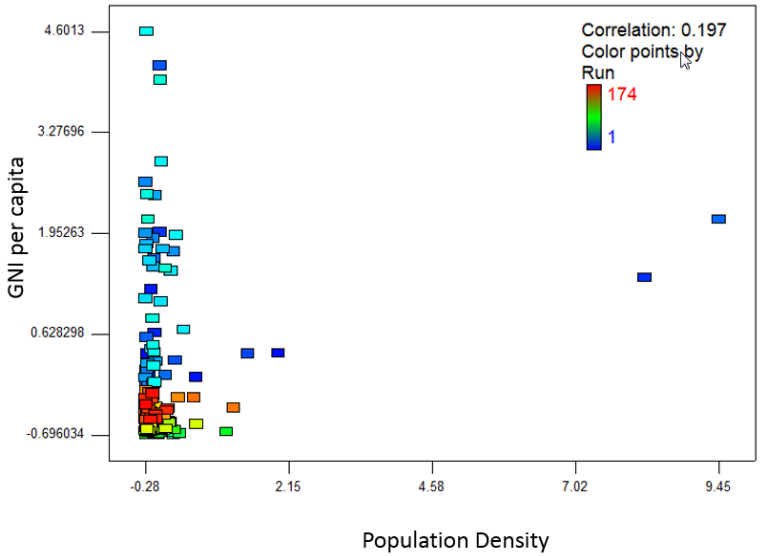


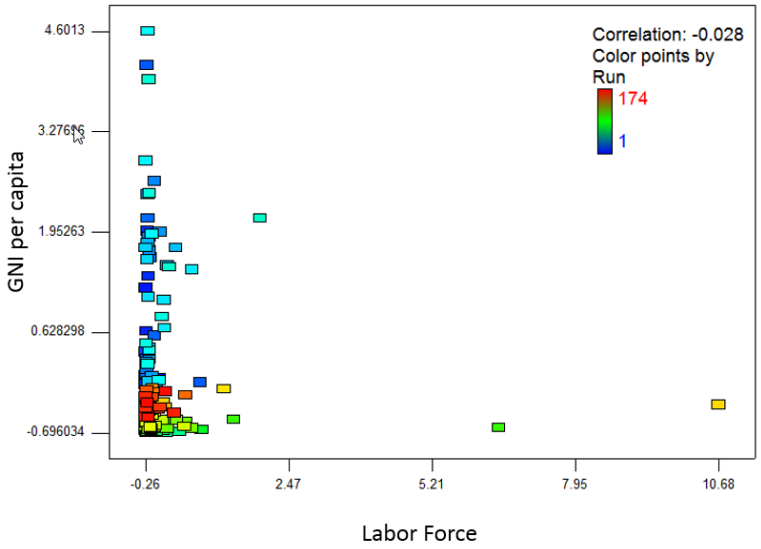


APPENDIX 9 – CORRELATION COEFFICIENT OF 10 SELECTED FEATURES TO GNI PER CAPITA (1/3)

