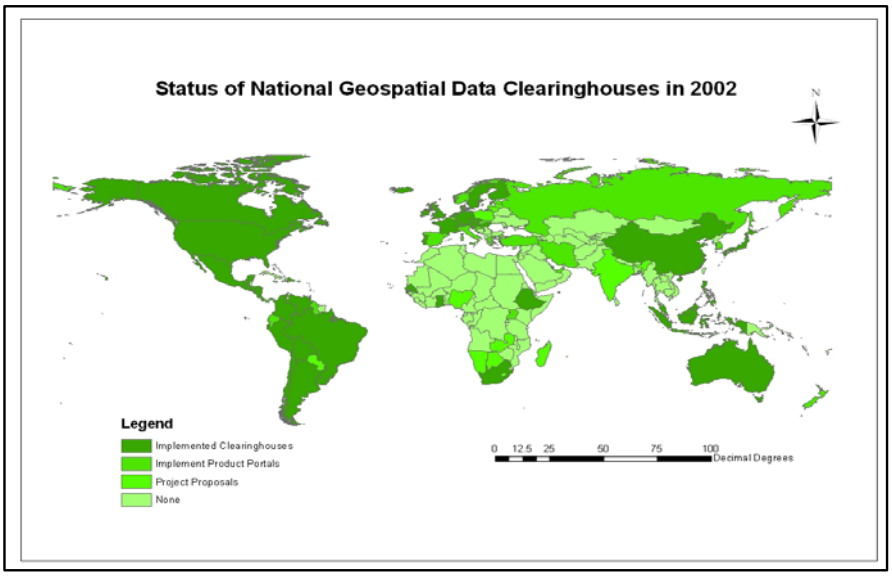
Centre for Geo-Information

Thesis Report GIRS-2005-012

Using data mining to determine the societal impact on national geospatial data clearinghouses

Jun Wang

Apirl 2005



WAGENINGEN UNIVERSITY
WAGENINGEN UR

Registration number: 691229-927-050

Supervisors:
Dr. Monica Wachowicz
Ir. Joep Crompvoets

A thesis submitted in partial fulfillment of the degree of Master of Science at Wageningen University and Research Centre, The Netherlands.

Apirl 2005
Wageningen, The Netherlands

Thesis code number: GRS-2005-012
Wageningen University and Research Centre
Laboratory of Geo-Information Science and Remote Sensing
Thesis Report: GIRS-2005-012

ACKNOWLEDGEMENT

During the period of my study, I have received various assistances and supports form many people and Wageningen University that I am grateful very much.

First and for most, with the deepest sense of honor I would wish to extend my zealous thanks to my worthy supervisors, Dr. Monica Wachowicz and Ir. Joep Crompvoets. Under whom affectionate supervision, accommodative attitude, continuous inspiration, encouragement and scholastic guidance, this research was planned, executed and completed. I will never forget their kind attitude, keen interest and way of teaching. They are so friendly that visiting them for discussions without prior appointments was never a problem.

I would like to express my heartfelt thanks to my teachers for giving me the academic and professional supports to peruse my study.

My thanks also will go to all my sincere classmates and friends, thanks for their truly friendship and warmly accompany in this study. I wish you all the best, nothing but the best. It has been an honor and pleasure to know you.

However, I could not finish without my parents, nothing could ever eclipse the true love, the unconditional dedication, and the assistant support they had. I could never have achieved my dreams without you.

Last but not least, I wish to express my sincere indebtedness to my wife. I thank you for your love and understanding.

ABSTRACT

The new technology revolution featured by information technology enables people to obtain enormous amount of information about the earth and human society. The geospatial information, which is an important part of the overall global information resources, has been given great attention and widely used. A new infrastructure, which is the spatial data infrastructure (SDI), has been vigorously developed. Now, many countries have a spatial data infrastructure of themselves named national data infrastructures (NSDI). In order to share the information, the geospatial data clearinghouse (GDC) is created. It is the core element of NSDI. Whether the GDC is successful has great influence to the development of the NSDI. This thesis explores the impact of characteristics of society on GDCs using data mining and predicts whether the countries that do not have the GDC now, can create the successful ones. The decision tree method of data mining was selected for the study. The study area includes 193 countries and regions. The data set includes the 22 attributes of GDCs and 252 attributes of society. All data are about the end of December 2002. One decision tree was used to assess societal impact on the existence of GDCs. Two decision trees were used to determine the societal impact on the success of GDCs. The prediction about whether the countries are able to have the successful GDCs has been done on 126 countries that do not have GDCs now. Based on the expert knowledge, the results seem to be reasonable. It is recommended that frequent updates are needed as the statuses of clearinghouses change at different years.

Key words: data mining, decision tree, geospatial data clearinghouse, impact analysis, NSDI

Table of Contents

List of Figures

List of Tables

List of Equations

List of abbreviations

CART	Classification and Regression Trees
CHAID	Chi-squared Automatic Interaction Detection
GDC	Geospatial data clearinghouse
GDP	Gross domestic product
EK	Expert knowledge method
KDD	Knowledge discovery in databases
LHS	Left hand side
MP	Median partition method
NSDI	National spatial data infrastructure
OLAP	On-Line Analytical Processing
RHS	Right hand side
SDI	Spatial data infrastructure
SQL	Structured Query Language

CHAPTER 1
INTRODUCTION

1.1 Background

Today, the world is entering the information age where the new technology revolution featured by information technology enables people to obtain enormous amount of information about the earth and human society. The information can then be integrated and analyzed organically for addressing such globally concerned issues as resources, environment, population and disasters to face the challenge of the changing world. In this process, the geospatial information, which is an important part of the overall global information resources, has been given great attention and widely used. At the same time, it has become an increasing concern how to realize the data sharing. A new infrastructure, which is the spatial data infrastructure (SDI), has been vigorously developed accordingly (Groot and McLaughlin, 2000). Now, many countries have spatial data infrastructures of themselves named national data infrastructures (NSDI). The goals of these Infrastructures are to make geospatial data more accessible to the public, to increase the benefits of using available data, to reduce duplication of effort among agencies, to improve quality and to reduce costs related to geospatial data (FGDC, 2004). In order to complete these objectives, the geospatial data clearinghouse (GDC) is necessary.

Fortunately, as the Internet developments, the need becomes feasible to create an agency or an electronic catalog service to unify and supply the information of definition of geospatial data and geospatial data. So the GDC comes out. The GDC becomes central to the establishment of the NSDI. An excellent GDC facilitates the sharing data, and saving the time used in collecting data and economizing the costs. But the implemrntation of the excellent GDC is determined by the national economy, politics, policy, technology, culture and education and so on (Groot and McLaughlin, 2000). How to make a GDC successful? The societal aspects of nations (economy, politics, technology, culture and education and so on) should function importantly. Which aspects have the great impacts on the GDC? The analysis of action of societal aspects on the GDCs can help us to direct and predict the development and implementation of GDCs.

Data mining is a good technique and tool, because it refers to the process of extracting interesting, non-trivial, implicit, previously unknown and potentially knowledge or patterns from data (Han and Kamber, 2001). A significant distinction between data mining and other analytical tools is the approach used in exploring the data relationships. Many of the analytical tools available support a verification-based approach, in which the user hypothesizes about specific data relationships and then uses the tools to verify or refute those hypotheses. This approach relies on the intuition of the analyst to pose the original question and refine the analysis based on the results of potentially complex queries against a database. The effectiveness of this verification-based analysis is limited by a number of factors, including the ability of the analyst to pose appropriate questions and quickly return results. Data mining, in contrast to these analytical tools, uses discovery-based approaches in which

classification and other methods are employed to determine the key relationships in the data (Moxon, 1996). It does not need any hypothesis.

What we want to know is what and how the aspects of society have an impact on the GDCs. Therefore a lot of data about the aspects of society have been collected. The analysis is not based on any hypothesis about which aspects have more relevant and how they could have an impact on GDCs. It is difficult to create the hypothesis depended on the intuition of analyst. But we hope to find the relationships among aspects of society and GDCs by exploring the various data about the society. The discovery based approach was the best suitable for the study. As a result, the data mining technique was selected.

1.2 Problem definition

During the last years, many GDCs have been implemented around the world. They help to satisfy the objective of geospatial data sharing across networks. A successful GDC facilitates the objectives of sharing geospatial data and saving costs and the time used in collecting data. But the developments of GDCs are very different in different countries. Because the existence and development of GDCs depend on the society, the aspects of society become the important studied objects to assess the GDCs. So the criteria based on aspects of society can be used to assess whether a GDC is implemented in a country or not and whether a GDC is successful or not. They can also indicate us the trace of the development of GDC and helps us to make the precise prediction of the progress of a GDC.

Every aspect of society has many attributes. Therefore the criteria were selected on the series of these attributes of societal aspect. The classification function of data mining was used to generate these criteria. At this moment, no study has been done on such a set of criteria based on societal aspects in order to assess the existence and success of GDCs. This thesis is a attempts toward this orientation

1.3 Objectives

The objectives are:

To discover the set of attributes and their correspondent criteria that can be used to determine the impact of societal aspects have on the existence of GDCs in different countries in the world.
To discover the set of attributes and their correspondent criteria that can be used to determine the impact of societal aspects have on the success of GDCs in different countries in the world.
To predict which countries will be able to implement successful GDCs in the future.

1.4 Research questions

This research focuses on these questions:

How to apply the data mining (decision tree approach) to determine the impact of society on the existence and success of GDCs?

Sub- research questions:

What are the criteria that can be used to assess the existence of GDCs?
What are the criteria that can be used to assess the success of GDCs?
What kinds of attributes are important in these criteria?
Is it possible to make the prediction about the success of GDCs?

1.5 Outline of the thesis

This thesis report is written as follows:

Chapter 2 provides a literature review that focuses on the definitions, functions, relationships and development of NSDI and GDC.

Chapter 3 provides a literature review that introduces the definition, classification of data mining and focuses on data preprocessing and decision tree approach.

Chapter 4 describes the study area, dataset and detailed procedures of the study methodology.

Chapter 5 explains the outcomes of the study. The results are shown and discussed.

Chapter 6 presents the conclusions and recommendations.

CHAPTER 2
Spatial Data Infrastructures

This chapter introduces NDSI including the definition of NSDI, the function of the NSDI, what drives the development of NSDI and how the NSDI develops and GDC. GDCs are described including the definition, the developments of GDCs, the relationships between NSDI and GDC and the classification of GDCs by their status in 2002.

2.1 The definition of geospatial data and NSDI

Geospatial data are items of information related to a location on earth, particularly information on national phenomena, culture and human resources. Examples are topography, including geographic features, place names, height data, land cover, hydrography, cadastre; administrative boundaries; resources and environment; socio-economic information, including demographics (OSDM, 2004). Governments use these data for their own purposes in legislative and policy development for the allocation and management of natural resources, for defense and public safety purpose, in support of a variety of regulatory activities, and in promoting a better understanding of the physical, economic and human geography of the nation. A vast array of private or common agencies and organizations also collect geospatial data for a wide variety of commercial, social and environmental applications (Groot and McLaughlin, 2000).

The development of application of geospatial data gives an opportunity for the emergence of SDI. A SDI first emerged in the earlier 1980s in Canada as mechanism to provide an effective collection, processing, store, management, accessing and sharing of this data (Canadian Government, 1986).

"National Spatial Data Infrastructure" (NSDI) concerns the various geo-information in a nation. It means the technology, policies, standards, and human resources necessary to acquire, process, store, distribute, and improve utilization of geospatial data (Executive Order, 1994).

2.2 The structure of NSDI

The applications specific modules, communications network, GDC, metadata, framework, geospatial data and standards construct the practical application structure of NSDI (ISRO, 2001). The Figure 2.1 shows the structure of NSDI.

NSDI Standards

In order to share the geospatial data, the NSDI requires a major effort at standardizing content and schemas, design and process, network protocols, exchange and transfer. The standardization can enable “user transparency” to information access. The standardization includes database standardization - formats, exchange and interoperability; Networks-gateways and protocols; communication equipment,

software standards, etc. Standards enable applications and technology to work together (FGDC, 2004).

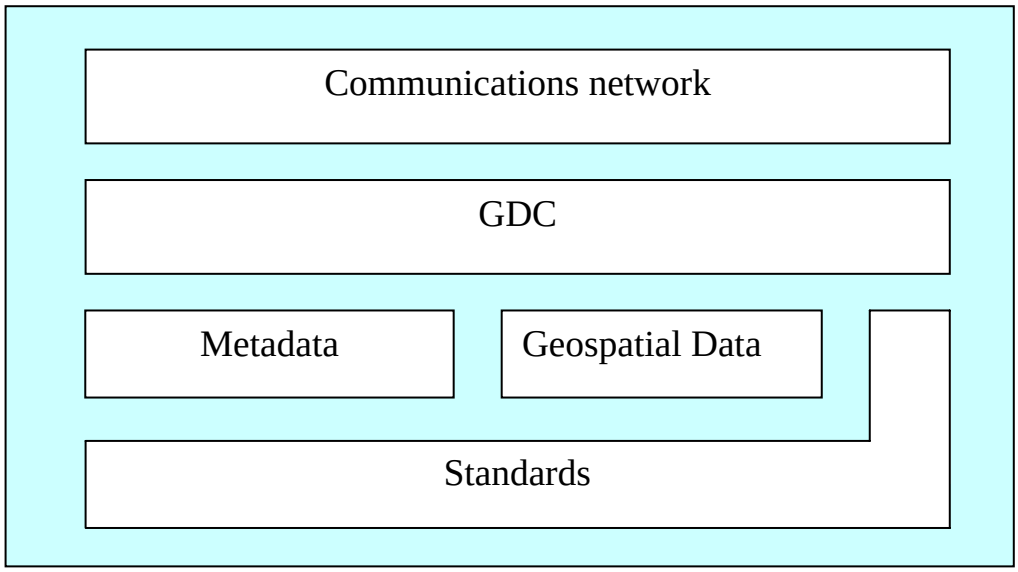


Figure 2.1 Constructional elements of NSDI (Source: ISRO, 2001)

Metadata of the NSDI

Metadata is “data about data”. Normally, the information of metadata includes nine categories: data set identification, data set overview, data set quality indictors, geo spatial reference system, extent, data definition, classification, administrative metadata and metadata reference (CEN, 1996).

Communication network

Communication network connects the computers and servers. It is hardware and software to connect the users and information resources. It is the fundament of exchanging the information.

GDC.

The GDC is the core of NSDI. GDC is the mechanism to provide access to the metadata and finally to the actual data sets. The GDC is a system to authenticate data requests and respond the requests. The GDC uses communication network to connect the users and geospatial data resources and uses the access protocols engines to look for and discover geospatial data.

2.3 The functions of NSDI

Sharing spatial data

Collecting, storing, managing and updating the geospatial data are the expensive and use amount of human resources and time. In order to avoid duplication of expenses and time, people generate the desirability to share the spatial data. Often the geospatial data for one application can be applied in others. This situation provides the possibility for sharing geospatial data. For many organizations, building and using a database of geospatial data requires large quantities of current and accurate digital data. They can save significant time, money and effort when they share the burden of data collection and maintenance with others. This is important, not only to the organizations looking for the data, but also for the organizations with the data. The more partners there are, the more the savings and the greater the efficiency. Sharing data can also improve data quality by increasing the number of individuals who find and correct errors (Rajabifard, 2003).

Supporting the decision making

Decision-making can broadly be defined to include choice or selection of alternative course of action (Malczeweki, 1999). A preliminary step toward achieving decision – making for complex problems has been increasing recognition of the role of geospatial information to generate knowledge, provide added value to identify problems, assist in proposing alternatives and defining a course of action, information discovery, assess and use. The need to integrate geospatial data from different sources gain momentum due to the growing attention at the end of previous century for sustainable development. The importance of geospatial information to support decision-making and management of growing national, regional, and global issues, such as deforestation and pollution, has been become one of themes on sustainable development (CSD, 2001).

2.4 Forces driving the development of NSDI

There are two forces driving the development of NDSI (Rajabifard *et al*, 2003).

The first is a growing need for government and business to improve their decision-making and increase their efficiency with the help of proper geospatial analysis (Gore, 1998). In most of developed countries it is widely acknowledged that NSDI is part of the national infrastructure and extensive efforts are being expended on this (Clarke, 2000). In the last two decades nations have made unprecedented investments in information and the means to assemble, store, process, analyze and disseminate it. Many organizations, agencies and departments in all levels of government, private and non-profit sectors and academia throughout the world spend billions of dollars each year producing and using geospatial information (FGDC, 1997).

The second force is the advent of cheap, powerful information and communications technology, which facilitate the more effective handling of large quantities of

geospatial data (Openshaw, 1993). The rapid advancement in geospatial data capture technologies has made the capture of digital data a relatively quick and easy process, such as satellite imagery with digital image processing techniques as well as using global positioning system (Openshaw, 1993).

2.5 The development of NSDI

So far, the NSDIs have developed two generations (Rajabifard *et al*, 2003). The first generation of NSDI development was emerged in the mid-1980s. At this time, countries developing NSDI on any jurisdictional level had only very limited ideas and knowledge about different dimensions and issues of the NSDI concept, and rather less experience of such development.

In this period, countries designed and developed NSDI based on their specific requirements and priorities and nationally specific characteristics. The ultimate objectives of the NSDI initiatives in this generation were to promote economic development, to stimulate better government and to foster environmental sustainability (Masser, 1999). Since then these countries have become more aware of different dimensions of SDI development and have therefore been able to identify emerging issues and challenges involved in the NSDI concept.

A significant milestone overcame by the first generation, for whom there were few experiences and existing NSDI developments from which to learn, was the documentation of researchers' and practitioners' experiences and status reports on their NSDI initiatives and as part of that report on their GDC activities which facilitated their NSDI initiatives. This achievement not only gave countries a knowledge-base from which to learn and/or develop their initiatives, providing exposure to the developmental strengths and weaknesses of different NSDI initiatives, but it provided social capital to share and foster NSDI development in other countries.(Rajabifard *et al*. 2002).

The second generation started from the year 2000 when some of the leading nations on NSDI development changed their development strategies and update their NSDI conceptual models. This led to a rapid increase in the number of countries becoming involved in NSDI development, fostered by the definition of an SDI community where experiences could be shared and exchanged experiences. This shows the continuum of strategic geospatial data development. In second-generation NSDI, the strategy for NSDI development is changing towards a more process-based approach from data-based approach. (Rajabifard *et al*. 2003).

The second generation of NSDI developments characteristically falls into two groups: those countries that started to develop an NSDI initiative during the period of the first generation and are gradually modifying and upgrading the initiative, as well as those countries that have recently decided to design and develop an NSDI for their respective countries and/or have just commenced doing so (Hyman and Lance, 2001).