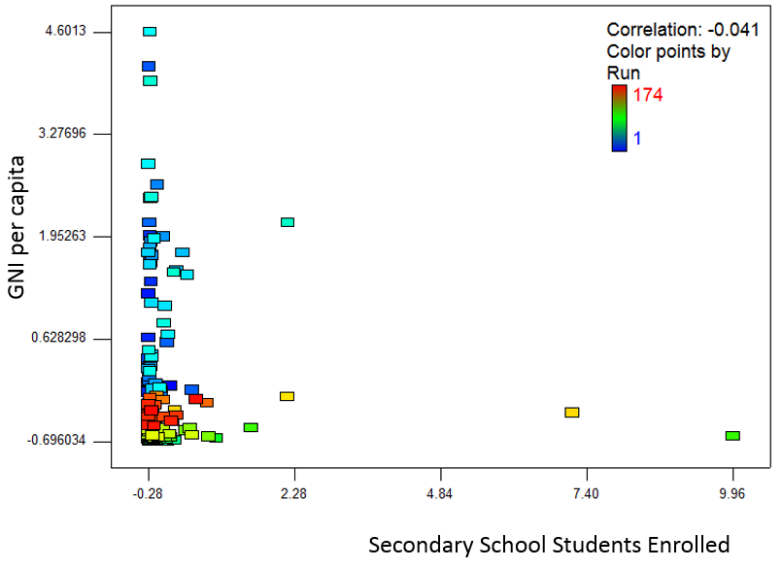


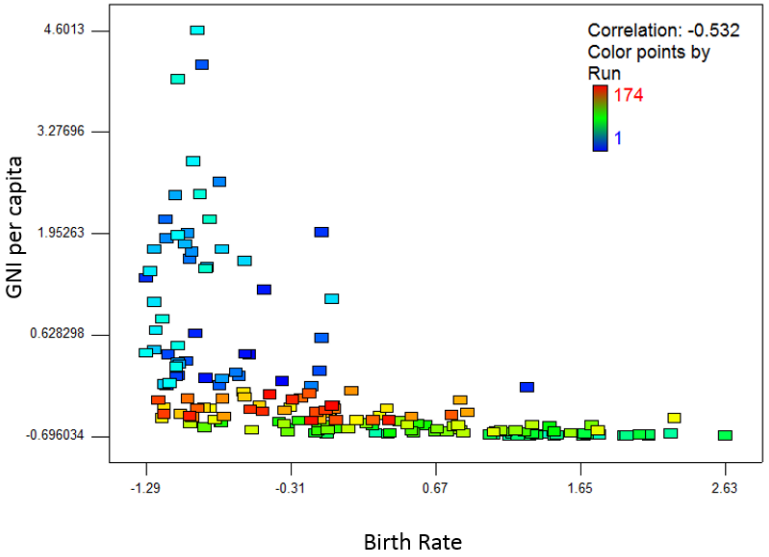


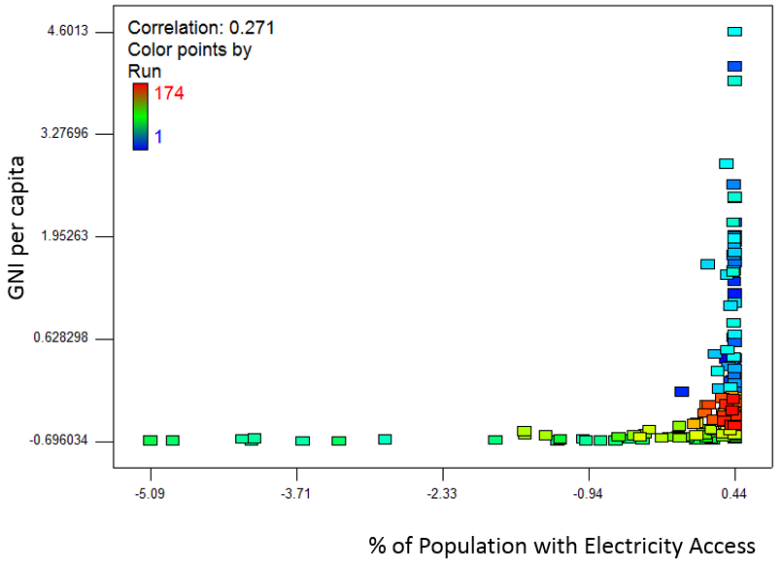


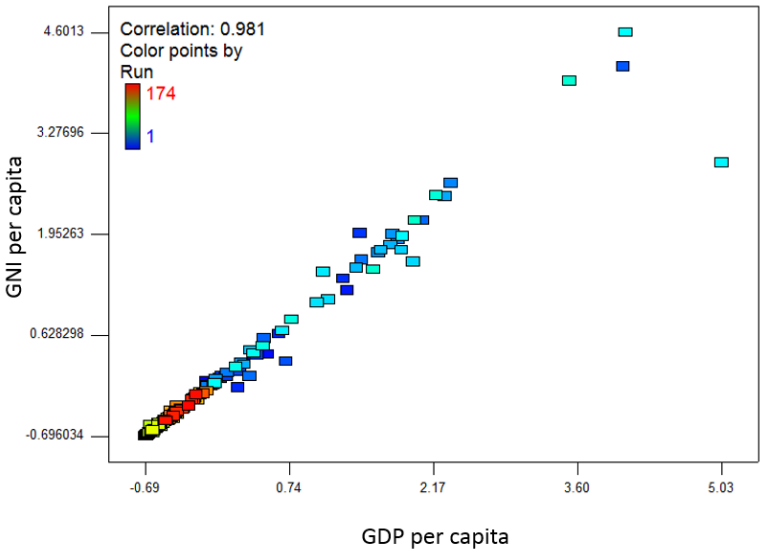


APPENDIX 9 – CORRELATION COEFFICIENT OF 10 SELECTED FEATURES TO GNI PER CAPITA (2/3)

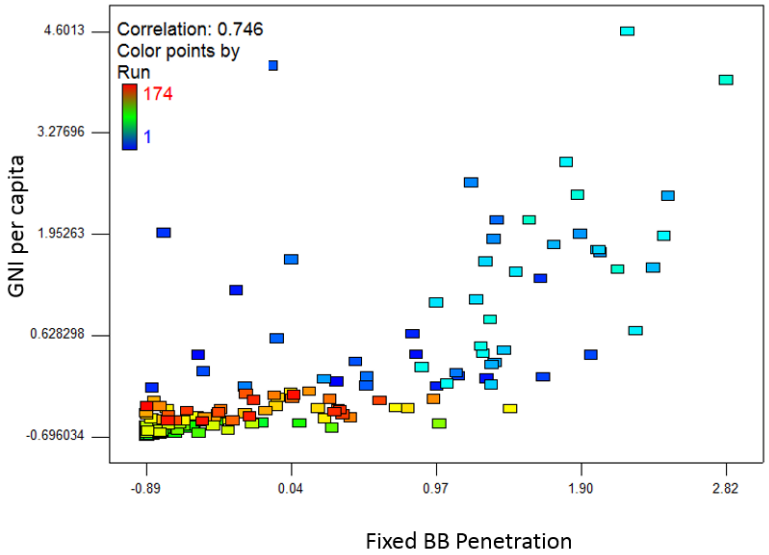


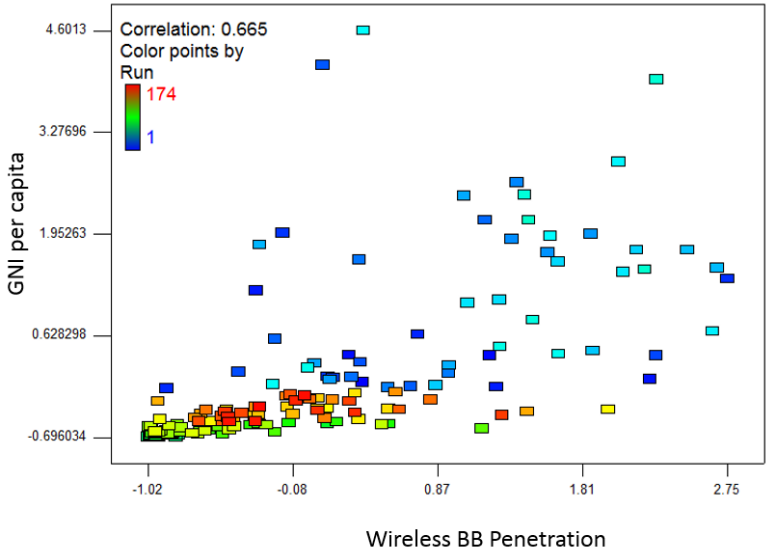






APPENDIX 9 – CORRELATION COEFFICIENT OF 10 SELECTED FEATURES TO GNI PER CAPITA (3/3)





APPENDIX 10 – SVM TRAINING AND TESTING RESULTS

- **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%

APPENDIX 11 – TEMPLATE FOR ROI SIMULATION

Months In A Year = 12		Yr 1	Yr 2	Yr 3	Yr 4	Yr 5	Yr 6	Yr 7	Yr 8	Yr 9	Yr 10
CapEx (Electronics + Site Infrastructure)											
Electronics (radio, backhaul, switches)											
Site Infrastructure (civil, mechanical, electrical, site acquisition, logistics, permitting, project management, etc.)											
Depreciation/Ammortization Cost											
OpEx (site rental, preventive maintenance, utilities, sales commission, device subsidy, marketing cost, logistics cost, etc.)											
Subscribers											
ARPU											
Revenue From Mobile Broadband											
Revenue From Fixed Broadband											
EBITDA											
Profits Before Taxes											
Taxes											
PAT											
Cashflow	**Yr 0**	**Yr 1**	**Yr 2**	**Yr 3**	**Yr 4**	**Yr 5**	**Yr 6**	**Yr 7**	**Yr 8**	**Yr 9**	**Yr 10**
Operating Cash		-	-	-	-	-	-	-	-	-	-
Capex											
Taxes											
FCFF											
Cummulative FCFF											

WACC	IRR	NPV	Peak Funding	Payback Period

APPENDIX 12 – EXAMPLES OF HYBRID KERNELS

Below are some examples of hybrid kernels:

- Linear + RBF:

$$k = (x_i \cdot x_j) + \exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right)$$

- RBF + polynomial:

$$k = exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right) + (x_i \cdot x_j)^d$$

- RBF + modified RBF + constant:

$$k = \exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right) + \left(\exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right)\right)^2 + C$$

- Polynomial + Linear + RBF:

$$k = \alpha(x_i \cdot x_j)^d + \beta(x_i \cdot x_j) + \exp\left(-\gamma \left|\left|x_i - x_j\right|\right|^2\right)$$

REFERENCE

[1] ITU, “Broadband: A Platform for Progress, a report by the Broadband Commission for Digital Development,” 2011, p. International Telecommunications Union.

[2] Malaysian Economic Planning Unit, “11th Malaysia Plan,” Prime Minister’s Department, 2015.

[3] L. Roller and L. Waverman, “Telecommunications infrastructure and economic development: A simultaneous approach,” in *American Economic Review (AER),* no. 91, 2001, pp. 909–923.

[4] Kuppusamy, Raman, and Lee, “Whose ICT Investment Matters to Economic Growth: Private or Public? The Malaysian Perspective,” *Electron. J. Inf. Syst. Dev. Ctries. (EJIDC),* pp. 37, 7, 1–19, 2009.

[5] A. Shiu and P. L. Lam, “Causal Relationship between Telecommunications and Economic Growth: A Study of 105 Countries,” *17th Biennial Conference of the International Telecommunications Society (ITS)*. Montreal, pp. 24–27, 2008.

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

[7] N. Czernich, O. Falck, T. Kretschmer, and 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,” *The Global Information Technology Report*. World Economic Forum, 2013.

[9] J. Bergendahl, “Broadband changes society,” 2010. [Online]. Available: www.ericsson.com/news/101027_broadband_bergendahl_244218599_c.

[10] R. L. Katz and P. Koutroumpis, “Measuring Socio-Economic Digitization: A Paradigm Shift,” 2012.

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

[12] R. L. Katz and T. A. Berry, *Driving Demand for Broadband Networks and Services*. Springer, 2014.

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

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

[15] M. Kuppusamy and B. Shanmugam, “Information-Communication Technology and Economic Growth in Malaysia, International Association for Islamic Economics,” *Rev. Islam. Econ.*, vol. 11, no. 2, pp. 87–100, 2007.

[16] L. Waverman, M. Meschi, and M. Fuss, “The impact of telecoms on economic growth in developing countries,” in *Africa: The impact of mobile phones, The Vodafone Policy Paper Series*, no. 2, 2005, pp. 10–23.

[17] C. Chakraborty and B. Nandi, “Privatization, telecommunications and growth in selected Asian countries: An econometric analysis,” *Commun. Strateg.*, vol. 52, no. 4, pp. 31–47, 2003.