For the first generation, data were the key driver for SDI development and the focus of initiative development. For the second generation, the use of that data (and data applications) and the needs of users are the driving force for NSDI development. In summary, second-generation SDI development has been relatively quick due to the concept gaining momentum and because of the existence of early prototypes, clarification on many initial design issues, increased sharing and documentation of experiences to facilitate implementation and face the complexity of decision-support challenges (Rajabifard *et al*. 2003).

2.6 Geospatial data clearinghouses

2.6.1 The definition of GDC

A GDC is a system of software and organization which is in the intermediary between the users and the suppliers to facilitate the discovery, evaluation and downloading of geospatial data (FGDC, 2004). GDC provides the information about geospatial data over the Internet. It contains the information of metadata that will be used to query geospatial data by users. It could be a distributed network of geospatial data producers, managers, and users linked electronically (Executive Order, 1994).

The GDC is implemented using a multi-tier software architecture that includes a Client tier, a middleware or "Gateway" tier, and a server tier, as is illustrated in Figure 2.2(USGS, 2005).

The client tier consists of a traditional Web browser or a native search client application. The Web browser uses conventional Hypertext Transport Protocol (HTTP) communications, whereas the native search client uses some protocols directly against a set of servers.

The middle tier in the architecture includes a World Wide Web to a gateway. The gateway provides the transformation of protocols, transition of query and search results. The gateway gives the uniform portal of users and query forms which can be downloaded from gateway that lets the user define the geographic area and time period of interest, search against text fields or full-text, and select which servers to search.

At the bottom tier of the architecture are the servers. These servers provide the results of query and geospatial data demanded by users. The servers are responsible for the geospatial database and metadata management.

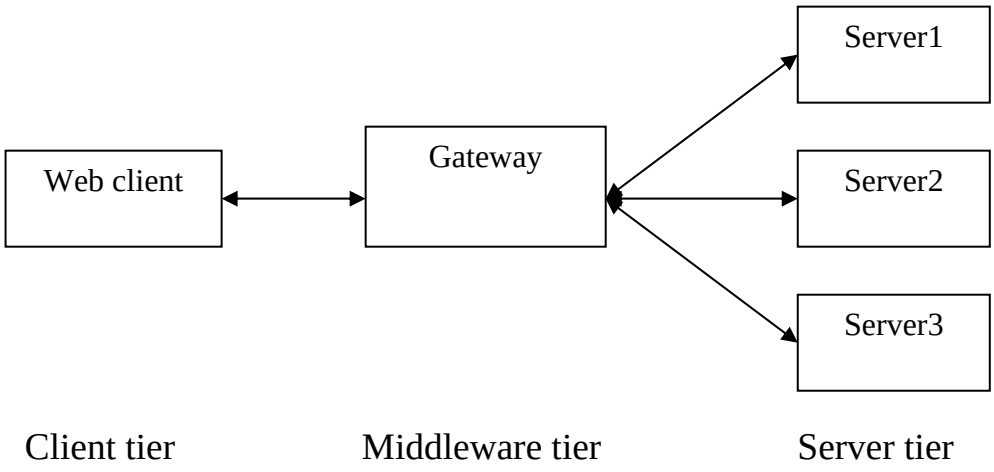


Figure 2.2 Architecture of the GDC (Source: USGS, 2005).

2.6.2 The development of GDCs

In 1994, the NSDI of the U.S. was officially launched to coordinate the geospatial data collection and management activities between governmental and non-governmental organizations in the United States. This means the beginning of the implementation of the first GDC (FGDC 2004).

When the first GDC has been set up, the FGDC focuses on the extension of the GDC network that provides 'one-stop' access to standardized geospatial data, applications, programs and products from all federal agencies and incorporates similar non-federal information, and it establishes web mapping and online data services to meet general requirements of government and citizens users (FGDC 2004).

Generally, the community of geospatial data providers and users is loose. So the GDC becomes a decentralized system of servers located on the Internet. Metadata are collected in a standard format in order to facilitate query and consistent presentation across multiple participating sites

Primarily designed to facilitate sharing of data collected and managed by U.S. Federal government activities, the GDC has been widely deployed in the U.S. and other countries, linking geospatial data users with the geospatial data providers of all types.
The main works on the GDC are improved the metadata and search engine for different countries.

2.6.3 The relationship between GDC and NSDI

The GDC has two main objectives and functions: the metadata manager and the query processor. As the metadata manager, the GDC announces the strict standards to the

data providers, and let them to make their geospatial data as these standards. The geospatial data providers can describe the available geospatial data on the GDC by electronic form, as the same time, the provider can offer the users access his geospatial data. As the query processor, the GDC becomes center point to supply all of information of geospatial data. It provides the name, attributes, scale, price and how to communicate to the provider of data *et al*. GDC even can provide hypertext linkages within their metadata entries that enable users to directly download the digital data set in one or more formats.

For the NSDI, the GDC is the core element. It provides a virtual consolidated information space across which searches of geospatial data may be conducted through a single query. The development of GDC provides much better metadata management and query process, which are the main functions of NSDI. As a result, the sharing geospatial data becomes more efficient and ease. So the development of GDC can be considered the driving force that will strengthen the progress of NSDI.

2.6.4 The classification of GDCs

Due to the difference of technological, economical level, the GDCs stay at the different status. The GDCs can be divided according to their status into: clearinghouse, product portal, and project.

The clearinghouse status implies that a nation implemented a National geospatial data clearinghouse for managing metadata, processing query, exchanging geospatial data, regulating delivery and reporting trading data. It most supports a software architecture with the whole functions (The Pit Master, 2005).

The product portal status can be considered as a simple GDC. It is a geospatial data product gateway. As a World Wide Web site, it is or proposes to be a major starting site for users when they connect to the Web and want to search some geospatial data. It provides some query form, a directory of Web sites that have geospatial data. But it has not the other functions that an implemented GDC has (Answers, 2005).

Project is the GDC or product portal that is being developed and has not been implemented.

CHAPTER 3
Knowledge Discovery

Knowledge discovery in database (KDD) is a process. The data mining is a part of it. The decision tree is one of classification methods of data mining. This chapter introduces the process of KDD, and mainly focuses on methods of processing the missing values, methods of data mining and the theory of the decision tree approach.

3.1 The Knowledge Discovery Process

The steps in the knowledge discovery process are defining the problem, collecting and preprocessing data, data mining, validating the models, deploying the model, monitoring (Hamilton, 2001).

Defining the problem is to identify the goals of the knowledge discovery project and verify that the goals are actionable (Fayyad, 1996).

Collecting and preprocessing data are to collect data form heterogeneous data sources, create the database and process the incomplete, noisy and inconsistent data in the data base, reduce the errors and make the data suit data mining.

Data mining is to build the model. The model-building step involves selecting data mining tools, generating samples (as necessary) for training, testing samples and, finally, using the tools to build, test and select models. (StatSoft, Inc., 2004)

Validating the models is to test the model for accuracy on an independent dataset that has not been used to create the model, assess the sensitivity of a model and pilot of testing the model for usability. At most time, data mining and validating model are the unified and iterative process.

Deploying the model, for a predictive model, is to use the model to predict results for new cases, then to use the prediction to alter organizational behavior. Deployment may require building computerized systems that capture the appropriate data and generate a prediction in real time so that a decision maker can apply the prediction.

Monitoring models are necessary, if the models can be used. Whenever you are modeling, it is likely to change over time. So the model that was correct yesterday may no longer be very good tomorrow. Monitoring models requires constant revalidation of the model on new data to assess if the model is still appropriate.

3.2 Data preprocessing

Data preprocessing is to resolve representation and encoding differences, join data from various tables to create a homogeneous source, check and resolve data conflicts,

outliers (unusual or exception values), missing data, and ambiguity, use conversions and combinations to generate new data fields such as ratios or rolled-up summaries after the data were collected from various internal and external sources. These steps require considerable effort, often as much as much as 70 percent or more of the total data mining effort.

Data preprocessing has 5 tasks that should be done if it is necessary. They are data cleaning, data transformation, data integration, data reduction and data discretization (Han and Kamber, 2000).

Data cleaning task is to fill in missing values, identify outliers and smooth out noisy data, correct inconsistent data and resolve redundancy caused by data integration.

Data transformation task is to aggregate data, generalize data, normalize data and construct the new attributes from the given ones.

Data integration task is to identify real world entities from multiple data sources and combine data from multiple sources into a coherent store.

Data reduction task is to reduce the Dimensionality that means, remove unimportant attributes, compress data, reduce numerosity to fit data into models

Data discretization task is to reduce the number of values for a given continuous attribute by dividing the range of the attribute into intervals. Interval labels can then be used to replace actual data values.

In order to study the effects of societal aspects on GDC, the tasks of the collecting data and most of preprocessing data have been carried out by my supervisors except for the processing of the missing values. Therefore the next section will focus on the methods developed for processing the missing values.

3.2.1 Processing missing values

Missing values maybe exist in most of data used by data mining project. Missing values may be due to various reasons. For example: equipment malfunctions lose data, some data is deleted because of inconsistent with other recorded data, data not entered due to misunderstanding and data may not be considered important at the time of entry and so on (Han and Kamber, 2000).

In order to reduce the influence of missing values, one strategy is to exclude records (normally, it is row in the table) with any missing data from database. These records are called incomplete cases. The records that do not have missing values are called complete case. While using only complete cases makes the processing missing values simplicity, the information in the incomplete cases is lost. This approach also ignores the possible systematic difference between the complete cases and incomplete cases, and the resulting inference may not be applicable to the population of all cases, especially with a smaller number of complete cases (Bao,2002)..

Another strategy is single imputation, in which a value substitutes the missing value. Analysis procedures can then be used with the filled-in data set. For example, each missing value can be imputed from the attribute mean of the complete cases, or it can be imputed from the mean conditional on observed values of other attributes. This approach treats missing values as if they were known in the analysis. Single imputation does not reflect the uncertainty about the predictions of the unknown missing values. Because the missing values are replaced by artificially created "average" data point, single imputation may considerably change the values of correlations (Bao,2002).

Four methods will be used to processing the missing values to reduce the influence of using values to substitute the missing values in this study. They are mean method, median method, field mean method and field median method. The mean method and median method are using the attribute mean or median to replace the missing values in this attribute. The field mean method and the field median method are using the mean or median of all samples belonging to the same class in the attribute to substitute the missing values in this class. They are simply imputation methods.

3.3 Data mining

Data mining is an information extraction activity whose goal is to discover hidden facts contained in databases. Using a combination of machine learning, statistical analysis, modeling techniques and database technology, data mining finds patterns and subtle relationships in data and infers rules that allow the prediction of future results (Two Crows Corporation, 2004).

Data mining methods may be classified into 4 kinds by the function. They are classification, association, sequential or temporal patterns and clustering or segmentation (Dilly, 1995)

Classification

Classification refers to the data mining problem of attempting to predict the category of categorical data by building a model based on some predictor variables. Classification is the most common types of problems to which data mining is applied today. Data miners use classification to find the rule of development of cases and predict the happening of the cases. Various data mining techniques are available for classification, and some techniques have several algorithms. These techniques produce very different models (Dilly, 1995).

Associations

Associations creates rules that describe how often events have occurred together. Given a collection of items and a set of records, each of which contains some number of items from the given collection, an association function is an operation against this set of records that return affinities or patterns that exist among the collection of items. These patterns can be expressed by rules such as "72% of all the records that contain items A, B and C also not contain items D and E." The specific percentage of occurrences (in this case 72%) is called the confidence factor of the rule. Also, in this

rule, A, B and C are said to be on an opposite side of the rule to D and E. Associations can involve any number of items on either side of the rule.

Sequential/Temporal patterns

Sequential/temporal pattern functions analyze a collection of records over a period of time for example to identify trends. Where the identity of a customer who made a purchase is known an analysis can be made of the collection of related records of the same structure (i.e. consisting of a number of items drawn from a given collection of items). The records are related by the identity of the customer who did the repeated purchases. Such a situation is typical of a direct mail application where for example a catalogue merchant has the information, for each customer, of the sets of products that the customer buys in every purchase order. A sequential pattern function will analyze such collections of related records and will detect frequently occurring patterns of products bought over time. A sequential pattern operator could also be used to discover for example the set of purchases that frequently precedes the purchase of a microwave oven.

Clustering/Segmentation

Clustering and segmentation are the processes of creating a partition so that all the members of each set of the partition are similar according to some metric. A cluster is a set of objects grouped together because of their similarity or proximity. Objects are often decomposed into an exhaustive and/or mutually exclusive set of clusters.

Clustering according to similarity is a very powerful technique, the key to it being to translate some intuitive measure of similarity into a quantitative measure. When learning is unsupervised then the system has to discover its own classes i.e. the system clusters the data in the database. The system has to discover subsets of related objects in the training set and then it has to find descriptions that describe each of these subsets.

There are a number of approaches for forming clusters. One approach is to form rules that dictate membership in the same group based on the level of similarity between members. Another approach is to build set functions that measure some property of partitions as functions of some parameter of the partition.

3.3.1 Main modeling techniques of classification

There are four techniques that dominate the available classification tools today (Brand and Gerritsen, 1998). Briefly, they are:

Decision tree technique

A decision tree is a technique that generates a graphic representation of the model it produces. As the name implies, the graphic output is similar in structure to a tree. It is also usually accompanied by rules of the form "if condition then outcome" which constitute the text version of the model. Decision trees have become very popular tools

because users easily understand the results. In this study, the criteria and their relevant societal attributes need to be found. The decision tree and the rules of the decision tree can be considered as the appropriate criteria. The relevant attributes can also be obtained from the rules of the decision tree. Finally, the decision tree can also be used to carry out the prediction.

Neural network technique

Neural networks are based on an early model of human brain function, and they are equally effective for classification and regression. Often referred to as a "black box" technology, neural networks are more complicated than other techniques. They require setting numerous training parameters and, unlike decision trees, provide no easily understandable output. The output from a neural network is purely predictive. Because the neural network model cannot provide the description of component, a neural network's choices are harder to understand, and as a result, the validation becomes difficult, because we cannot provide enough data to validate the neural network.

Naive-Bayes technique

Naive-Bayes is a technique that limits its inputs to categorical data, and it is applicable only to classification. (This technique is named after Bayes’s theorem, and it is also called Simple Bayes. The modifier "simple" or "naive" is used because the algorithm assumes variables are independent when they may not be.) Simplicity and speed make Naive-Bayes an ideal exploratory tool. The technique is based on a simple concept: conditional probabilities derived from observed frequencies in the training data. This is valid only if the assumption of statistical independence between the various independent variables. In this study, it was impossible to realize that the attributes of societal aspects were statistically independence.

K-nearest neighbor technique

K-nearest neighbor (K-NN) differs from the other techniques mentioned in that it has no distinct training phase because the training data is actually the model. Predictions for a new case are made by finding a group with the most similar cases ("k" refers to the number of items in this group) and using their predominant outcome for the predicted value. The nearest neighbor technique has a drawback when compared to the decision tree. The drawback is a lack of descriptive output, which might influence the founding of the important attributes of societal aspects that affect the development of GDCs.

3.3.2 The decision tree approach

A decision tree is a model that is both predictive and descriptive. It is called a decision tree because the resulting model is presented in the form of a tree structure. The visual presentation makes the decision tree model very easy to understand and assimilate. As a result, the decision tree has become a very popular data mining technique. Decision trees are most commonly used for classification. The decision tree method

encompasses a number of specific algorithms, including ID3, C4.5 and C5 (Brand and Gerritsen, 1998).

Decision trees graphically display the relationships found in data. Most products also translate the tree-to-text rules. The training process that creates the decision tree is usually called induction.

The top node is called the root node. A decision tree grows from the root node, so you can think of the tree as growing upside down, splitting the data at each level to form new nodes. The resulting tree comprises many nodes connected by branches. Nodes that are at the end of branches are called leaf nodes. The lower nodes are children of the node which is above these nodes, is named parent node. There are two phases, when a decision tree is created. They are growing phase and pruning phase.

Growing phase

The growing phase is an iterative process that involves splitting the data into progressively smaller subsets. Each of iterations considers the data in only one node. The first iteration considers the root node that contains all the data. Subsequent iterations work on derivative nodes that will contain subsets of the data.

One important characteristic of the tree splitting algorithm is greedy. Greedy algorithms make decisions locally rather than globally. When deciding on a split at a particular node, a greedy algorithm does not look forward in the tree to see if another decision would produce a better overall result. Once a node is split, the same process is performed on the new nodes, each of which contains a subset of the data in the parent node. The variables are analyzed and the best split is chosen. This process is repeated until only nodes where no splits should be made remain.

Pruning Trees

After a tree grows, an analyst must explore the model. Exploring the tree model may reveal nodes or sub trees that are undesirable because of over fitting, or may contain rules that the domain expert feels are inappropriate. Pruning is a common technique used to make a tree more general. Pruning removes splits and the sub trees created by them. In some implementations, pruning is controlled by user configurable parameters that cause splits to be pruned because, for example, the computed difference between the resulting nodes falls below a threshold and is insignificant. With such algorithms, users will want to experiment to see which pruning rule parameters result in a tree that predicts best on a test dataset. Algorithms that build trees to maximum depth will automatically invoke pruning. In some products users also have the ability to prune the tree interactively.

In data mining, the ID3, C4.5 and C5 are the available algorithms using the decision tree approach. The C4.5 and C5 are extension of ID3 that accounts for unavailable values, continuous attribute value ranges, pruning of decision trees and rule derivation. The C4.5 algorithm was used in the study. The algorithm considers all the possible tests that can split the data set and selects a test that gives the best information gain. For each discrete attribute, one test with outcomes as many as the number of distinct values of the attribute is considered. For each continuous attribute, binary tests

involving every distinct values of the attribute are considered. In order to gather the entropy gain of all these binary tests efficiently, the training data set belonging to the node in consideration is sorted for the values of the continuous attribute and the entropy gains of the binary cut based on each distinct values are calculated in one scan of the sorted data. This process is repeated for each of continuous attributes.

3.3.3 Theoretical concepts

The mathematic theory used by the C4.5 algorithm is information gain.

Entropy and information gain

If there are n equally probable possible messages, then the probability p of each is 1/n and the information conveyed by a message is -log(p) = log(n). That is, if there are 16 messages, then log(16) = 4 and we need 4 bits to identify each message.

In general, if we are given a probability distribution P = (p1, p2, .., pn) then the Information conveyed by this distribution, also called the Entropy of P, is:

$$I(P) = -(p1*\log(p1) + p2*\log(p2) + .. + pn*\log(pn)) \quad (3.1)$$

For example, if P is (0.5, 0.5) then I(P) is 1, if P is (0.67, 0.33) then I(P) is 0.92, if P is (1, 0) then I(P) is 0. [Note that the more uniform is the probability distribution, the greater is its information.]

If a set T of records is partitioned into disjoint exhaustive classes C1, C2, .., Ck on the basis of the value of the categorical attribute, then the information needed to identify the class of an element of T is Info(T) = I(P), where P is the probability distribution of the partition (C1, C2, .., Ck):

$$P = (|C1|/|T|, |C2|/|T|, ..., |Ck|/|T|) \quad (3.2)$$

If we first partition T on the basis of the value of a non-categorical attribute X into sets T1, T2, .., Tn then the information needed to identify the class of an element of T becomes the weighted average of the information needed to identify the class of an element of Ti, i.e. the weighted average of Info(Ti):

$$Info(X,T) = \text{Sum for i from 1 to n of } \frac{|Ti|}{|T|} * Info(Ti) \quad (3.3)$$

Consider the quantity Gain(X,T) defined as

$$Gain(X,T) = Info(T) - Info(X,T) \quad (3.4)$$

This represents the difference between the information needed to identify an element of T and the information needed to identify an element of T after the value of attribute X has been obtained, that is, this is the gain in information due to attribute X.

We can use the notion of gain to rank attributes and to build decision trees where at each node is located the attribute with greatest gain among the attributes not yet considered in the path from the root. (Joshi, K., 1997).

Confidence

Confidence of rule "B given A" is a measure of how much more likely it is that B occurs when A has occurred. It is expressed as a percentage, with 100% meaning B always occurs if A has occurred. Statisticians refer to this as the conditional probability of B given A (Two Crows Corporation, 2004).

Support

Support of rule “B give A” is a measure of how much more likely it is that A occurs. It is expressed as a percentage. With 100% meaning A always occurs in all cases.

Class variable

Class variable is the variable whose values are to be modeled and predicted by other variables. It is analogous to the dependent variable in linear regression. There must be one and only one target variable in a decision tree analysis (DTREG, 2005).

Predictor variable

Predictor variable is a variable whose values will be used to predict the value of the class variable. It is analogous to the independent variables in linear regression. There must be at least one predictor variable specified for decision tree analysis; there may be many predictor variables. If more than one predictor variable is specified, decision tree approach will determine how the predictor variables can be combined to best predict the values of the class variable (DTREG, 2005).

Confusion matrix

A confusion matrix is a visualization tool typically used in supervised learning. Each column of the matrix represents the instances in a predicted class, while each row represents the instances in an actual class. One benefit of a confusion matrix is that it is easy to see if the system is confusing two classes (WIKIPEDIA, 2005).

3.4 Validation

A model built by any technique needs to be validated, which means calculating an error rate based on data independent of that used to estimate the model. This exercise gives a statistically valid estimate of the true error rate that the modeling procedure produces. It does not guarantee that the model is correct in any way. It simply says that if the same technique was used on a succession of databases to build a model, the average error rate would be close to the one obtained this way.

The most basic testing method is called simple validation. To carry this out, a percentage of database is set aside and is not used in any way in the model building and estimation. This percentage is usually small, perhaps 5% to 20%, but there is no fixed rule (Two Crows Corporation, 2004). For all the future calculations to be correct, the division of the data into two groups must be random.

After building and estimating the model on the main body of the data, the model is used to predict the classes of the validation database. The number of correct and incorrect classifications can counted up and an error rate is calculated.

CHAPTER 4
Materials and Methods

The chapter introduces the materials used in this study and the procedures carried out in the study. It includes how to pre-process the data, and how to create the decision trees.

4.1 Study Area

There are 193 countries and regions in my study area. The Figure 4.1 illustrates the distribution of GDCs in the whole word. Four classes of countries were used to decide the status of GDCs in 2002:

1. Clearinghouse: the counties have the GDCs, and the GDCs have the whole functions described in the section 2.6.4.
2. Product portal: the counties have the GDCs that only provide the data portal and simple information about the data.
3. Project: the counties have the projects to implement the GDCs.
4. None: the countries do not have GDCs and projects.

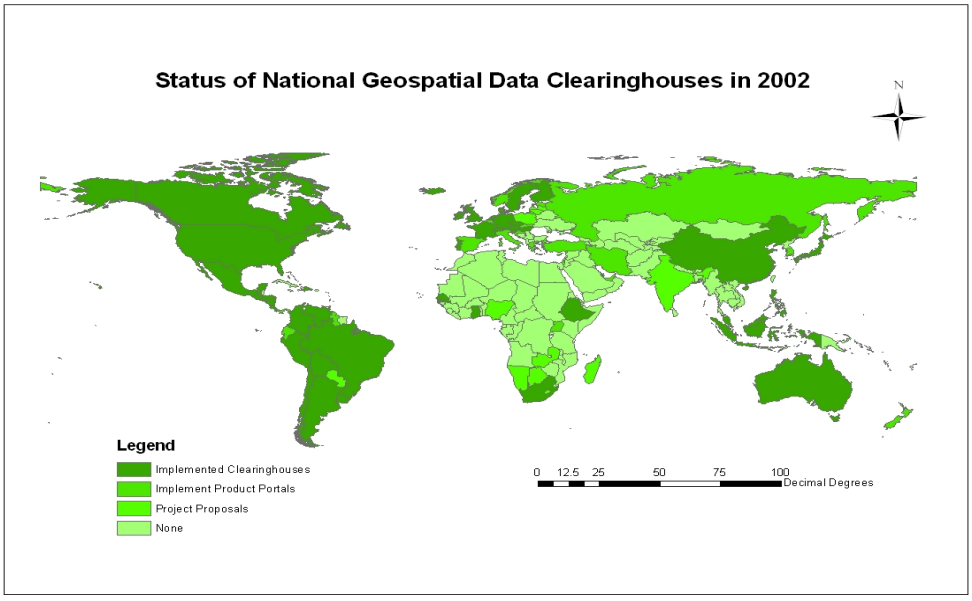


Figure 4.1 The distribution of GDCs

Data About GDCs

In 2002, 67 countries had GDCs in the world, in which 48 countries had clearinghouses and 19 countries had product portals. There are 22 attributes to describe the characteristics of GDCs. They are history, people, data, access network, standard and policy that are the six main aspects of NSDIs and GDCs.

Data about society

10 societal aspects (tables) are described by 252 attributes in the dataset. There are 27 attributes about demography, 76 attributes about economy, 11 attributes about legal, 44 attributes about technology, 5 attributes about culture, 17 attributes about network readiness, 6 attributes about institution, 15 attributes about education, 38 attributes about environment and 13 attributes about health. These data had been collected for 193 that had clearinghouses, product portals, project or none. These data were the latest data till the end of 2002.

4.2 Overview of procedures

The procedures of study mainly included data preprocessing, setting up the decision tree to determine the existence of GDCs, setting up decision trees to determine the success of GDCs and using the decision trees to predict the development of GDCs. Figure 4.2 shows the detail steps of this study work.

Data pre-processing

The dataset used in this study was about the various aspects of society and does not contain redundant attributes. Values of every record had been corrected. But some values were missing when the data were collected. In order to run the decision tree software well and reduce the bias produced by missing values, four methods that have introduced in section 3.2.1 were used individually to find the best method to process the missing values.

Creating the decision tree to determine the existence of GDCs

The attributes of societal aspects were the predictor variables. The class variable was whether the GDC exists. The class variable came from the status of GDCs (clearinghouse, product portal, project and none).

Creating the decision tree to determine the success of GDCs

The attributes of societal aspects were the predictor variables in the decision trees. The class variable were success and failure. How to confirm class variable? That means which GDCs are successful and which GDCs are failure. The answers will quite different depended on the individual experience and perspective. There are not uniform standards for GDCs. So the successful GDCs in this study were the relative group against the fail one. In the study, two methods were used to look for the successful GDCs. One is named using expert knowledge method (EK method). The EK method is to use the expert knowledge to partition the GDCs into two groups (success and failure). Each group is represented by a value. For example: ‘1’ is successful group and ‘2’ is the failure group. The other method is median partition method (MP method). The MP method is using median values of the attributes of GDCs to divide the GDCs into two groups (success and failure). The study range included the GDCs whose statuses are clearinghouse or product portal.

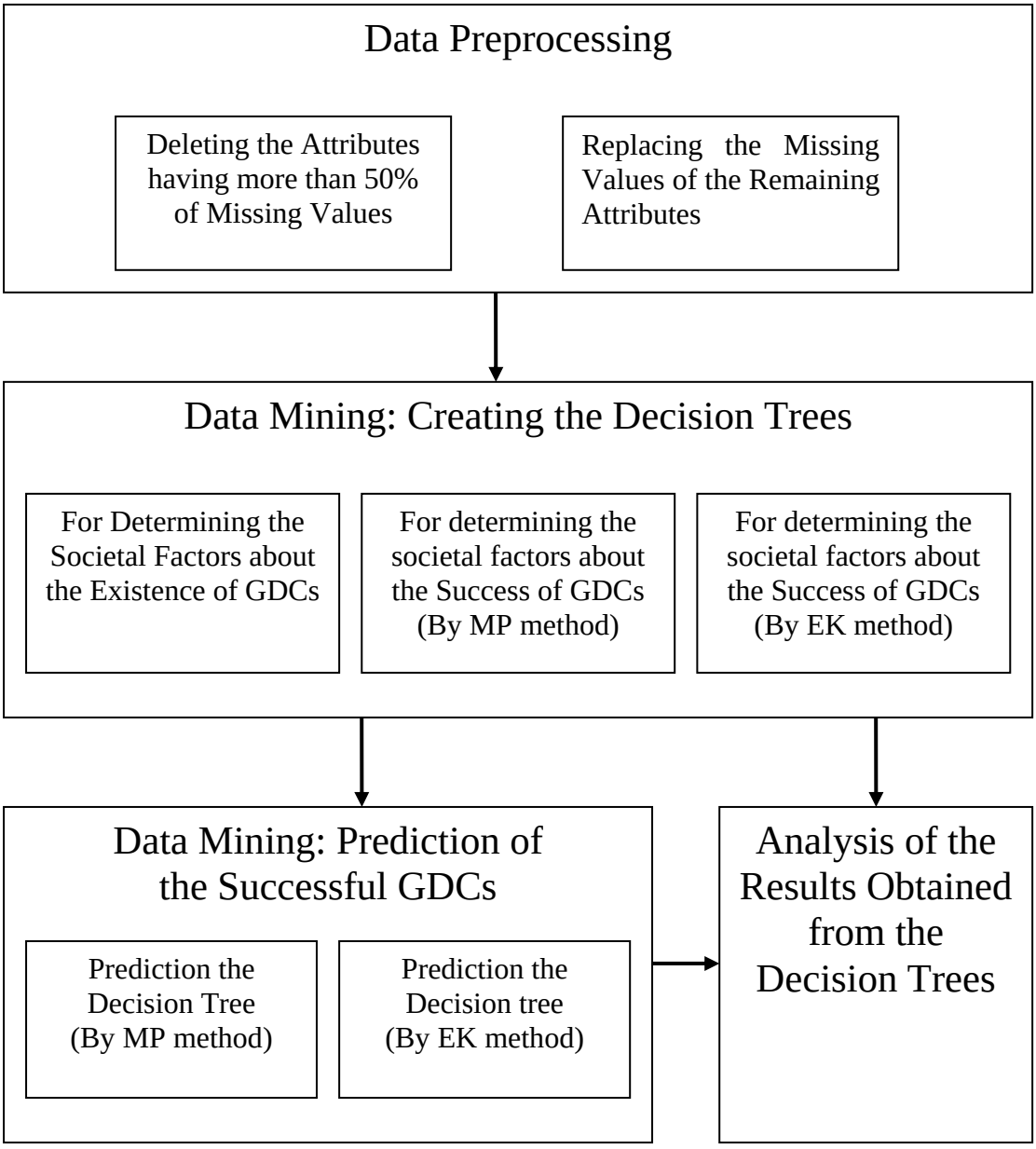


Figure 4.2 Working flow of research

Prediction the development of GDCs

There are 126 countries that do not have GDCs in 2002. The models created for the assessing the success of GDC will be used to predict these countries to look for whether they will be successful if they are created now.

4.3 Data preprocessing

Because the attributes of GDCs are the fundamental variables to describe the GDCs, and there are only 67 records (countries) in the table. If too many missing values are in an attribute, the attribute will give a great error or bias for the description of DGCs. So the attribute will lose the ability of expressing the real situation of GDCs. Therefore a criterion is set up: if the attribute has more than 20 percent of missing values, it will be ignored. Figure 4.3 shows the different percentage of missing values of the attributes of GDCs.

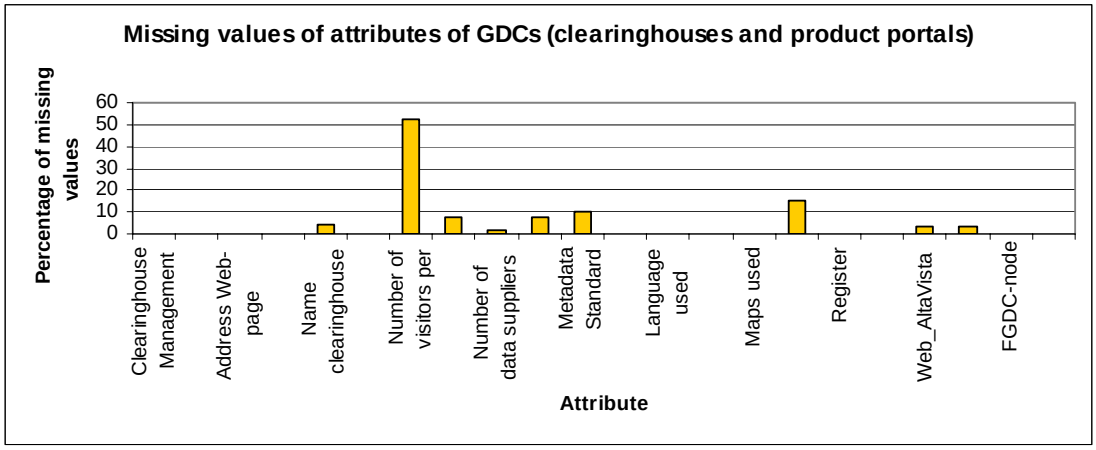


Figure 4.3 The percentage of missing values of the attributes of GDCs

Figure 4.4 shows the percentage of missing values of different societal aspects. Figure 4.5 shows the percentage of missing values per countries that belong to the different classes. From the Figure 4.4 and 4.5, we can see there are a lot of missing values in the 252 attributes of society. How to avoid or reduce the influence of missing values becomes the biggest problem in the study. In order to let more attributes have opportunities to participate in the creating the decision tree, and let the useful information more then the noise in the attribute, the attribute is ignored if it has more than 50 percent of missing values

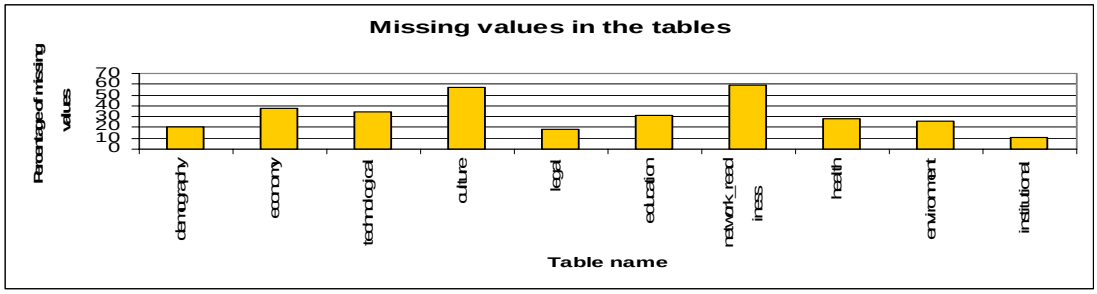


Figure 4.4 The percentage of missing values of different societal aspects

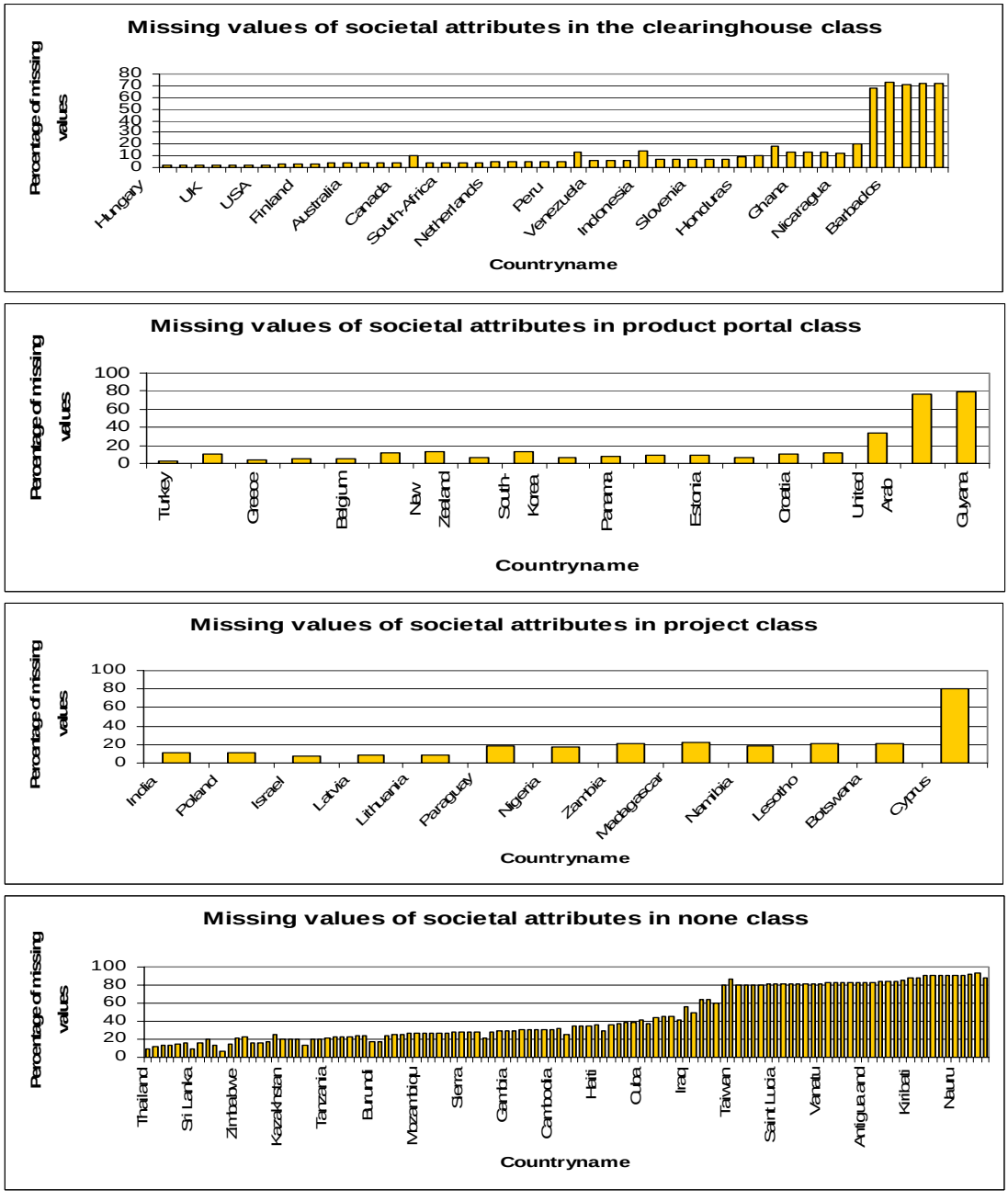


Figure 4.5 The percentage of missing values per countries

Although eliminating the attributes that have more than 50 percent of missing values, we have still left with approximately 25% of the data that are missing values in the dataset (see Appendix A). In order to create the decision trees by these incomplete data, the missing values needed to be processed. As be said in the literature review, the methods (mean, median, field mean, field median) can be used to replace the missing values in the data set after the selecting the data.