[18] P. Stenberg and M. 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*, Information

Science Reference, USA, 2011, p. 157.

[19] J. Karner and R. Onyeji, “Telecom private investment and economic growth: The case of African and Central & East European countries,” Jönköping University, 2007.

[20] C. Kongaut, I. K. Rohman, and E. 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.

[21] K. Sabbagh, B. El-Darwiche, R. Friedrich, and M. Singh, “Maximizing the impact of digitization,” Strategy&, PwC, 2012.

[22] M. Kirlidog and S. Little, “Regional – National ICT Strategies,” in *Sherif Kamel: E-Strategies for Technological Diffusion and Adoption: National ICT Approaches for Socioeconomic Development*, Information Science Reference, USA, 2010, p. 66.

[23] UNCTAD Secretariat, “Background Paper for Expert Meeting in Support of the Implementation and Follow-up of WSIS: USING ICTs TO ACHIEVE GROWTH AND DEVELOPMENT,” in *United Nations Conference on Trade and Development*, 2006.

[24] P. Curwen and J. Whalley, *Telecommunications in a High Speed World – Industry structure, Strategic Behavior and Socio-Economic Impact*. UK Gower Publishing, 2010.

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

[26] F. J. Cronin, E. B. Parker, E. K. Colleran, and M. A. Gold, “Telecommunications

infrastructure investment and economic development," *Telecomm. Policy*, vol. 17, pp. 415–430, 1993.

[27] F. J. Cronin, E. K. Colleran, P. L. Herbert, and S. Lewitzky, "Telecommunications and growth: The contribution of telecommunications infrastructure investment to aggregate and sectoral productivity," *Telecomm. Policy*, vol. 17, pp. 677–690, 1993.

[28] Y. Wolde-Rufael, "Another look at the relationship between telecommunications investment and economic activity in the United States," *Int. Econ. J.*, vol. 21, pp. 199–205, 2007.

[29] S. L. Arlinghaus, *PHB Practical Handbook of Curve Fitting*. CRC Press, 1994.

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

[31] D. C. Li and I. H. Wen, "A genetic algorithm-based virtual sample generation technique to improve small data set learning," *Neurocomputing*, vol. 143, pp. 222–230, 2014.

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

[33] X. Li, D. Lord, Y. Zhang, and Y. Xie, "Predicting motor vehicle crashes using Support Vector Machine models," *Accid. Anal. Prev.*, vol. 40, no. 4, pp. 1611–1618, 2008.

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

[35] H. F. Liau and D. Isa, “Feature selection for support vector machine-based face-iris multimodal biometric system,” *Expert Syst. Appl.*, vol. 38, no. 9, pp. 1105–11111, 2011.

[36] C. T. Cheng, Z. K. Feng, W. J. Niu, and S. L. Liao, “Heuristic Methods for Reservoir Monthly Inflow Forecasting: A Case Study of Xinfengjiang Reservoir in Pearl River China,” *Water*, vol. 7, no. 8, pp. 4477–4495, 2015.

[37] “Communications & Multimedia Selected Facts & Figures Q1 2010.” Malaysian Communications and Multimedia Commission, 2010.

[38] “Communications & Multimedia Pocket Book of Statistics Q3 2014.” Malaysian Communications and Multimedia Commission, 2014.

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

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

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

[42] “National Transformation Programme Annual Report 2016.” Prime Minister’s Office, Malaysia, 2017.

[43] “Universal Service Provision Annual Report 2008.” Malaysian Communications and Multimedia Commission, 2009.

[44] “Economic Transformation Programme Annual Report 2011.” Prime Minister’s Office, Malaysia, p. 13, 2012.

[45] R. Katz, “Investment, infrastructure and competition in European telecom,” *Intermedia*, vol. 41, no. 2, 2013.

[46] K. Weisul, “Economic indicators: Hot or not?,” *Fortune*, 2010. [Online]. Available: http://archive.fortune.com/2010/06/01/news/economy/economic_indicator.fortune/index.htm.

[47] R. Yu and M. Abdel-Aty, “Utilizing support vector machine in real-time crash risk evaluation,” *Accid. Anal. Prev.*, vol. 51, pp. 252–259, 2013.

[48] A. Azadeh, M. Saberi, A. Kazem, V. Ebrahimipour, and Z. Nourmohammadzadeh, A. Saberi, “A flexible algorithm for fault diagnosis in a centrifugal pump with corrupted data and noise based on ANN and support vector machine with hyper-parameters optimization,” *Appl. Soft Comput.*, vol. 13, no. 3, pp. 1478–1485, 2013.

[49] L. Han, L. Han, and H. Zhao, “Orthogonal support vector machine for credit scoring,” *Eng. Appl. Artif. Intell.*, vol. 26, no. 2, pp. 848–862, 2013.

[50] I. Jala, “We must not forget the villages while developing Malaysia,” *The Star*, 2014. [Online]. Available: https://www.thestar.com.my/business/business-news/2014/07/28/we-must-not-forget-the-villages/.

[51] “World Urbanization Prospects.” United Nations, 2014.

[52] N. Hashim, “The Planning & Implementation of Urban Agglomeration,” in *BENGKEL INISIATIF 22 : MEMACU PERTUMBUHAN MELALUI AGLOMERASI BANDAR*, 2011.