Mean Method: The mean of the attribute is used to replace the missing values in this attribute, and the procedure is done in all dataset. But there is a limit for nominal and ordinal values. Because these values do not have means, the majority values will be used to replace the missing values.

Median Method: It is using the median of the attribute to replace the missing values in this attribute, and this procedure is done in all dataset. Because the nominal values have not the medians, the majority values will be used to replace the missing values.

Field Mean Method: Field mean is feature and class mean. Firstly, The attribute was divided into different groups by the classes (clearinghouse, product portal, project and no). Secondly, the mean of every group is used to replace the missing values in this group. And the procedure was done for the whole dataset. Because the nominal and ordinal values do not have mean, the feature and class's majority values are used to replace the missing values.

Field Median Method: It is the same as the field mean method. Because the ordinal values do not have the field median, the nominal missing values are replaced by the feature and class majority values.

Which method is the best one for processing the data? Because the range and scale of the data are very different and the information gain only concerns the information in the attributes, and does not concern the distribution of the data, such the normal criteria as standard deviation, T test and F test are unuseful to assess which method is the best. Therefore the least percentage of misclassification errors of the decision trees is the criterion used to choose the best method. The four methods are used to process every table that is about each aspect of society. The attributes of every table processed by each method were the predictors to create the decision tree, while the four class of countries will be the class variable. Comparing the percentage of misclassification errors of the decision trees created with same table and different data preprocessing method, the method that has the least percentage of misclassification errors is the choice to process the missing values in this table. In this procedure, totally 40 decision trees were created to obtain the least percentage of misclassification errors. Appendix J shows the example of these decision trees including decision tree, results, node view and rules section.

4.4 Configuration of parameters of the decision tree software

In order to let the software work well, the decision tree software needs to configure the parameters before input data to compute. These are some parameters needed to set necessarily:

Leaf node criteria

While growing the tree, whether to stop splitting a node and declare the node as a leaf node will be determined by the following criteria:

- Minimum node size means to stop splitting a node if number of records in that node is less or equal than a critical value. The critical value was defined 5 in the study.
- Maximum purity means to stop splitting a node if its purity is larger or equal than a critical value. The critical value was defined 100% in the study.
- Maximum depth means to stop splitting a node if its depth is larger or equal than a critical. The critical value was defined 10 in the study (Root node has Depth 1. Any node's depth is it's parent's depth + 1).

Criteria of rule cleaning option

- Minimum Confidence means to create the rule if the classification confidence larger or equal than a critical value. The critical value was defined 50% in the study. If the confidence of a rule is less than 50 percent, it means that more than 50% errors in the rule. The rule is inappropriate.

In this study, the Confusion Matrix was used to represent the accuracy of decision tree.

4.5 Creating decision tree for the existence of GDCs

When the data preprocessing was finished, data mining about the existence of GDCs was done. The 193 countries and regions were partitioned into two groups randomly. 70 percent of countries and regions were assigned to the training group and the other 30 percent were testing group.

Class variable

The statuses of GDCs were used as class variable. The countries, which had the clearinghouses or product portals, became the HAVING class. The countries, which had the projects and had not anything about GDC, became the NO class.

Predictor variable

There were totally two steps in order to set up the decision trees about the existence GDC. One step was selecting the predictor variables. The other was setting up the final decision tree.

Selecting the predictor variables: totally 252 attributes took part in the decision trees, but the maximum number of predictor variables allowed by the decision tree software is 49. The 252 attributes had to be divided into a few parts to make the decision trees

firstly. The attributes were selected according to the rules made by these decision trees to make the final decision tree. According to the reality, there were 10 aspects of society in the dataset, so the 10 decision trees were created for the 10 aspects individually. The attributes that took part in the final decision tree were selected from the 10 decision trees. The criteria of selecting the predictor variables for the final decision tree were the predictor variables in the rules created by the 10 decision trees. Appendix K shows the example of these decision trees including decision tree, results, node view and rules section.

The other step was setting up the final decision tree. Appendix L shows the decision tree including decision tree, results, node view and rules section.

4.6 Creating the decision trees for success of GDCs

In order to create the decision trees for success of GDCs, the 67 countries were partitioned into two groups randomly. 54 countries were assigned to the training group and the other 13 countries were testing group. The two classes were defined as being success and failure. There were two methods used to confirm the class variables.

The MP method

The hypothesis was: Since the weightiness of the attributes of GDC was not known, I supposed the attributes had the same weightiness. 16 attributes were used to assess the GDCs in the 21 attributes of GDCs. Two groups were separated by the median value for every attribute. The weight 2 was given to the countries that have the higher value than median. The weight 1 was given to others. That means to replace all the values of every attribute to the weight 1 or 2 in term of the median of every attribute. For example, the median of attribute “Number of Datasets” was 62. If the value of “Number of Datasets” of a GDC was higher than 62, the weight 2 was given to “Number of Datasets” of the GDC. If the value of “Number of Datasets” of a GDC was lower than 62, the weight 1 was given to “Number of Datasets” of the GDC. The specific criteria about how to allocate the weight to every value of every attribute were given in the Appendix B. After the work of replacing the original values by weight, an attribute was created for the sum of the weights of attributes of every GDC.

The first quartile, median and the third quartile of the 67 sums of weights were used individual to divide the 67 GDCs into two groups. For example, when the first quartile was used to split the GDCs into two groups, the GDCs, of which the values of sums of weights were larger than the first quartile, became a group. The GDCs were SUCCESS group. The others were FAILURE group. This was a scenario. So totally, The first quartile, median and the third quartile generated three scenarios. Every scenario as the class variables took part in the creating the decision trees.

The EK method

The hypothesis was: the expert knowledge was used to weight the attributes of GDCs. Because the GDC has very close relationship with the five components of SDI (people, data, access network, policy and standards), according to the expert knowledge, the 15 attributes of GDCs were grouped into six aspects: history, people, data, access network, policy and standards. Every aspect received different weight (The criteria of weight of every aspect are shown in Appendix C). Every attribute was also received different weight within its aspect (The criteria of weight of every attribute are shown in the Appendix D). All the expert knowledge came from Ir. Joep Crompvoets and the survey, which had been done on the SDI course of GIS department Wageningen University (The question of survey and the results are shown in the Appendix E). According to the criteria in the appendix D, the weights of every attribute replaced the values of the attribute. The sums of all weights of groups multiplying weights of attributes were obtained for every GDC. 67 sums of weights were obtained for 67 GDCs.

The first quartile, median and the third quartile of the 67 sums of weights were used individual to divide the 67 GDCs into two groups. The approach was the same as it used in the MP method. So totally, the first quartile, median and the third quartile generated three scenarios. Every scenario as the class variables took part in the creating the decision trees.

Class variable

The six scenarios obtained from the MP and EK methods were used as class variable. Every scenario became the class variables individually.

Predictor variable

There were totally two steps in order to set up the decision trees about the success of GDCs. One step was selecting the predictor variables. The selecting of predictor variables depended on the every scenario individually. That means different scenario selected different predictor variables from the 252 societal attributes. The approaches to select the predictor variables in the study of success of GDCs were the same as them in the study of existence of GDCs. How to selecte the predictor variables had been described in the section 4.5. In this procedure, totally 60 decision trees were created to obtain the attributes to set up the final decision trees. Appendix M shows the example of these decision trees including decision tree, results, node view and rules section. The other step was setting up the final decision tree.

When the six decision trees were created, a decision tree, which had the least percentage of misclassification errors in the three scenarios of MP method, was considered as the decision tree for the success of GDCs. a decision tree will be selected from the three scenarios of EK method too. Appendix N shows the example of decision trees created by MP method including decision tree, results, node view and rules section. Appendix O shows the example of decision trees created by EK method including decision tree, results, node view and rules section.

4.7 Prediction

Decision trees have obvious value as both predictive and descriptive models. We have seen that prediction can be done on a case-by-case basis by navigating the tree. More often, prediction is accomplished by processing multiple new cases through the decision tree or rule set automatically and generating an output file with the predicted value or class appended to the record for each case.

In order to determine the predicted value of a case, the prediction process begins with the root node, and then decides whether to go into the left or right child node based on the value of the splitting variable. This process Continues with using the splitting variable for successive child nodes until it reaches a terminal, leaf node. The value of the class variable shown in the leaf node is the predicted value of the case.

126 countries and regions did not have the GDCs in the world in 2002. How about the GDCs are, if these countries and regions create the GDCs. The two final decision trees chosen form the six scenarios were as the model to analyze the society's attributes of 126 countries and regions and gave the results of prediction

Using the two decision trees as the model individually, data of 126 countries as the predicted data were inputted into the decision tree software. The prediction function of the software outputted the prediction results.

CHAPTER 5
Results and discussion

The chapter describes all the results obtained from every procedure. All the results are introduced in terms of the order of the study process from selecting the data to the prediction of development of GDCs.

5.1 Data preprocessing

Attributes of GDC

From the total of 22 attributes, the attribute “number of visitors per month” had more than 20 percent of missing values, so it was ignored. Therefore, 21 attributes were used in the study.

Attributes of society

In the 10 aspects (tables), 55 attributes were ignored because the percentage of missing values was larger than 50 %. (There was 1 in the table of demography, 15 in economy, 11 in technology, 2 in legal, 5 in culture, 2 in education, 18 in network readiness and 1 in health). From the Appendix A, we can see the percentage of missing values of all attributes. In the case of culture and network readiness were they larger than 50%, and the most missing values concentrated on the class “no”. So they did not take part in the construction of decision tree about the existence of GDC. However they had very low percentage of missing values on the class “clearinghouse” and “product portal”. Therefore they were included in the construction of the decision tree for analyzing the success of GDC.

Dealing with missing values

In order to look for the best method to process the missing data, the decision trees were created for every aspect of society (demography, economy, education, legal, institution, technology, Health and environment). The Table 5.1 shows the percentage of misclassification errors of every decision.

When we look at the first row on the table. It is about the demography. We can see the percentage of misclassification errors is 2.59% for the mean method, 3.11% for the median, 1.55% for the field mean, 2.07% for the field median. Since the percentage of misclassification errors of field mean method was the least, the field mean method was considered as the best method for the attributes of demography aspect. So we can get from the table: the field mean method was selected for economy, education, environment, and health to process the missing values. The field median method was used for legal, technology and institutional. From the results of selecting the method, we can see that the field mean method and field median were better than the mean and median method

Table 5.1 The percentage of misclassification errors found after classification

	Mean	Median	Field mean	Field median
Demography	2.59	3.11	1.55	2.07
Education	2.59	4.66	2.59	3.63
Environment	2.07	2.07	1.55	2.07
Health	3.63	5.7	0.52	1.04
Legal	10.36	8.29	5.18	4.15
Network readiness				
Culture				
Technology	3.11	1.04	1.55	0
Institutional	7.25	7.25	5.7	5.18
Economy	1.55	2.07	0.52	0.52

There are two empty rows in the table. The reason is that the percentage of the number of missing values was larger than 50% for every attribute in culture and network readiness aspects. Because the two aspects were only used to analyze the success of GDCs, although the methods of processing missing values were the same as the other aspect they were only about the class “clearinghouse” and “product portal”. The Table 5.2 shows the percentage of misclassification errors of the four methods about the two aspects. The field mean method was used for network readiness and culture to process the missing values.

Table 5.2 The percentage of misclassification errors about the Network readiness and Culture.

	Mean	Median	Field mean	Field median
Network	6.22	5.7	2.07	2.59
Culture	6.22	5.7	0.52	1.04

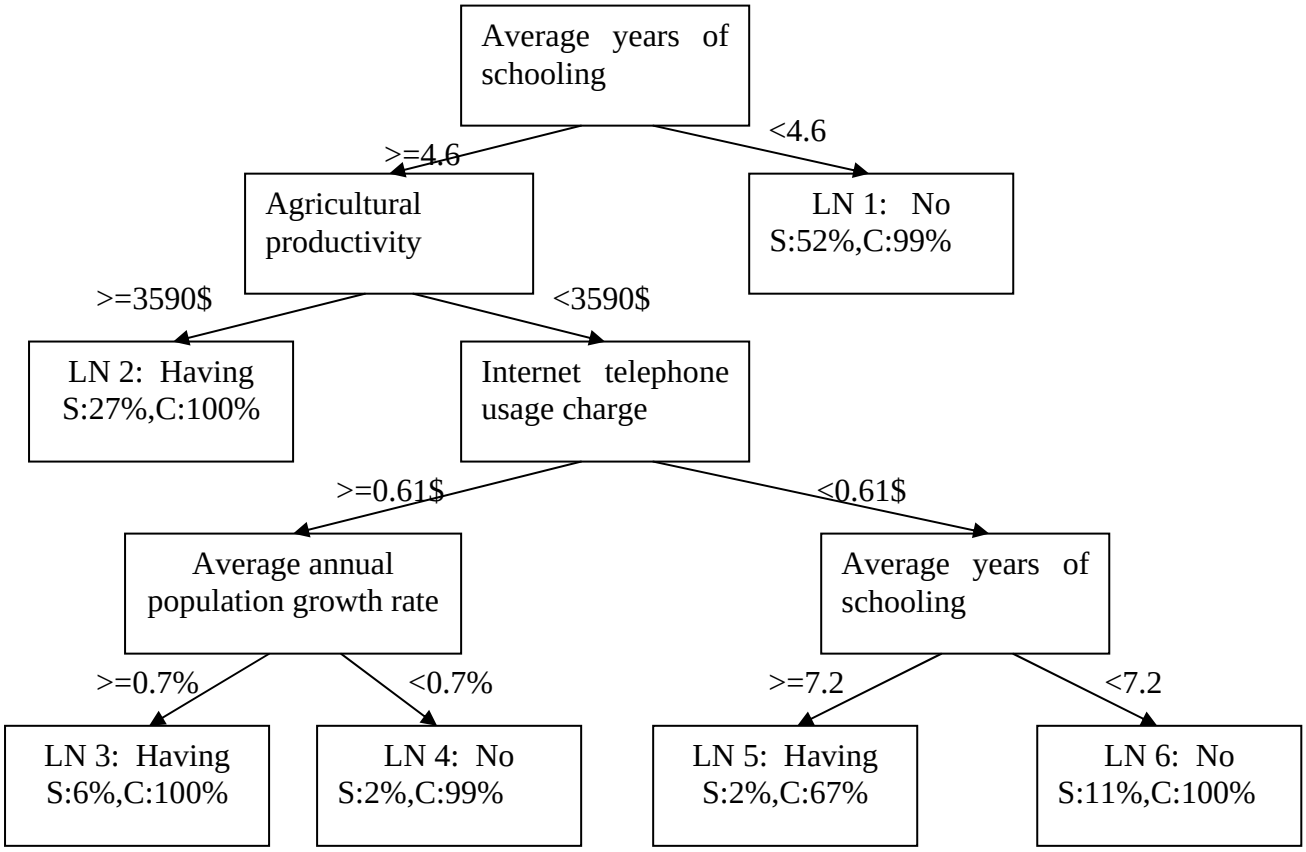
5.2 Decision tree about the existence of GDCs

Result of selecting the attributes

In summary, 40 attributes were selected from the 10 aspects of society (Appendix F) in such a way that 5 attributes belong to demography, 9 attributes belong to economy, 4 attributes belong to technology, 3 attributes belong to legal, 3 attributes belong to institution, 5 attributes belong to education, 4 attributes belong to health and 7 attributes belong to environment.

Decision tree

The 193 observations were divided into two groups in such a way that 142 (70%) observations were training group and 51 (30%) observations were testing group. 40 attributes of society as the predictors took part in the creation of the final decision tree. Figure 5.1 shows the final decision tree about the existence of GDCs. There were 11 nodes in the decision tree: 6 nodes were the leaf nodes. The percentage of misclassification errors of training data was 1.41%. The percentage of misclassification errors of testing data was 9.8%. The confusion matrices (Figure 5.2) show the misclassification errors of decision tree.



Note:

LN: Leaf node S: Percentage of support C: Percentage of confidence

Figure 5.1 The final decision tree about the existence of GDCs

In Figure 5.1, 105 observations (countries) were allocated in the leaf node 1; 49 observations were allocated in the leaf node 2; 11 observations were allocated in the leaf node 3; 3 observations were allocated in the leaf node 4; 4 observations were allocated in the leaf node 5 and 21 observations were allocated in the leaf node 6.

It is clear that the decision tree and every leaf node have very high confidence values. The least one was 67%. Because the total number of the records was 193, the number

of records was too small, and this may be a reason for the very high confidence. But on the other hand, it has provided us with very high accuracy and the attributes in the decision tree, Which really had important effects on the existence of the GDCs. Therefore the decision tree is suitable to determine whether the GDCs exist or not.

Confusion Matrix Total

Training Data

True Class \ Predicted Class	Having	No	Total
Having	49	1	50
No	1	91	92
	50	92	142

Test Data

True Class \ Predicted Class	Having	No	Total
Having	13	4	17
No	1	33	34
	14	37	51

Figure 5.2 Confusion Matrices of decision tree about existence of GDCs

From Figure 5.1, we can get 4 factors that have an impact on the existence of GDCs.
1. Average years of schooling, total, 2000
2. Agricultural productivity, Agriculture, value added per worker, $, 1999-2001
3. Internet, Monthly off-peak access charges, Telephone usage charge, $, 2001
4.Average annual population growth rate, %, 2001

These factors belong to the educational, economical, demographical and technological aspects. So the four aspects have bigger effects on the existence or no existence of GDCs than the other aspects.

How about these factors affect the GDCs specifically? Some statements had been made according to the decision tree. Four of them had very high confidence. The statement: “if Average years of schooling <4.6 then the country does not have the GDC” had 96% confidence and 54% support of 193 countries. The statement: “if Average years of schooling >=4.6 then the country has the GDC” had 71% confidence and 46% support of 193 countries. The statement: “If Agricultural productivity, Agriculture, value added per worker>=3590$ then the country has the GDC” had 89% confidence and 28% support. The statement: “If Agricultural productivity, Agriculture, value added per worker <3590$ then the country does not the GDC” had 86% confidence and 72% support. Compared with the statements about the other factors, we can say these two factors were the most important for the existence of GDCs. Therefore the education and economy were the most important aspects to affect the existence of GDCs. In order to illuminate the statement, Figure 5.3 shows the comparison among the existence part of status of GDCs and the positive and negative impact of the four factors. Figure 5.4 shows the no existence part of status of GDCs and the positive and negative impact of the four factors.

From the confidence and support of the four statement, Figure 5.3 and Figure5.4, it is almost true that the higher the values of “Agricultural productivity, Agriculture, value added per worker” and “Average years of schooling” are, the more possibility of existence of GDCs. for the other factors, we did not obtain this kind of relationships.

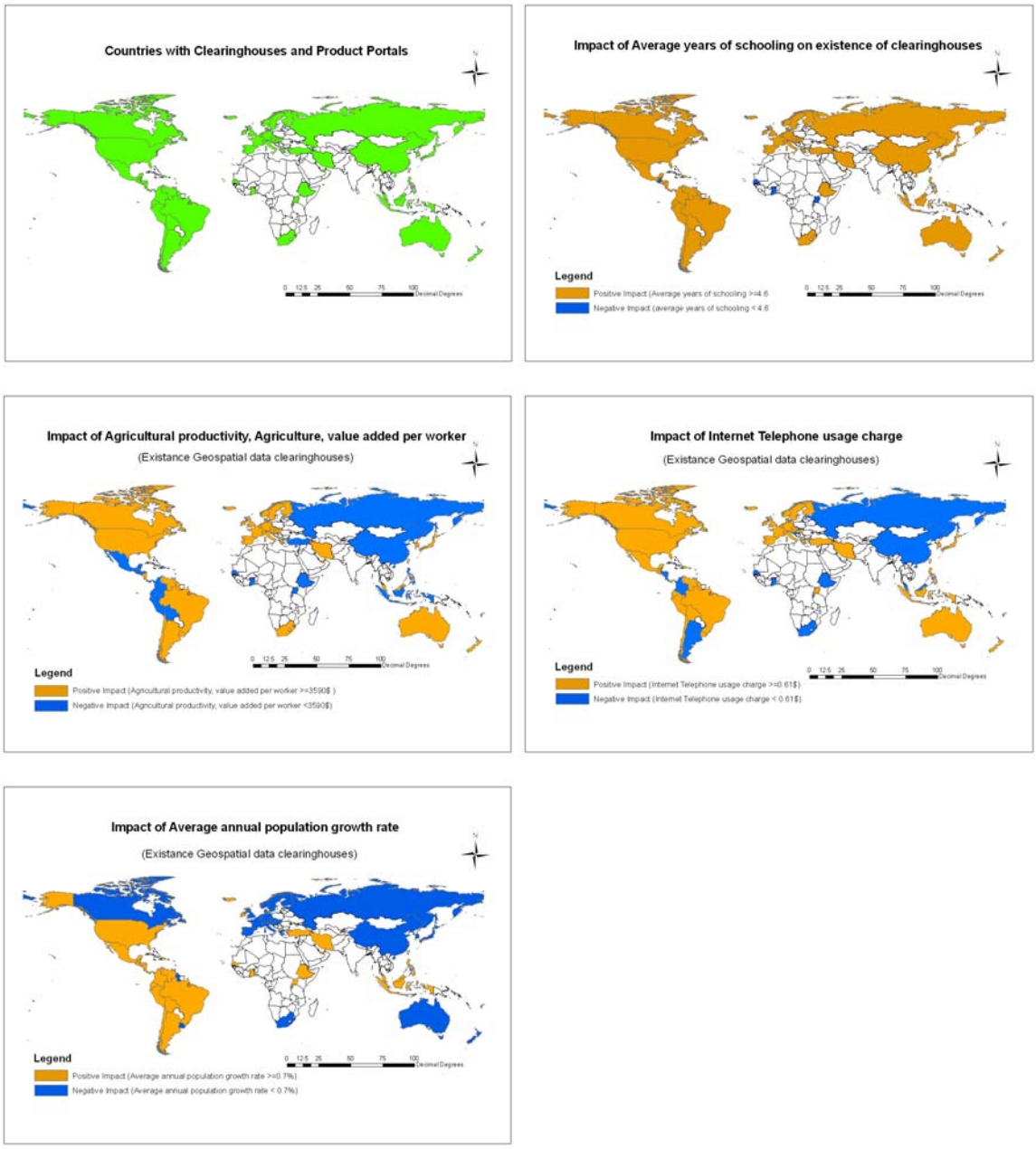


Figure 5.3 Comparison among the existence part of status of GDCs and effect of the four factors.

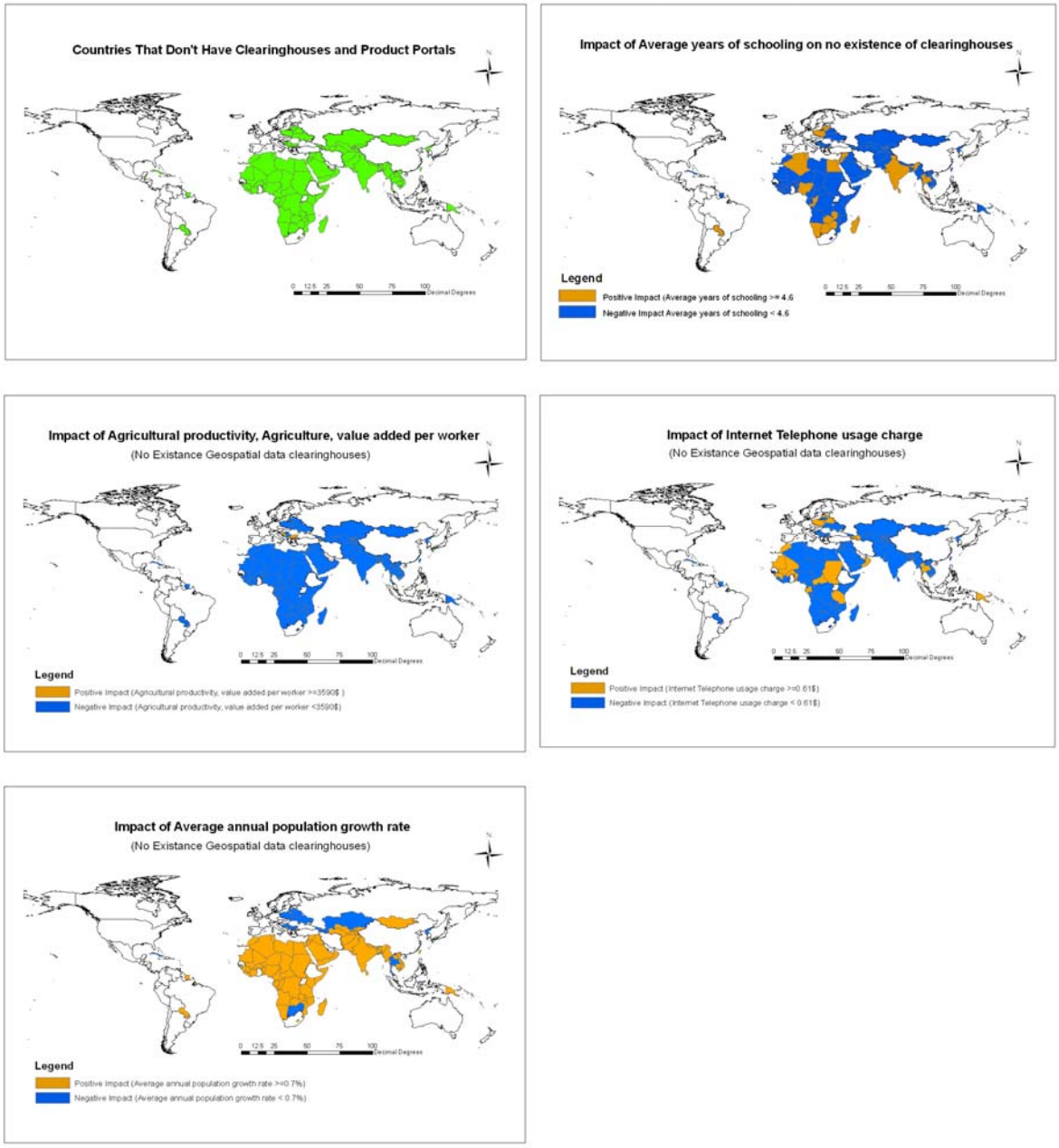


Figure 5.4 Comparison among the existence part of status of GDCs and effect of the four factors.

5.3 Results of the decision trees about the success of GDCs using the MP method

Creating and selecting the scenarios

The sums of weights of every GDC were obtained for MP method. The Appendix G shows these results. For the MP method, the first quartile value was 20; the median is 22 and the third quartile is 24. As be said in the section 4.6, three scenarios were created with the first quartile, median and the third quartile as the split point. In order to look for the best scenarios, the decision trees had been created using every scenario as the class variables. The Median scenario was selected for MP method. The Table 5.3 shows the percentage of misclassification errors of different decision trees created for the three scenarios.

Table 5.3 The percentage of misclassification errors of decision trees for MP method.

	First quartile	Median	Third quartile
MP method	3.7	1.85	5.56

In summary, 48 attributes were selected from the 10 aspects of society (Appendix H) in such a way that 7 attributes belong to demography, 9 attributes belong to economy, 3 attributes belong to education, 5 attributes belong to environment, 3 attributes belong to health, 3 attributes belong to legal, 4 attributes belong to network readiness, 4 attributes belong to culture, 4 attributes belong to technology and 5 attributes belong to institution.

The 67 observations were divided into two groups randomly in such a way that 54 (80%) observations were training group and 13 (20%) observations were testing group. The 48 attributes of society as the predictors took part in the creation of the decision. There were 15 nodes in the decision tree: 8 nodes were the leaf nodes. The percentage of misclassification errors of training data was 1.85%. The percentage of misclassification errors of testing data was 23.08%. Figure 5.5 shows the confusion matrices. Though the confidence of testing data was lower than the confidence of training data, 76.9% of accuracy was very high for the testing. In the decision tree, 19 observations (countries) were allocated in the leaf node 1; 5 observations were allocated in the leaf node 2; 5 observations were allocated in the leaf node 3; 8 observations were allocated in the leaf node 4; 1 observation was allocated in the leaf node 5; 19 observations were allocated in the leaf node 6; 4 observations were allocated in leaf node 7 and 6 observations were allocated in leaf node 8. Figure 5.6 shows the final decision tree about success or failure of GDC with the MP method. The lowest confidence of the leaf nodes was 80%. The high confidences of the decision tree and the leaf nodes illuminated that the decision tree was suitable to determine whether the GDCs are successful or failure.

Confusion Matrix

Training Data

True Class	Predicted Class: Failure	Predicted Class: Success	
Failure	30	0	30
Success	1	23	24
	31	23	54

Test Data

True Class	Predicted Class: Failure	Predicted Class: Success	
Failure	7	2	9
Success	1	3	4
	8	5	13

Figure 5.5 Confusion Matrixes of decision tree about success of GDCs (MP method)

From Figure 5.6, we obtained 7 factors that have an impact on the success of GDC.
1. Health Expenditure, Total % of GDP, 1997-2000
2. Internet, Hosts Total, 2002
3. Median Age, years, 2002
4. Land use, Arable land, % of land area, 2000
5. Crude death rate (/1000people), 2001
6. Gross domestic product, % growth, 2000-2001
7. Rural population % (2001)

In order to obtain the most important factors, the statements about the 7 factors were created. I found that the statements about the "Health Expenditure, Total % of GDP" and "Rural population" were important. The statement: "if Health Expenditure, Total % of GDP>=7.1% then the GDC is successful" had 73% confidence and 45% support of 67 countries. The statement: "if Health Expenditure, Total % of GDP<7.1% then the GDC is failure" had 84% confidence and 55% support of 67 countries. The statement: "if Rural population<30% then the GDC is successful" had 60% confidence and 42% support of 67 countries. The statement: "if Rural population>=30% then the GDC is failure" had 72% confidence and 58% support of 67 countries. The four statements had the highest confidence than the others. In order to explain the accuracy of the statements and the effects of the 7 factors, Figure 5.7 shows the comparison among the successful part of status of DGCs determined by MP method and the positive and negative impact of the 7 factors on this part. Figure 5.8 shows the comparison among the failure part of status of DGCs determined by MP method and the positive and negative impact of the 7 factors. From the statements and the Figure 5.7 and 5.8, we can see that the factor "Health Expenditure, Total % of GDP" was higher than 7.1% for most of the successful GDCs and lower than 7.1% for most of failure GDCs. Rural population was lower than 30% for most successful GDCs and higher than 30% for most successful GDCs.

.