[53] “ITU Facts and Figures 2016.” International Telecommunication Union, 2016.

[54] M. Awad and R. Khanna, *Efficient Learning Machines: Theories, Concepts, and Applications for Engineers and System Designers*. Apress Media, 2015.

[55] T. Mitchell, *Machine Learning*. McGraw-Hill, 1997.

[56] “Genetic Algorithm: Find global minima for highly nonlinear problems.” [Online]. Available: https://www.mathworks.com/discovery/genetic-algorithm.html.

[57] A. Wahab, M. N., S. Nefti-Meziani, and A. Atyabi, “A comprehensive review of swarm optimization algorithms,” *PLoS One*, vol. 10, no. 5. e0122827., 2015.

[58] D. A. Coley, “An Introduction to Genetic Algorithms for Scientists and Engineers.” World Scientific Publishing, Singapore, 1999.

[59] D.-C. Li, C.-S. Wu, T.-I. Tsai, and Y.-S. Lina, “Using mega-trend-diffusion and artificial samples in small data set learning for early flexible manufacturing system scheduling knowledge,” *Comput. Oper. Res.*, pp. 966–982, 2007.

[60] R. Lanouette, J. Thibault, and J. L. Valade, “Process modeling with neural networks using small experimental datasets,” *Comput. Chem. Eng.*, no. 1167–1176, 1999.

[61] P. Niyogi, F. Girosi, and T. Poggio, “Incorporating Prior Information in Machine.” IEEE, 1998.

[62] G. Sarp and M. Ozcelik, “ScienceDirectWater body extraction and change detection using time series:A case study of Lake Burdur, Turkey,” *J. Taibah Univ. Sci.*, vol. 11, pp. 381–391, 2017.

[63] L. O’Gorman, “What is Pattern Recognition,” vol. 25, no. 1. International

Association for Pattern Recognition, 2003.

[64] N. Nandhakumar and J. K. Aggarwal, “The Artificial Intelligence Approach to pattern Recognition – A perspective and an overview,” *Pattern Recognit.*, vol. 18, no. 6, p. 383, 1985.

[65] A. K. Jain, R. P. W. Duin, and J. Mao, “Statistical Pattern Recognition: A Review,” *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 1, 2000.

[66] Y. Abu-Mostafa., “Caltech’s Machine Learning Course - CS 156.” Caltech Academic Media Technologies, 2012.

[67] F. Khan, F. Enzmann, and M. Kersten, “Multi-phase classification by a least-squares support vector machine approach in tomography images of geological samples,” *Solid Earth*, vol. 7, pp. 481–492, 2016.

[68] W. Yu, T. Liu, R. Valdez, M. Gwinn, and M. J. Khoury, “Application of support vector machine modeling for prediction of common diseases: the case of diabetes and pre-diabetes,” *BMC Med. Inform. Decis. Mak.*, 2010.

[69] M. Hassan, “Improvement of Support Vector Machine Classification by Implementing Kalman Filter as Pre-processing Technique.” 2015.

[70] B. Wilson, “Introduction to Predictive Learning LECTURE SET 4 Statistical Learning Theory.” Electrical & Computing Engineering, University of Minnesota.

[71] A. Ghose, “Explain VC dimension and shattering in lucid Way?,” *2013*. [Online]. Available: https://www.quora.com/Explain-VC-dimension-and-shattering-in-lucid-Way. [Accessed: 15-Dec-2018].

[72] KEVINBINZ, “Data Partitioning: Bias vs Variance,” 2014. [Online]. Available: https://kevinbinz.com/2014/08/24/data-partitioning-bias-vs-variance/. [Accessed:

01-Dec-2018].

[73] C. M. Bishop, *Pattern Recognition and Machine Learning*. Springer, 2006.

[74] Ashanira Mat Deris et al., “Overview of Support Vector Machine in Modeling Machining Performances,” in *International Conference on Advances in Engineering*, 2011.

[75] S. Amari and S. Wu, “Improving Support Vector Machine Classifiers by Modifying Kernel Functions,” *Neural Networks*, pp. 783–789, 1999.

[76] C.-C. Chang and C.-J. Lin, “LIBSVM : a library for support vector machines,” *ACM Trans. Intell. Syst. Technol.*, vol. 2, no. 3, 2011.

[77] M. G. Genton, “Classes of Kernels for Machine Learning: A Statistics Perspective,” *J. Mach. Learn. Res.*, no. 2, pp. 299–312, 2001.

[78] K. Korjus, M. N. Hebart, and R. Vicente, “An Efficient Data Partitioning to Improve Classification Performance While Keeping Parameters Interpretable,” *PLoS One*, vol. 11, no. 8, 2016.

[79] “Cross-Validation Explained.” [Online]. Available: http://genome.tugraz.at/proclassify/help/pages/XV.html.

[80] T. V. Gestel, M. Espinoza, J. A. K. Suykens, and C. Brasseur, *Bayesian Input Selection for Nonlinear Regression with LS-SVMS*. IFAC System Identification, 2003.

[81] C. J. C. Burges, “A Tutorial on Support Vector Machines for Pattern Recognition.” Kluwer Academic, 1998.

[82] Y. Su, H. Xu, and L. Yan, “Support vector machine-based open crop model

(SBOCM): Case of rice production in China," *Saudi J. Biol. Sci.*, pp. 537–547, 2017.

[83] "Manual for Measuring ICT Access and Use by Households and Individuals." International Telecommunication Union, 2014.

[84] B. Baumohl, *The Secrets of Economic Indicators, 3rd edition*. New Jersey: Pearson Education, Inc., 2013.

[85] R. L. Katz and P. Koutroumpis, "Measuring digitization: A growth and welfare multiplier," *Technovation*, vol. 33, no. 10–11, pp. 314–319, 2013.

[86] Sprague *et al.*, "Offline and falling behind: Barriers to Internet adoption," McKinsey & Company, 2014.

[87] S. Vargas, "National Broadband Deployment and the Digital Divide." New York: Nova Science Publishers, 2015.

[88] M. Minges, "Exploring the Relationship between Broadband and Economic Growth," in *Background Paper prepared for the World Development Report 2016: Digital Dividends*, The World Bank, Washington DC., 2016.

[89] S. Avgousti, "Mobile video tele-echography robotic platform over 4G-LTE network," PhD Dissertation, University of Orleans & Cyprus University of Technology, 2016.

[90] Andrew Dymond, "Universal Service: Trends, opportunities & best practices for Universal Access to Broadband services," in *8th Annual OOCUR Conference*, 2010.

[91] C. Bouras, A. Gkamas, and T. Tsiatsos, "Best Practices and Strategies for Broadband Deployment: Lessons Learned from Around the World," in *Yogesh*

Dwivedi: Adoption, Usage, and Global Impact of Broadband Technologies, Information Science Reference, USA, 2011, pp. 128–131.

[92] Economic Planning Unit of Malaysian, “Malaysia: 30 Years of Poverty Reduction, Growth and Racial Harmony.” International Bank for Reconstruction and Development, The World Bank, 2004.

[93] Malaysian Communications and Multimedia Commission, “REGULATORY FRAMEWORK FOR COMMUNICATIONS AND MULTIMEDIA INDUSTRY IN MALAYSIA,” in *ITU Workshop on Telecommunication Policy and Regulation for Competition*, 2005.

[94] “South East Asia and Oceania, 2015, Ericsson Mobility Report.” Ericsson, Sweden, 2015.

[95] T. Irawan, “ICT and economic development: comparing ASEAN member states,” *Int. Econ. Econ. Policy*, vol. 11, no. 1–2, pp. 97–114, 2014.

[96] O. Teppayayon and E. Bohlin, “Broadband Universal Service in Europe: A Review of Policy Consultations 2005-2010,” *Commun. Strateg.*, vol. 80, no. 4, 2010.

[97] “Broadband market developments in the EU - Digital Agenda Scoreboard 2015,” 2015.

[98] United Nations Country Team - Malaysia, “Malaysia Achieving the Millennium Development Goals - Successes and Challenges.” United Nations Development Programme, United Nations, 2005.

[99] P. L. Parcu, “Study on Broadband Diffusion: Drivers and Policies.” Independent Regulators Group, following the terms of reference within IRG(11)11, Florence School of Regulation, 2011.

[100] “Handbook for the Collection of Administrative Data on Telecommunication and ICT.” International Telecommunication Union, 2011.

[101] Intel Leap Ahead Whitepaper, “Expanding Universal Service/Access Funds to Help Bridge the Digital Divide.” Intel Corporation, 2008.

[102] “A4AI Affordability Report,” Alliance for Affordable Internet (A4AI), Washington DC, USA., 2017.

[103] Chalita Srinuan, “Understanding the digital divide: Empirical studies of Thailand,” PhD Dissertation, Chalmers University of Technology, Sweden, 2012.

[104] K Salemink et al., “Rural development in the digital age: A systematic literature review on unequal ICT availability, adoption, and use in rural areas,” *Journal of Rural Studies*, vol. 54. Elsevier Ltd., pp. 360–371, 2015.

[105] J. E. Prieger, “The broadband digital divide and the economic benefits of mobile broadband for rural areas,” *Telecommunications Policy*, vol. 37. Elsevier Ltd., pp. 483–502, 2013.

[106] Y. Kim, T. Kelly, and S. Raja, “Building broadband: Strategies and policies for the developing world.” Global Information and Communication Technologies (GICT) Department, World Bank, 2010.

[107] “Working Together to Connect the World by 2020 – Reinforcing Connectivity Initiatives for Universal and Affordable Access,” *2016*. International Telecommunication Union.

[108] R. LaRose et al., “The impact of rural broadband development: Lessons from a natural field experiment,” *Government Information Quarterly*, vol. 28. Elsevier Inc., pp. 91–100, 2011.

[109] B. Whitacre et al., “How much does broadband infrastructure matter? Decomposing the metro–non-metro adoption gap with the help of the National Broadband Map,” *Government Information Quarterly*, vol. 32. Elsevier Inc., pp. 261–269, 2015.

[110] V. B. Haynesworth and A. R. Harrison, “GAO Report (2006) - Telecommunications: Broadband Deployment is Extensive Throughout the United States, but it is Difficult to Assess the Extend of Development Gaps in Rural Areas,” in *Maximizing Broadband Services to Rural Communities*, New York: Nova Science Publishers, 2009, pp. 105–106.

[111] R. Cadman and C. Dineen, “Price and Income Elasticity of Demand for Broadband Subscriptions: A Cross-Sectional Model of OECD Countries.” Strategy and Policy Consultants Network Ltd., United Kingdom, 2008.