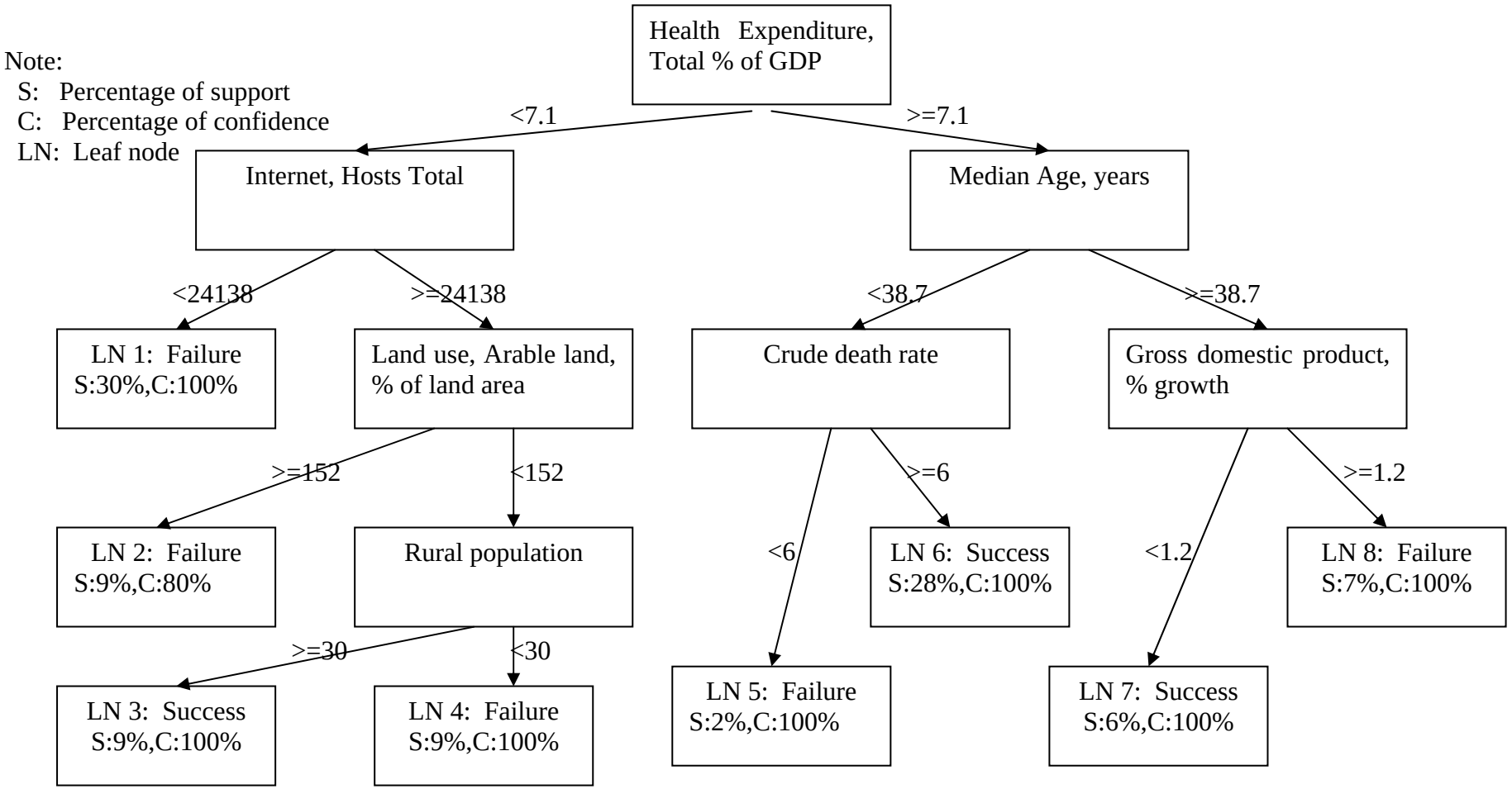


Figure 5.6 The final decision tree about success of GDCs with MP method

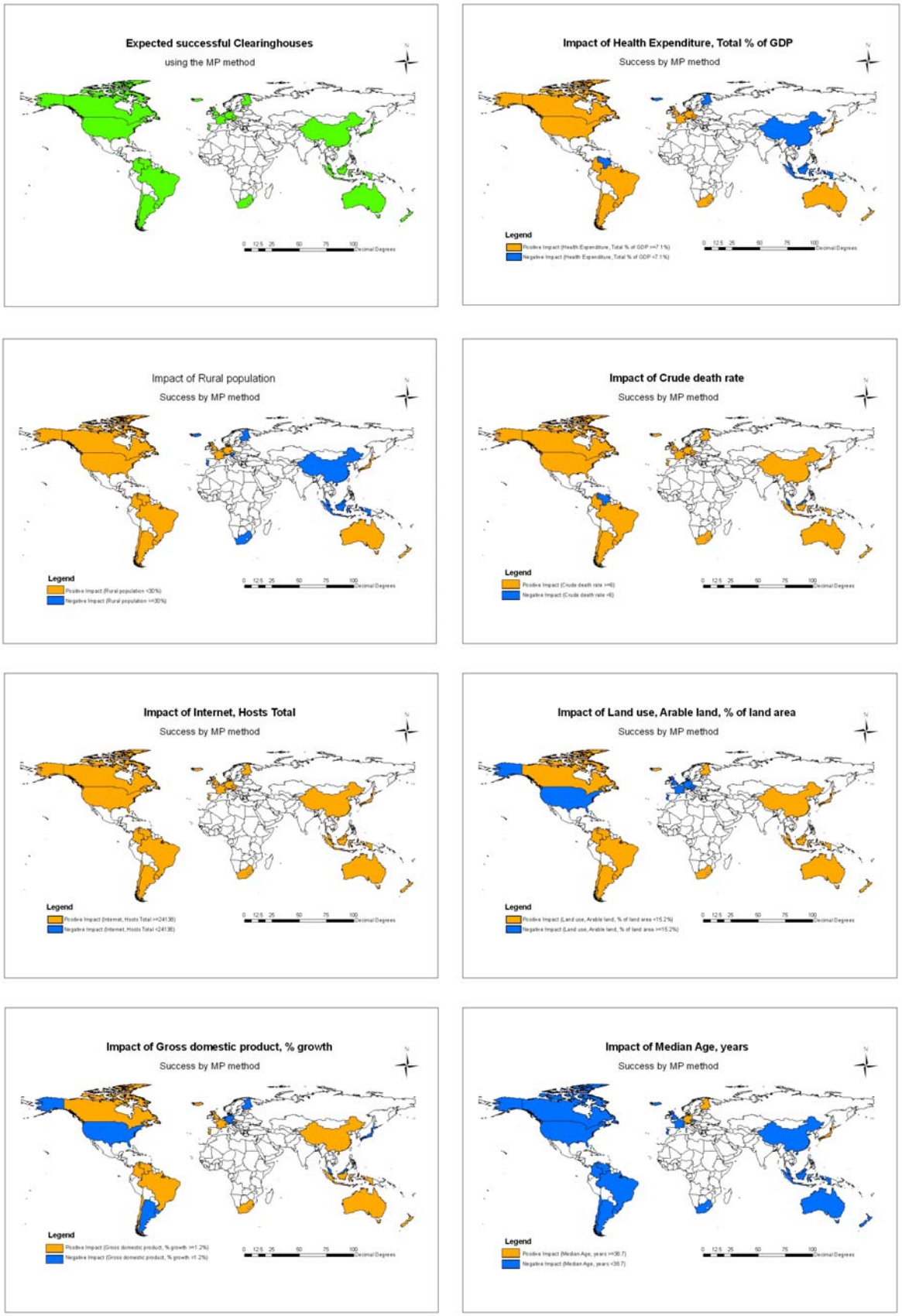


Figure 5.7 Comparison among the successul part of status of GDCs and effect of the 7 factors

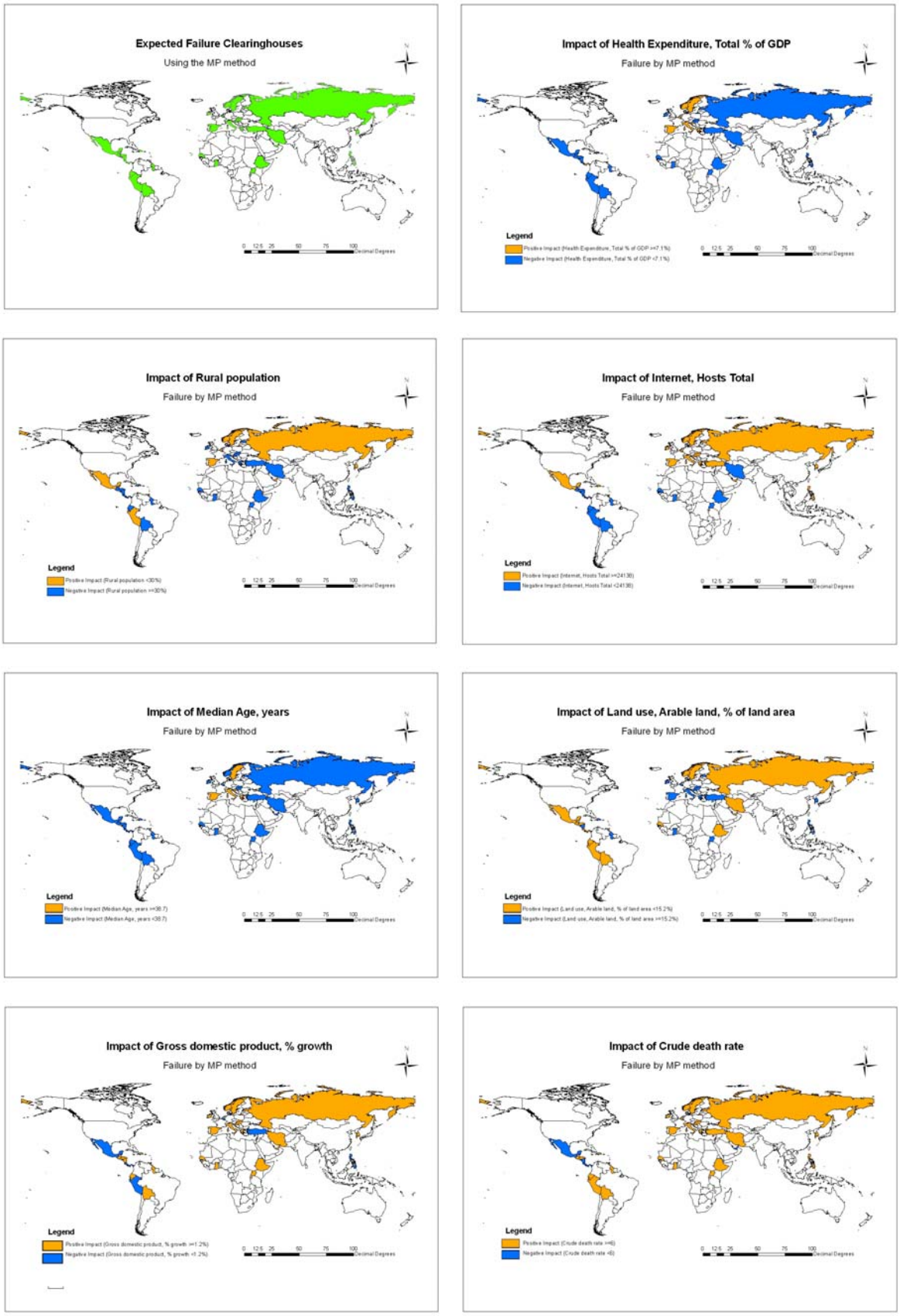


Figure 5.8 Comparison among the failure part of status of GDCs and effect of the 7 factors

The seven factors come from Health, demography, economy, environment and technology aspects of society. I wanted to explain something about "Health Expenditure, Total % of GDP". The total health expenditure is the sum of public and private health expenditure. It covers the provision of health service (preventive and curative), family planning activities, nutrition activities, and emergency aid designated for health. It is the financial spending in the health sector (World Bank, 2003). The GDP is the sum of value added by all resident producers plus any product taxes (less subsidies) not included in the valuation of output. It is the most important indicators of development of economy (World Bank, 2003). "Health expenditure Total,% of GDP" represents the relationship between the spending of health sector and development of economy very well. The "Health expenditure Total, % of GDP" is not only an attribute of health aspect, but also an attribute of the development of economy. So "Health expenditure Total, % of GDP" was considered as the attribute of both health and economy aspect. Therefore health, economy and demography were the most important aspects to determine whether the GDCs are success or failure.

5.4 Results of the decision trees about the success of GDCs using the EK method

The sums of weights of every GDC were obtained by the EK method. The Appendix G shows these results. For the EK method, the first quartile value is 28.5; the median is 37 and the third quartile is 46. As be said in the section 4.6, three scenarios were created with the first quartile, median and the third quartile as the split point. In order to look for the best scenarios, the decision trees had been created using every scenario as the class variables. The third quartile scenario was selected for EK method. The Table 5.4 shows the percentage of misclassification errors of different decision trees created for the three scenarios.

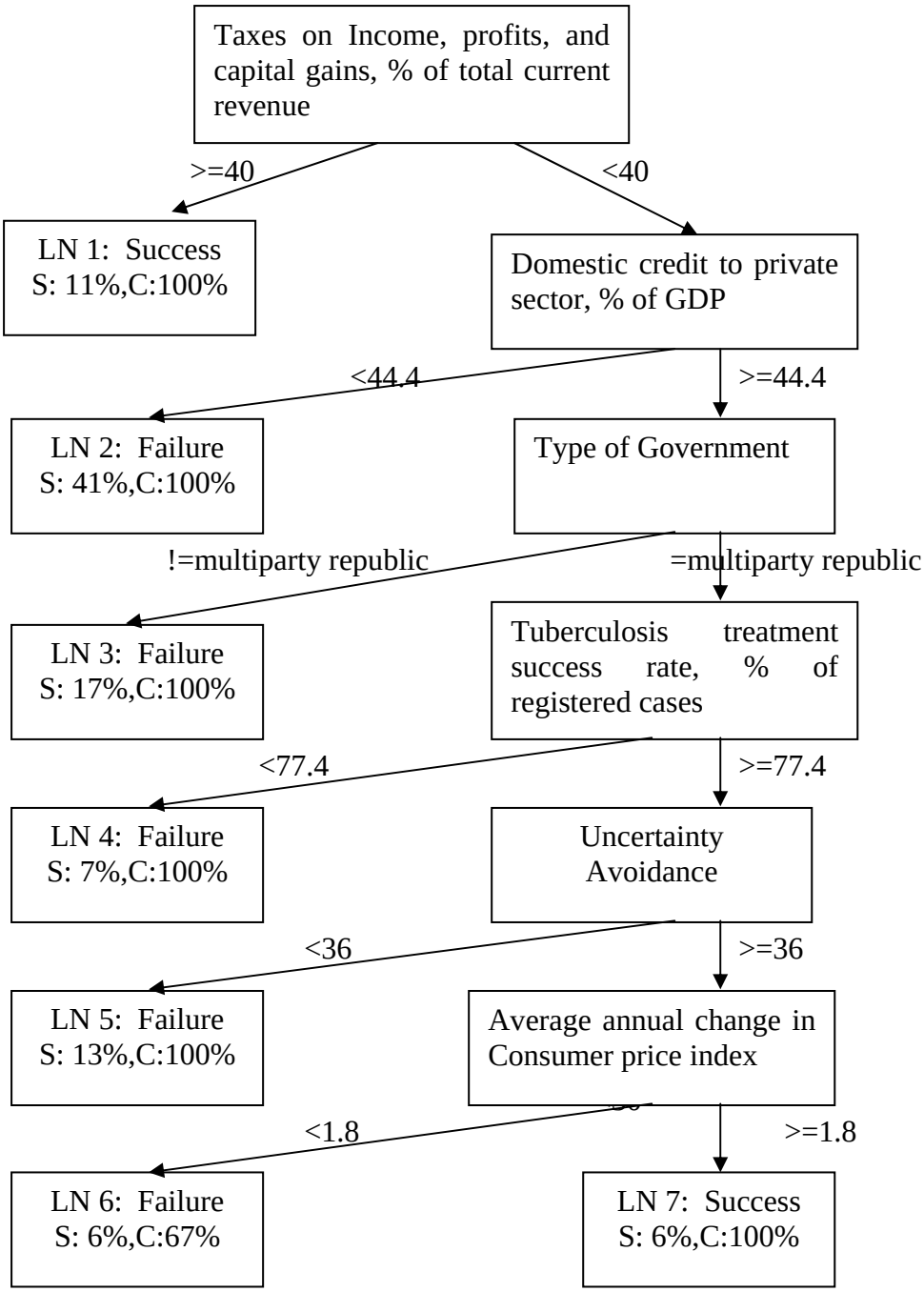
Table 5.4 The percentage of misclassification errors of decision trees for EK method

	First quartile	Median	Third quartile
EK method	5.57	5.56	1.85

In summary, 41 attributes were selected from the 10 aspects of society (Appendix H): 3 attributes belong to demography; 9 attributes belong to economy; 3 attributes belong to education; 4 attributes belong to environment; 3 attributes belong to health; 2 belong to legal; 4 attributes belong to network readiness; 3 attributes belong to culture; 5 attributes belong to technology and 5 attributes belong to institution.

The 67 observations (countries) were divided into two groups randomly. 54 (80%) observations were training group. 13 (20%) observations were testing group. The 41 attributes of society as the predictors took part in the creation of the decision. There were 13 nodes in the decision tree: 7 nodes were the leaf nodes. The percentage of misclassification errors of training data was 1.85%. The percentage of misclassification errors of testing data was 23.08%. The final decision tree shows at Figure 5.9. The confusion matrices of this decision tree shows at Figure 5.10.

In Figure 5.9, 8 observations were allocated in the leaf node 1; 25 observations were allocated in the leaf node 2; 14 observations were allocated in the leaf node 3; 6 observations were allocated in the leaf node 4; 3 observations were allocated in the leaf node 5; 3 observations were allocated in the leaf node 6 and 8 observations were allocated in leaf node 8. The confidences of the leaf nodes were all 100%, except for the leaf note 6. The confidence about the testing data is 76.9%. Therefore the decision tree created by EK method can give a good assessing about whether the GDCs are successful or failure.



Note:

LN: Leaf node S: Percentage of support C: Percentage of confidence

Figure 5.9 The final decision tree about success of GDCs with EK method

Confusion Matrix

Training Data

True Class \ Predicted Class	Failure	Success	
Failure	40	0	40
Success	1	13	14
	41	13	54

Test Data

True Class \ Predicted Class	Failure	Success	
Failure	9	2	11
Success	1	1	2
	10	3	13

Figure 5.10 Confusion Matrixes of decision tree about success of GDCs by EK method

From Figure 5.9, we obtained 5 factors that had effect on the success of GDCs.

1. Taxes on Income, profits, and capital gains, % of total current revenue, 2000
2.Domestic credit to private sector, % of GDP, 2001
3.Type of Government, 2002
4.Uncertainty Avoidance
5.Average annual change in Consumer price index, %, 2000-2001

The five factors belong to economy, culture and institution. So the 3 societal aspects have more effect than other societal aspects. In order to find the important factors, many statements were created. The statement: “if Taxes on Income, profits, and capital gains, % of total current revenue >=40% then the GDC is successful” had 88% confidence and 12% support of 67 countries. The statement: “if Taxes on Income, profits, and capital gains, % of total current revenue <40% then the GDC is failure” had 85% confidence and 88% support of 67 countries. The two statements about “Taxes on Income, profits, and capital gains, % of total current revenue” had the highest confidence in the statements. The statements of other factors had less than 60% confidence. So only this factor was selected. Figure 5.11 shows the comparison among the successful part of status of DGCs determined by EK method and the positive and negative impact of the 5 factors on this part. Figure 5.12 shows the comparison between the failure part of status of DGCs determined by EK method and the positive and negative impact of the 5 factors on this part. It was true that the attribute “taxes on Income, profits, and capital gains, % of total current revenue” was higher than 40% in most successful countries and it was lower than 40% in most failure countries. . From the statements and the figures, it is obvious that economy aspect was the most important one to affect GDCs in this decision tree.

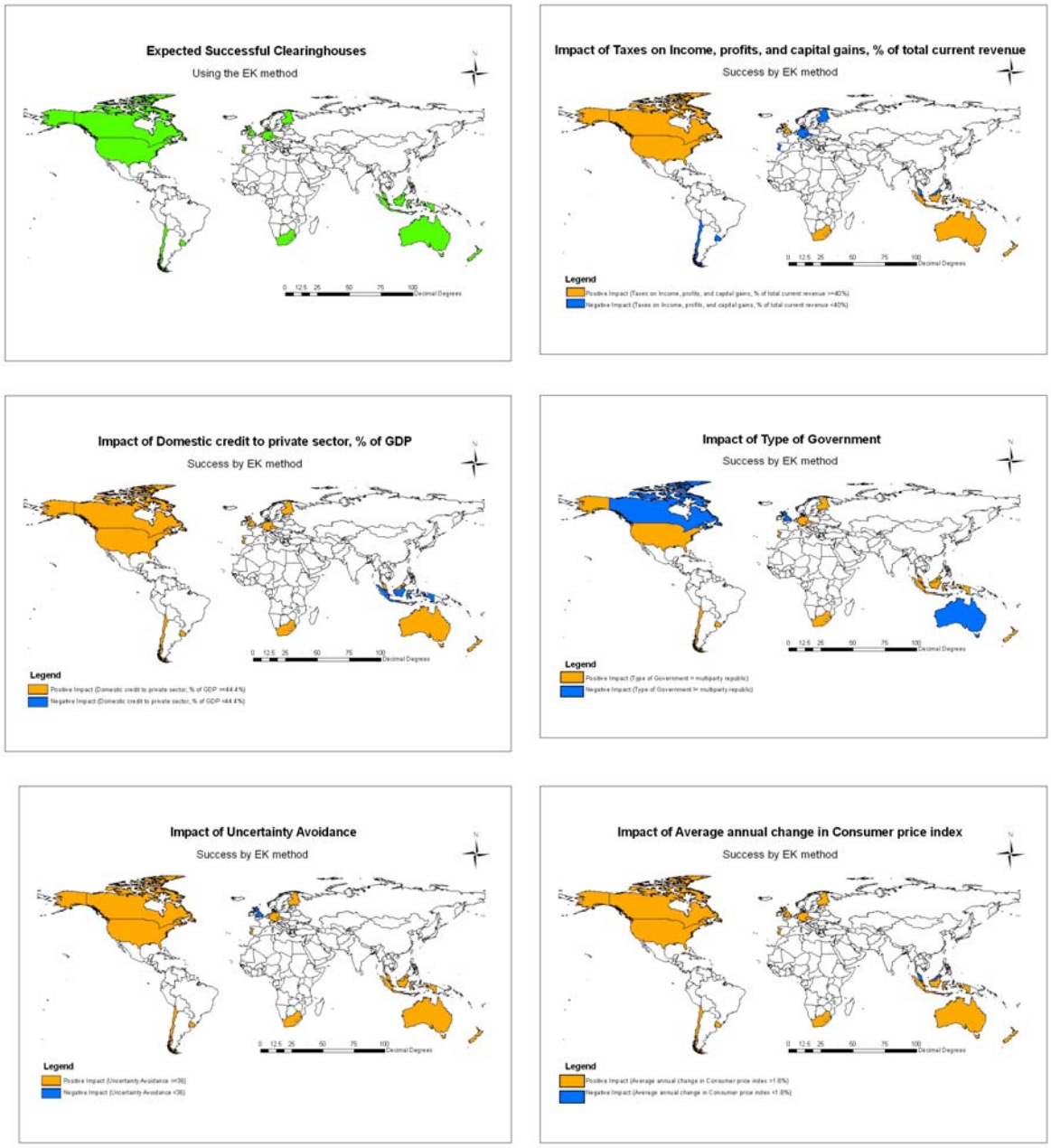


Figure 5.11 Comparison among the successful part of status of GDCs and effect of the 5 factors