[112] “A4AI Affordability Report,” Alliance for Affordable Internet (A4AI), 2013.

[113] World Bank Group, “Strategy and PPP Options for Supporting the ICT Sector and Broadband Connectivity in Somalia, Final Report,” World Bank, 2017.

[114] “Rural-Urban Dynamics and the Millennium Development Goals, Global Monitoring Report 2013 (Advance Edition).” World Bank and the International Monetary Fund, 2013.

[115] L. Hosman, “A National ICT-in-Education Initiative: Macedonia Connects,” in *Sherif Kamel: E-Strategies for Technological Diffusion and Adoption: National ICT Approaches for Socioeconomic Development*, Information Science Reference, USA., 2010, p. 4.

[116] Guldi, Melanie, and Chris M Herbst, “Offline effects of online connecting: the impact of broadband diffusion on teen fertility decisions,” *J. Popul. Econ.*, vol. 30, no. 1, pp. 69–91, 2017.

[117] H. Desalegn, “Techno-Economic Analysis of LTE Deployment: A Case Study of Addis Ababa, Ethiopia,” MSc Dissertation, Addis Ababa University, Ethiopia, 2014.

[118] A. M. Elvidge and J. Martucci, “Telecommunications network total cost of ownership and return on investment modelling,” *BT Technol. J.*, vol. 21, no. 2, 2003.

[119] Analysis Mason, “Models for efficient and effective public-sector interventions in next generation access networks,” *Report prepared for the UK Broadband Stakeholders Group*. 2008.

[120] M. Falch and A. Henten, “Public private partnerships as a tool for stimulating investments in broadband,” *Telecomm. Policy*, vol. 34, no. 9, pp. 496–504, 2010.

[121] C. Blackman and L. Srivastava, *Telecommunications Regulation Handbook - 10th anniversary edition*. Washington, DC.: World Bank, 2011.

[122] T. Kelly, Rossotto, and Carlo Maria, *Broadband Strategies Handbook*. World Bank, 2011.

[123] G. J. Gulati and D. J. Yates, “Different paths to universal access: The impact of policy and regulation on broadband diffusion in the developed and developing worlds,” *Telecommunications Policy*, vol. 36. Elsevier Ltd., pp. 749–761, 2012.

[124] F. Belloc et al., “Whither policy design for broadband penetration? Evidence from 30 OECD countries,” *Telecommunications Policy*, vol. 36. Elsevier Ltd., pp. 382–398, 2012.

[125] Intelecon Research & Consultancy Ltd., “Workshop on Universal Access & Service (UAS) & Broadband Development.” World Bank, Washington D.C., 2009.

[126] Press Release, “Maxis’ 2013 full year net profit rises to RM2.1 billion,” *Maxis Bhd*, 2014. [Online]. Available: https://www.maxis.com.my/en/about-maxis/media-centre/press-releases/2014/02/20140211-en.html.

[127] E. Low, “Maxis’ Sulin Lau on rethinking telco marketing,” *Marketing-Interactive.com*, 2015. [Online]. Available: Maxis’ Sulin Lau on rethinking telco marketing.

[128] Press Release, “CELCOM TO INVEST IN QUALITY TO FUEL GROWTH FOLLOWING STELLAR PERFORMANCE IN Q4, 2013,” *Celcom Axiata Berhad*, 2014. [Online]. Available: https://www.celcom.com.my/Web_Center_Sites/PBO/Files_Corporate/Media_Release/2014/2014-Mar-4.pdf.

[129] “Celcom invest RM100m in flood-prone east coast states,” *The Sun Daily*, 2015. [Online]. Available: http://www.thesundaily.my/news/1298034.

[130] Press Release, “Digi delivers Malaysia’s widest 4G LTE network,” *Digi Telecommunications Sdn Bhd*, 2015. [Online]. Available: http://www.digi.com.my/aboutus/media/press_release_detail.do?id=8812&page=1&year=2015.

[131] T. E. Goh, “U Mobile on the verge of operational profitability,” *Digital News Asia*, 2014. [Online]. Available: https://www.digitalnewsasia.com/mobile-telco/u-mobile-on-the-verge-of-operational-profitability.

[132] Kugan, “Telekom Malaysia buys 57% of Packet One (P1), 4G LTE launch this year,” *Malaysian Wireless*, 2014. [Online]. Available: https://www.malaysianwireless.com/2014/03/telekom-malaysia-buys-57-packet-one-p1/.

[133] Soyacincao, “Altel to invest RM1b in 5 years to roll out LTE network,” *Soyacincao*,

2014. [Online]. Available: https://www.soyacincau.com/2014/04/23/altel-to-invest-rm1b-in-5-years-to-roll-out-lte-network/.

[134] T. E. Goh, “YTL Comms to be profitable soon?,” *Digital News Asia*, 2014. [Online]. Available: https://www.digitalnewsasia.com/mobile-telco/ytl-comms-to-be-profitable-soon.

[135] “Malaysia - Rural population,” *Index Mundi*. [Online]. Available: https://www.indexmundi.com/facts/malaysia/rural-population.

[136] J. R. Schneir and Y. Xiong, “A cost study of fixed broadband access networks for rural area,” *Telecommunications Policy*, vol. 40. Elsevier Ltd., pp. 755–773, 2016.

[137] “Global mobile trends.” GSMA Intelligence, London, 2016.

[138] M. Abbas, “Connecting the Unconnected: Bridging the Digital Divide Using WiMAX,” in *Wireless World CeBIT*, 2009.

[139] A. Muente-Kunigami and J. Navas-Sabater, *Options to Increase Access to Telecommunications Services in Rural and Low-Income Areas*. The World Bank, Washington DC., 2009.

[140] Economics and Development Resource Center, “GUIDELINES FOR THE ECONOMIC ANALYSIS OF TELECOMMUNICATIONS PROJECTS.” Asian Development Bank, 1997.

[141] M. Kantor *et al.*, “General framework for techno-economic analysis of next generation access networks,” *12th International Conference on Transparent Optical Networks*. Munich, pp. 1–4, 2010.

[142] Sofia Verbruggee, K. Casier, J. Van Oofeghem, and B. Lanno, “White paper: Practical Steps in Techno-Economic Evaluation of Network Deployment

Planning," Department of Information Technology (INTEC), UGent/IBBT, 2009.

[143] B. T. Olsen, L. Henden, A. F. Hansen, and M. Lähteenoja, "The Core of Techno-Economics." Telenor R&L, 2009.

[144] "Whitepaper: The Economics of Networking," 2011. [Online]. Available: https://www.cisco.com/c/dam/en/us/solutions/collateral/enterprise-networks/white_paper_c11-687149.pdf.

[145] T. Smura, "Techno-Economic Analysis of IEEE 802.16a-Based Fixed Wireless Access Networks," MSc Thesis Dissertation, Aalto University, Helsinki, 2007.

[146] R. Prasad and F. J. Velez, "Business Models and Cost/Revenue Optimization," in *WiMAX Networks: Techno-Economic Vision and Challenges*, Springer Science + Business Media, 2010, pp. 395 – 421.

[147] A. Firli, I. Primiana, and U. Kaltum, "The Impact of Increasing CAPEX on Customer Number, Profit, and ROI in Indonesia Telecommunication Industry," *Am. J. Econ.*, vol. 5, no. 2, pp. 135–138, 2005.

[148] M. Gaynor, "Network Services Investment Guide – Maximizing ROI in Uncertain Times." Wiley Publishing, Inc., Indiana., 2003.

[149] "Design-Expert Software by Stat-Ease, Inc." [Online]. Available: https://www.statease.com/. [Accessed: 12-Dec-2018].

[150] "World Development Indicators," *World Bank*. [Online]. Available: http://databank.worldbank.org/data/source/world-development-indicators.

[151] J. D. Evans, *Straightforward statistics for the behavioral sciences*. Pacific Grove : Brooks/Cole, 1996.

[152] G. Nargund, “Declining birth rate in Developed Countries: A radical policy re-think is required,” *F, V V ObGyn*, vol. 1, no. 3, pp. 191–193, 2009.

[153] “World Fertility Patterns 2015,” Department of Economic and Social Affairs, Population Division. United Nations, 2015.

[154] “The Lancet. ‘Life expectancy set to increase in developed nations, potentially surpassing 90 years in some countries.,’” *ScienceDaily*, 2017. [Online]. Available: www.sciencedaily.com/releases/2017/02/170221222524.htm.

[155] A. Khan, S. Khan, and M. Khan, “Factors effecting life expectency in developed and developing countries of the world,” *Int. J. Phys. Educ. Sport.*, vol. 1, no. 1, pp. 04–06, 2016.

[156] “The Development of Fixed Broadband Networks,” *OECD Digital Economy Papers, No. 239*. OECD, 2014.

[157] A. Odusola and B. Abidoye, “Effects of Temperature and Rainfall Shocks on Economic Growth in Africa,” in *29th Triennial Conference of the International Association of Agricultural Economists (IAAE)*, 2015.

[158] M. Dell, B. F. Jones, and B. A. Olken, “Temperature Shocks and Economic Growth: Evidence from the Last Half Century,” *Am. Econ. J. Macroecon.*, vol. 4, no. 3, pp. 66–95, 2012.

[159] “What is MATLAB.” [Online]. Available: https://cimss.ssec.wisc.edu/wxwise/class/aos340/spr00/whatismatlab.htm. [Accessed: 09-Dec-2018].

[160] D. Gale Johnson, “Agricultural economics.” [Online]. Available: https://www.britannica.com/topic/agricultural-economics. [Accessed: 09-Dec-2018].

[161] D. Brain and G. I. Webb, “On the effect of data set size on bias and variance in classification learning,” in *Fourth Australian Knowledge Acquisition Workshop (AKAW) ’99*, 2000, pp. 117–128.

[162] J. E. JACKSON, *A User’s Guide To Principal Components*. A Wiley-Interscience Publication, 1991.

[163] N. Deng, Y. Tian, and C. Zhang, “Support vector machines: optimization based theory, algorithms, and extensions.” CRC Press, Taylor and Francis Group ISBN 978: 363., 2013.

[164] Z. Li, W. Zhang, L. He, and J. Liu, “Speaker recognition with kernel based IVEC-SVM,” *Acta Autom. Sin.*, vol. 40, pp. 780–784, 2014.

[165] “A Complete List of Kernels Used in Support Vector Machines,” *Biochem. Pharmacol. Open Access*, vol. 4, no. 5, 2015.