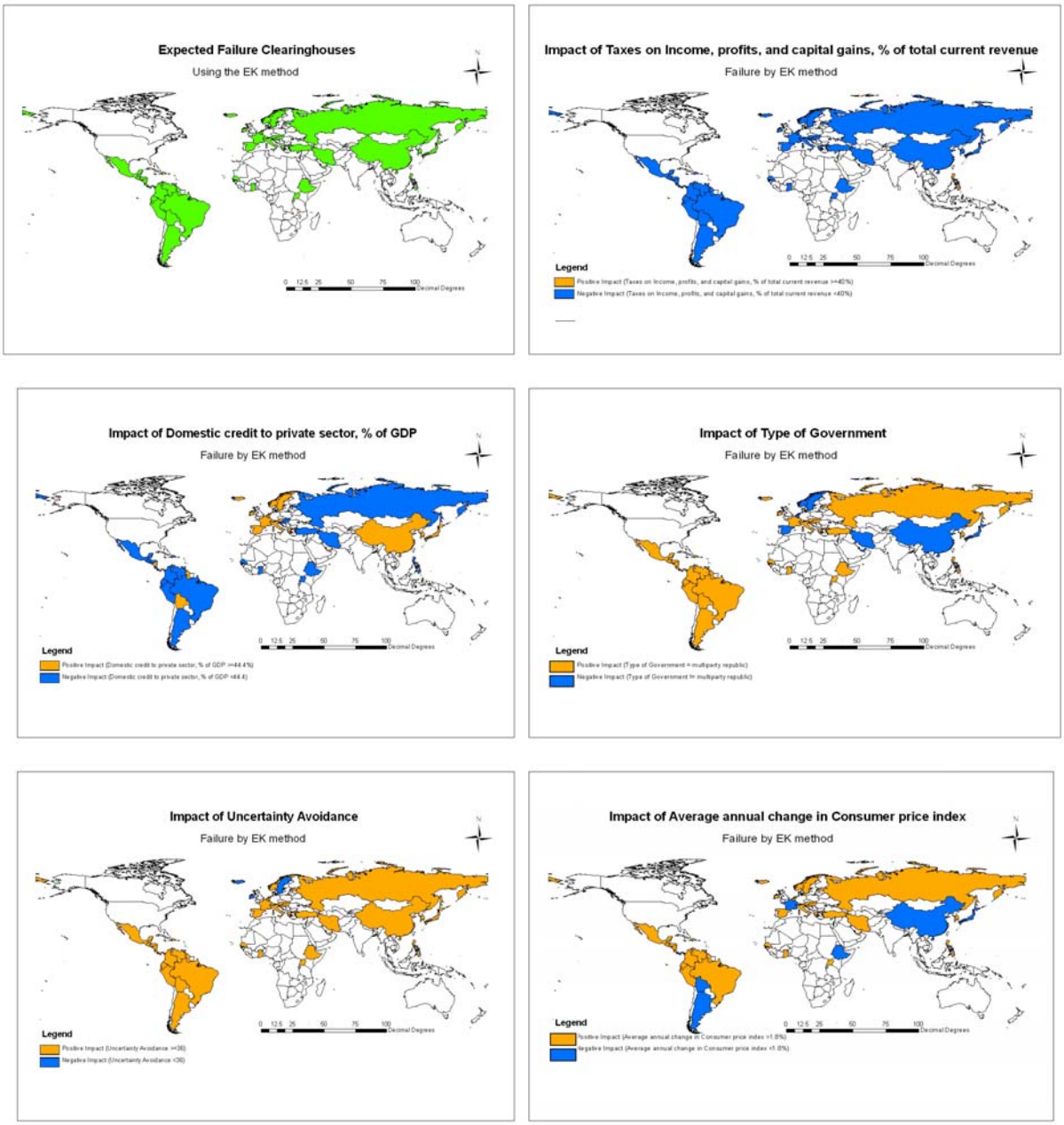


Figure 5.12 Comparison among the Failure part of status of GDCs and effect of the 5 factors

By analyzing and discussing the three decision trees about whether the GDCs exist or not and whether the GDCs are success or failure. The economy aspect was see in every decision tree, and the attributes of this aspect were the important factors in the three decision trees. The economy aspect was the only important aspect for all of the three decision trees. Therefore, if we want to find the most significant aspect to affect the development of GDCs, the economy aspect was the greatest one.

5.5 Results of prediction of success of GDCs using the two method

The 126 countries and regions took part in the prediction with the two methods. By using the decision tree shown in the Figure 5.6, 10 countries and regions were the successful if they created the GDCs. 116 countries were failure: 106 countries were allocated in the leaf node 1; 1 country was allocated in the leaf node 3; 8 countries were allocated in the leaf node 4; 2 countries were allocated in the leaf node 5 and 9 countries were allocated in the leaf node 6.

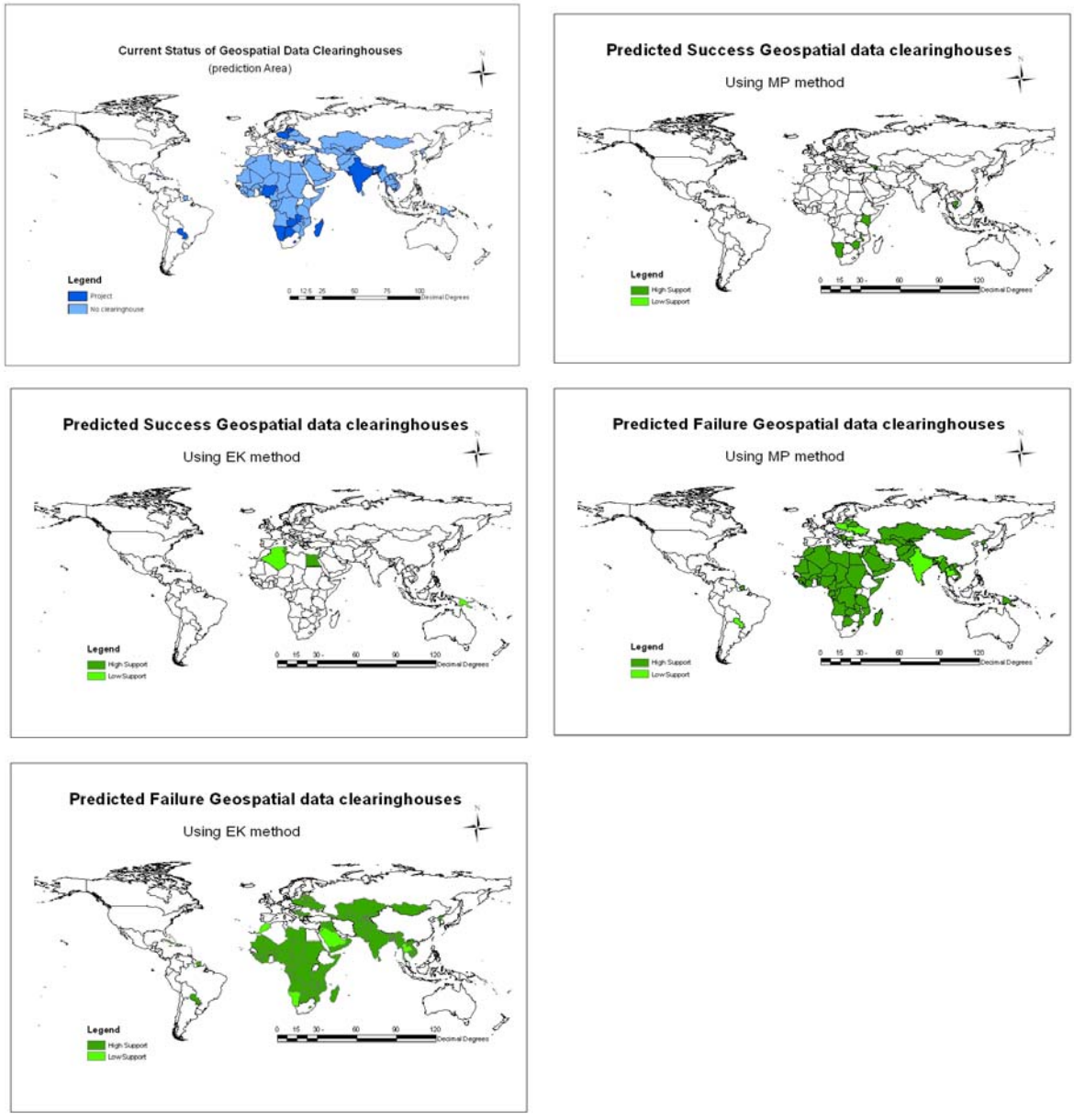


Figure 5.13 The comparison the status of GDCs and success part of the results of prediction by MP method

By using the decision tree shown at the Figure 9, 7 countries were the successful if they created the GDCs. 119 countries were failure: 3 countries were allocated in the leaf node 1; 113 countries were allocated in the leaf node 2; 5 countries were allocated in the leaf node 3; 1 country was allocated in the leaf node 4 and 4 countries were allocated in the leaf node 7. Figure 5.13 shows the predicted countries and the results of prediction by the two decision trees..

Table 5.5 The successful country predicted by two methods

Country name	MP method	EK method
Israel	X	X
Lebanon	X	X
Namibia	X	
Algeria		X
Armenia	X	
Cambodia	X	
Egypt, Arab Rep.		X
Georgia	X	
Kenya	X	
Malawi	X	
Mauritius		X
Papua New Guinea		X
Tunisia		X
Zimbabwe	X	
Taiwan	X	

From Figure 5.13 and Table 5.5, we can see 10 countries and regions were successful decided by MP method. 7 countries were successful decided by the EK method. The Israel and Lebanon were the successful countries by the two methods at the same time. The 111 countries were failure predicted by the two methods at the same time. 13 countries had different results of prediction by the two methods. The results of prediction using the two methods showed in Appendix I. Although the two decision trees came form different methods (MP and EK methods) and they were consisted of different attributes of society, it is obvious that the two decision trees were coherent.

CHAPTER 6
Conclusions and recommendations

This chapter has two parts. Answering some questions is the first part. The second part gives some recommendations.

6.1 Conclusions

Obviously, the society has great effect on the development of GDCs. The thesis tries to explore the impact of the various societal aspects on the GDCs. The research has answered the questions:

What are the criteria that can be used to assess the existence of GDCs?

The decision tree about the existence of GDCs provided the criteria to assess the existence of GDCs. The decision tree has 96% accuracy to describe the existence or no existence of the GDCs. It was consisted of four factors of society. The economy and education were the most important aspects of society.

What are the criteria that can be used to assess the success of GDCs?

There were two methods used for creating the decision trees about whether the GDCs are successful or failure. One was MP method, and the other was EK method. The decision tree of Mp method was consisted of 7 societal attributes and had 94% accuracy for the training and testing data. The health, economy and demography aspects were important in the decision tree. The decision tree of EK method was consisted of 6 societal attributes. Its accuracy was 94% for the training and testing data. The most important societal aspect was economy for this decision tree. The two decision trees generated the criteria that had good representation for their methods individually. The economy was the most important aspect in both of methods

What kinds of attributes are important in these criteria?

The “Average years of schooling”, “Agricultural productivity, Agriculture, value added per worker”, “Health Expenditure, Total % of GDP”, “Rural population %” and “Taxes on Income, profits, and capital gains, % of total current revenue” were the most important attributes to affect the development of GDCs. They belong to education, economy, demography and health. The economy was the most important aspect, because it was important in the three decision trees.

Is it possible to make the prediction about the success of GDCs?

Yes, it is possible. The decision trees provided the good prediction on their methods individually. 113 countries got the same results of prediction form the two different

decision trees. 13 countries got the different results of prediction. The coherence of the prediction of the two decision trees was good, because more than 90% of results of prediction are same by both of methods.

How to apply the data mining (decision tree approach) to determine the impact of society on the existence and success of GDCs?

This is the main question. I created decision trees to determine the impact of society on the existence and success of GDCs, after the data preprocessing. One decision tree was obtained to assess whether the GDCs exit or not. Two decision trees were obtained to assess whether the GDCs are successful or failure. The whole report of thesis had explained the detailed procedures step by step.

6.2 Recommendations

More than 50 percent of missing values were in my study area. The existence of them increases the uncertainty of the results. If it is possible, reducing the missing values may be a good method to increase the creditability.

The software used to do the study was not a very professional. It only provided limited functions. For instance, it only allowed 50 predictors in the data to analyze. The selection of more suitable attributes form 252 attributes has to be done first. The professional software can give us more functions and fewer limits to do the work.

Although the coherence of the two methods was good to analyze whether the GDCs are successful or failure, only 2 countries were successful predicted by the two decision trees at the same time. 13 countries have different results by the two decision trees. So the standards of success of GDCs were still a problem we should think about carefully. I thought that the much more expert knowledge would help us to sharp the standards of success of GDCs.

From the discussions and conclusions, the economy was the most important aspect of society for development of GDCs. It should be focused on. The more work need to do on it to look for the effect of economy on the GDCs.

Because the society should develop continually and the data about the every aspect of the society will change as the time changes, the trees obtained by the study should adjust in time in order to suit the development of society.

References

Answers, 2004, http://www.answers.com/topic/web-portal
Canadian Government, 1986. Management of Government: Major Suverys, A Study Team Report to the Task Force on Program Review, July 31,1985. Canadian Gov Publ Center, Ottawa, Canada.

CEN, 1996, DOC: prEN 287009 (comite europeen de normalization)

Clarke, D., 2000,The Global SDI and Emerging Nations-Challenges and Opportunities for Global Cooperation.15th UNRCC Conference and 6th PCGIAP meeting,11-14 April 2000, 2000http://www.gsi.go.jp/PCGIAP/kl/derekclarke.pdf

CSD,2001,Commission on Sustainable Development Global Issues, environment. Australian Department of Foreign Affairs and Trade. (http://www.dfat.gov.au/environment/csd.html)

Brand, E., and Gerritsen, R.,1998,predicting outcomes is the most popular application of data mining,DBMS,(available at http://www.dbmsmag.com/9807m00.html)

Dilly,R., 1995, Data Mining, An Introduction Student Notes, Parallel Computer Center, Queens University Belfast, Version 2.0, December1995, http://www.pcc.qub.ac.uk/tec/courses/datamining/stu_notes/dm_book_1.html)

DTREG, 2004, http://www.dtreg.com/vartype.htm Edelstein, H., A. (1999). Introduction to data mining and knowledge discovery (3rd ed). Potomac, MD: Two Crows Corp.

Executive Order, 1994, Coordinating geographic data acquisition and access, the National Spatial Data Infrastructure. Executive Order 12906, Federal Register 59, 1767117674, Executive Office of the President, USA

Famili, A., Shen, M., Weber, R., 1997, Data preprocessing and intelligent data analysis, Intell.Data Anal. 1 (1)(1997)3 –23.

Fayyad, U. M., Piatetsky-Shapiro, G., Smyth, P., & Uthurusamy, R. (1996). Advances in knowledge discovery & data mining. Cambridge, MA: MIT Press.

FGDC, 1997, Framework, introduction and guide. (Washington: Federal Geographic Data Committee),PP 106

FGDC, 2004, NSDI (Washington, DC: Federal Geographic Data Committee) (available at http://www.fgdc.gov /nsdi/nsdi.html)

FGDC, 2004, questions and answers about clearinghouses (Washington, DC: Federal Geographic Data Committee) (available at http://www.fgdc.gov/GDC/background.html)

Gore, A., 1998, The Digital Earth: understanding out planet in the 21st century. The Australian Surveyor 43(2): 89-91. http://www.ci.bakersfield.ca.us/gis/notes/misc/digital_earth.htm

Groot, R., McLaughlin, J., 2000, Geospatial data infrastructure: concepts, cases and good practice, oxford, 2000

Hamilton,H.,2001, Knowledge discovery in databases (available at http://www2.cs.uregina.ca/~hamilton/courses/831/index.html)

Han, J., Kamber, M., 2000. Data mining: Concepts and Techniques. New York: Morgan-Kaufman

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

Ho Tu Bao,2002, missing data,knowledge discovery and data mining techniques and practice (available at http://www.netnam.vn/unescocourse/knowlegde/2-3.htm#2.3%20%20%20Missing%20Data)

Hsu, Chung-Chian, 2004, Extending Attribute-Oriented Induction Algorithm for Major Values and Numeric Values, Expert Systems with Applications, Vol. 27, No. 2, pp. 187-202. (SCI, EI)

Hyman, G., and Lance, K., 20001, Adoption and implementation of national spatial data infrastructure in Latin America and the Caribbean,5th GSDI Conference, Cartagena de indias, Colombia.

Hyman, G., Perea, C., Rey, D., Lance, K., 2002, Survey of the Development of National Spatial Data Infrastructures in Latin America and the Caribbean (available at http://gis.esri.com/library/userconf/proc03/p1140.pdf) and implementation of national spatial

ISRO, 2001, NSDI strategy and action plan, India space research organization (available at http://www.isro.org/NSDI.pdf)

Joshi, K., 1997, Analysis of Data Mining Algorithms (available at http://userpages.umbc.edu/~kjoshi1/data-mine/proj_rpt.htm#data_class_mtd

Kantardzic, M., (2002) Data Mining Concepts, Models, Methods and Algorithms

Kulasekere, E., 2004 New Directions In Intelligent Systems (available at http://www.ent.mrt.ac.lk/~ekulasek/it6401/Building%20Classification%20Models%20ID3%20and%20C4_5.htm

Masser, I., 1999, All shapes and sizes: the first generation of National Spatial Data Infrastructures, International Journal of Geographical Information Science 13.

Moxon, B., 1996, Defining Data Mining, DBMS Data Warehouse Supplement. August 1996.

Malczeweki, J., 1999 GIS and Multicriteria Decision Analysis. (New York:John Wiley and Sons),392pp.

OSDM , 2004, Office of Spatial Data Management Glossary (available at http://www.osdm.gov.au/osdm/glossary.html)

Openshaw, S., 1993,Over twenty years of data handling and computing in Environment and Planing. Environment and Planning A, Anniversary Issue

Rajabifard, A., Feeney, M. E., and Williamson, I. P., 2002, Future directions for SDI Development. International Journal of Applied Earth Observation and Geoinformation, 4, 11–22.

Rajabifard, A., Feeney, M. E., and Williamson, I. P., 2003 Chapter two Spatial Data Infrastructures: Concept, Nature and SDI Hierarchy. In Developing Spatial Data Infrastructures: From Concept to Reality, edited by I. Williamson, A, Rajabifard and M. E. Feeney (London: Taylor &Francis)

Rajabifard, A., Feeney, M. E., Williamson, I. P. and Masser, I., 2003 Chapter Six National SDI-initiatives. In Developing Spatial Data Infrastructures: From Concept to Reality, edited by I. Williamson, A, Rajabifard and M. E. Feeney (London: Taylor & Francis), pp. 95–109.

StatSoft, Inc., 2004, Electronic Statistics Textbook. Tulsa (available at http://www.statsoft.com/textbook/stathome.html).

Two Crows Corporation, 2004, Data Mining Glossary (available at http://www.twocrows.com/glossary.htm#anchor314309)

The Pit Master, 2005, http://www.thepitmaster.com/otherresources/glossary.htm

USGS, 2005, http://geo-nsdi.er.usgs.gov/talk/dmt2003/dmt2003.html

WIKIPEDIA, 2004, http://en.wikipedia.org/wiki/Confusion_matrix
World Bank, 2003, *World Development Indicators*. http://www.worldbank.org, The World Bank Group: 2003

Appendices

Appendix A Attributes of society

Note:
Number (194): number of missing values in 194 records
Number (67): number of missing values in 67 records that have GDCs

Demography

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
Population (million), 2001	0	0.00	0	0.00
Population density (people/sq,km2001)	1	0.52	0	0.00
Life expectancy at birth, years, 2001	12	6.19	0	0.00
Life expectancy at birth, Male years, 2001	39	20.10	2	2.99
Life expectancy at birth, Female years, 2001	39	20.10	2	2.99
Gender-related development index (GDI), 2001	51	26.29	2	2.99
Gender empowerment measure (GEM), 2001	123	63.40	18	26.87
Promote gender equality, Ratio of female to male enrollments in primary and secondary school, 2000	61	31.44	13	19.40
Labor force gender parity index, 2001	46	23.71	7	10.45
Average annual population growth rate(% 1980-2001)	43	22.16	6	8.96
Average annual population growth rate(% 2001-15)	44	22.68	7	10.45
Median Age, years, 2002	2	1.03	0	0.00
Population under age 15, as % of total, 2001	25	12.89	1	1.49
Population between ages 15-64, %. 2001	43	22.16	6	8.96
Population ages 15-64, millions, 2001	44	22.68	7	10.45
Crude death rate(/1000people), 2001	44	22.68	7	10.45
Infant mortality rate, per 1000 live births, 2001	44	22.68	7	10.45
Under-five mortality rate per 1000, 2001	45	23.20	7	10.45
Crude birth rate(/1000people), 2001	44	22.68	7	10.45
Total fertility rate, births per woman, 2001	44	22.68	7	10.45
Adolescent fertility rate, births per 1000 women, ages 15-19, 2002	44	22.68	7	10.45
Labor force, total millions, 2001	45	23.20	7	10.45
Labor force, Average annual growth rate, %, 1980-2001	45	23.20	7	10.45
Rural population % (2001)	45	23.20	7	10.45
Urban population(% of total population2001)	45	23.20	7	10.45
Human development index trends(2001)	20	10.31	0	0.00
Life expectancy index, 2001	20	10.31	0	0.00
Total	1058	20.20	141	7.79

Appendix A
Note:
Number (194): number of missing values in 194 records
Number (67): number of missing values in 67 records that have GDCs

Economy

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
Gross national income, $ billions, 2001	19	9.79	3	4.48
Gross national income per capita $(calculated using the worldbank atlas2001)	21	10.82	4	5.97
Purchasing power parity (PPP) gross national income, $ billions	23	11.86	4	5.97
Purchasing power purchasing (PPP) gross national income, per capita, 2001	20	10.31	1	1.49
Purchasing power parity (PPP) conversion factor, local currency units to international $, 2001	55	28.35	9	13.43
Ratio of PPP conversion factor to official exchange rate, 2001	57	29.38	8	11.94
Real effective exchange rate, 1995 = 100, 2001	119	61.34	22	32.84
Gross domestic product, $ millions, 2001	52	26.80	9	13.43
Gross domestic product, average annual % growth, 1990- 2001	51	26.29	7	10.45
Gross domestic product, % growth, 2000-01	23	11.86	4	5.97
Gross domestic product, per capita, % growth, 2000-01	23	11.86	4	5.97
GDP implicit deflator, average annual % growth, 1990-2001	50	25.77	7	10.45
Unemployment: Total, % of total employment, 1998-2001	116	59.79	15	22.39
Long-term unemployment, % of total unemployment, total, 1998-2001	154	79.38	35	52.24
Average hours worked per week, 1995-99	139	71.65	27	40.30
Minimum wage, $ per year, 1995-99	135	69.59	31	46.27
Labor cost per worker in manufacturing, $ per year, 1995-99	114	58.76	17	25.37
Value added per worker in manufacturing, $ per year, 1995-99	125	64.43	19	28.36
Population below $2 a day, %	74	38.14	5	7.46
Poverty gap at $2 a day, %	75	38.66	6	8.96
Gini-index	72	37.11	8	11.94
Gross domestic savings, % of GDP, 2001	53	27.32	9	13.43

Appendix A
Note:
Number (194): number of missing values in 194 records
Number (67): number of missing values in 67 records that have GDCs

Economy

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
Gross national savings, % of GNI, 2001	60	30.93	9	13.43
Net national savings, % of GNI, 2001	59	30.41	9	13.43
Adjusted net savings, % og GNI, 2001	66	34.02	9	13.43
Energy depletion, % of GNI, 2001	52	26.80	8	11.94
Trade in goods, % of GDP, 2001	52	26.80	9	13.43
Trade in goods, % of goods GDP, 2001	94	48.45	28	41.79
Change in trade, % of GDP, 1990-2000	102	52.58	20	29.85
Gross private capital flows, % of GDP, 2001	77	39.69	10	14.93
Net private capital flows, $ millions, 2001	78	40.21	31	46.27
Domestic credit to private sector, % of GDP, 2001	56	28.87	9	13.43
Gross foreign direct investment, % of GDP, 2001	79	40.72	11	16.42
Foreign direct investment, $ millions, 2001	53	27.32	8	11.94
Aid dependency ratio, Aid as % of GNI, 2001	54	27.84	9	13.43
Aid dependency ratio, Aid as % of gross capital formation, 2001	60	30.93	10	14.93
Aid dependency ratio, Aid as % of imports of goods and services, 2001	53	27.32	8	11.94
Euromoney country credit-worthiness rating, September 2002	44	22.68	7	10.45
Value traded, % of GDP, 2001	103	53.09	16	23.88
Listed domestic companies, 2002	99	51.03	12	17.91
Tax revenue, % of GDP, 2001	106	54.64	27	40.30
Taxes on Income, profits and capital gains, % of total taxes, 2001	88	45.36	9	13.43
Taxes on Income, profits, and capital gains, % of total current revenue, 2000	95	48.97	19	28.36
Societal security taxes, % of total current revenue, 2000	84	43.30	8	11.94
Domestic taxes on goods and services, % of value added in industry and services, 2001	94	48.45	12	17.91

Appendix A
Note:
Number (194): number of missing values in 194 records
Number (67): number of missing values in 67 records that have GDCs

Economy

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
Taxes on goods and services, % of total current revenue, 2000	96	49.48	20	29.85
Taxes on international trade, % of total current revenue, 2000	96	49.48	20	29.85
Central government expenditures, Goods and services, % of total expenditure, 2000	100	51.55	20	29.85
Wages and salaries, % of total (governmental) expenditure, 2000	101	52.06	21	31.34
Subsidies and other current transfers, % of total (governmental) expenditures, 2000	100	51.55	20	29.85
Military expenditure, of % of GDP	63	32.47	12	17.91
Military expenditure, % of central government expenditure, 2001	92	47.42	14	20.90
Armed forces personnel, Total thousands, 1999	47	24.23	7	10.45
Armed forces personell, % of labor force, 1999	47	24.23	7	10.45
Services, average annual % growth, 1990-2001	59	30.41	8	11.94
Services value added, % of GDP, 2001	61	31.44	13	19.40
Net barter terms of trade, 2000	83	42.78	17	25.37
Fuels, % of total (national) merchandise exports	83	42.78	10	14.93
Fuels, % of total (national) merchandise imports	82	42.27	9	13.43
Consumer price index, average annual % growth, 1990-2001	60	30.93	8	11.94
Food price index, average annual % growth, 1990-2001	71	36.60	8	11.94
Household final consumption expenditure, % of GDP, 2001	52	26.80	8	11.94
Household final consumption expenditure, $ millions, 2001	57	29.38	9	13.43
Household final consumption expenditure, average annual, % growth, 1990-2001	67	34.54	8	11.94
Household final consumption expenditure per capita, average annual % growth, 1990-2001	69	35.57	8	11.94

Appendix A
Note:
Number (194): number of missing values in 194 records
Number (67): number of missing values in 67 records that have GDCs

Economy

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
General government final consumption expenditure, % of GDP, 2001	56	28.87	8	11.94
General government final consumption expenditure, average annual % growth, 1990-2001	68	35.05	8	11.94
Gross capital formation, average annual % growth, 1990-2001	60	30.93	8	11.94
Central government finances, Current revenue, % of GDP, 2000	86	44.33	11	16.42
Central government finances, Total expenditure% of GDP, 2000	86	44.33	11	16.42
Central government finances, Overall budget balance (including grants), % of GDP, 2000	95	48.97	20	29.85
Total external debt, $ millions, 2001	77	39.69	30	44.78
GDP index	20	10.31	0	0.00
Services: male, % of male employment, 1998-2001	132	68.04	19	28.36
Services: female, % of female employment, 1998-2001	132	68.04	19	28.36
Average annual change in Consumer price index, %, 2000-2001	51	26.29	4	5.97
Total	5627	38.16	937	18.40

Appendix A
Note:
Number (194): number of missing values in 194 records
Number (67): number of missing values in 67 records that have GDCs

Technology

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
Electricity production, billion kwh, 2000	72	37.11	8	11.94
Access to electricity, % of population, 2000	74	38.14	6	8.96
Electric power, consumption per capita kwh 2000	77	39.69	8	11.94
Electric power, Transmission and distribution losses, % of output, 2000	81	41.75	8	11.94
Motor vehicles, per 1000 people, 2000	118	60.82	21	31.34
Motor vehicles, per kilometer of road, 2000	104	53.61	15	22.39
Fuel prices, Super $ per liter, 2002	48	24.74	7	10.45
Fuel prices, Diesel $ per liter, 2002	49	25.26	7	10.45
Telecommunication $million, 2001	94	48.45	33	49.25
Roads, Total road network km, 1995-2000	44	22.68	7	10.45
Roads, % Paved roads, 1995-2000	52	26.80	9	13.43
Roads, Goods hauled million ton-km, 1995-2000	134	69.07	33	49.25
Railways, Rail lines, Total km, 1996-2001	103	53.09	21	31.34
Railways, Traffic Density traffic units per km, 1996-2001	104	53.61	23	34.33
Air, aircraft departures thousands, 2001	60	30.93	10	14.93
Total telephone subscribers, Total (k), 2002	16	8.25	0	0.00
Fixed line and mobile phone subscribers per 1000 people, 2001	58	29.90	9	13.43
Main telephone lines per 100 inhabitants, 2002	5	2.58	0	0.00
Main telephone lines per 100 inhabitants, Annual growth %, 1997-2002	16	8.25	0	0.00
Main telephone lines, (k), 2002	4	2.06	0	0.00
Main telephone lines, Annual growth %, 1997-2002	16	8.25	0	0.00
Mobile phone subscribers, (k), 2002	9	4.64	0	0.00
Mobile phone subscribers, Annual growth rate, %, 1997-2002	33	17.01	1	1.49
Mobile phones per 100 inhabitants, 2002	9	4.64	0	0.00

Mobile phone subscribers, % Digital, 2002	84	43.30	21	31.34	
Mobile phone, as % of total telephone subscribers, 2002	19	9.79	0	0.00	

Appendix A
Note:
Number (194): number of missing values in 194 records
Number (67): number of missing values in 67 records that have GDCs

Technology

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
International telecommunications, Outgoing traffic minutes per subscriber, 2001	59	30.41	9	13.43
Personal Computers, thousands, 2002	33	17.01	1	1.49
Personal computers per 1000 people, 2001	68	35.05	7	10.45
Personal computers in education, 2001	143	73.71	25	37.31
Internet, Hosts Total, 2002	21	10.82	0	0.00
Internet, hosts per 10000 inhab., 2002	24	12.37	0	0.00
Internet, Users thousands, 2001	46	23.71	7	10.45
Internet, Users per 10000 inhab., 2002	6	3.09	0	0.00
Internet, Monthly off-peak access charges, Service provider charge, $, 2001	77	39.69	18	26.87
Internet, Monthly off-peak access charges, Telephone usage charge, $, 2001	85	43.81	28	41.79
Internet, Secure servers, 2001	85	43.81	8	11.94
ICT-expenditures, % of GDP, 2001	145	74.74	27	40.30
ICT-expenditures per capita, $, 2001	146	75.26	27	40.30
Scientists and engineers in R&D, per million people, 1990-2000	98	50.52	13	19.40
Technicians in R&D, per million people, 1990-2001	103	53.09	17	25.37
Expenditures for R&D, % of GDP, 1989-2000	124	63.92	18	26.87
High-technology exports, $ millions, 2001	88	45.36	9	13.43
High-technology exports, % of manufactured exports, 2001	87	44.85	9	13.43
Total	2921	34.22	470	15.94

Appendix A
Note:
Number (194): number of missing values in 194 records
Number (67): number of missing values in 67 records that have GDCs

Culture

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
Power Distance	99	51.03	3	4.48
Uncertainty Avoidance	99	51.03	3	4.48
Long-term Thinking	156	80.41	48	71.64
Individualism/Collectivism	99	51.03	3	4.48
Masculinity/Feminity	99	51.03	3	4.48
Total	552	56.91	60	17.91

Legal

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
Legal system, 2002	8	4.12	1	1.49
Royalty and license fees, Receipts, $ millions, 2001	107	55.15	20	29.85
Royalty and license fees, Payments, $ millions, 2001	100	51.55	15	22.39
Patent applications, filed, residents, 2000	87	44.85	14	20.90
Patent applications filed, non-residents, 2000	85	43.81	13	19.40
International Convention on the Elimination of all Forms of Racial Discrimination 1966	2	1.03	0	0.00
International Covenant on Civil and Political Rights 1966	2	1.03	0	0.00
International Covenant on Economic, Societal and Cultural Rights1966	2	1.03	0	0.00
Convention on the Elimination of All Forms of Discrimination Against women 1979	2	1.03	0	0.00
Convention Against Torture and Other Cruel, Inhuman or Degrading Treatment or Punishment 1984	2	1.03	0	0.00
Convention on the Rights of the Child 1989	2	1.03	0	0.00
Total	399	18.70	63	8.55

Appendix A
Note:
Number (194): number of missing values in 194 records
Number (67): number of missing values in 67 records that have GDCs

Education

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
Adult literacy rate %age 15 and above 2001	4	2.06	0	0.00
Adult illiteracy rate, Male, % ages 15 and above, 2001	56	28.87	5	7.46
Adult illiteracy rate, Female, % ages 15 and above, 2001	56	28.87	5	7.46
Public expenditure on education % of GDP 2000	68	35.05	11	16.42
Public expenditure on education, per student % of GDP per capita, 2000	121	62.37	28	41.79
Public expenditure on education, % of total government expenditure, 2000	112	57.73	26	38.81
Primary pupil-teacher ratio, pupils per teacher, 2000	59	30.41	12	17.91
Combined primary, secondary and tertiary gross enrolment ratopm (%), 2000-01	13	6.70	0	0.00
Gross enrollment ratio, Primary % of relevant age group, 2000	55	28.35	10	14.93
Gross enrollment ratio, Tertairy % of relevant age group, 2000	68	35.05	15	22.39
Net enrollment ratio, Primary % of relevant age group, 2000	63	32.47	12	17.91
Primary completion rate, Total, 1995-2001	50	25.77	6	8.96
Average years of schooling, Total, 2000	95	48.97	14	20.90
Education expenditure, % of GNI, 2001	57	29.38	9	13.43
Education index	20	10.31	0	0.00
Total	897	30.82	153	15.22

Appendix A
Note:
Number (194): number of missing values in 194 records
Number (67): number of missing values in 67 records that have GDCs

Network Readiness

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
Networked Readiness Index, 2002	115	59.28	9	13.43
Network Use, 2002	116	59.79	10	14.93
Enabling Factors, 2002	116	59.79	10	14.93
Network access, 2002	116	59.79	10	14.93
Network Policy, 2002	116	59.79	10	14.93
Networked society, 2002	116	59.79	10	14.93
Networked economy, 2002	116	59.79	10	14.93
Info Infrastructure, 2002	115	59.28	10	14.93
Hardware, Software, and Support, 2002	116	59.79	10	14.93
ICT Policy, 2002	116	59.79	10	14.93
Business and Economic Environment, 2002	116	59.79	10	14.93
Networked learning, 2002	116	59.79	10	14.93
ICT Opportunities, 2002	116	59.79	10	14.93
Societal Capital, 2002	116	59.79	10	14.93
E-Commerce, 2002	116	59.79	10	14.93
E-government, 2002	116	59.79	10	14.93
General Infrastructure, 2002	116	59.79	10	14.93
Total	1970	59.73	169	14.84

Appendix A
Note:
Number (194): number of missing values in 194 records
Number (67): number of missing values in 67 records that have GDCs

Health

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
Private health expenditure, % of total, 2000	45	23.20	7	10.45
Public expenditure on health, % of GDP, 2000	44	22.68	7	10.45
Health Expenditure, Total % of GDP, 1997-2000	45	23.20	7	10.45
Health expenditure per capita, $, 1997-2000	45	23.20	7	10.45
Hospital beds, per 1000 people, 1995-2000	98	50.52	15	22.39
Child immunization rate, % of childern under age one, Measles, 2001	43	22.16	6	8.96
Child immunization rate, % of childern under age one, DTP, 2001	44	22.68	7	10.45
Tuberculosis treatment success rate, % of registered cases, 1999	87	44.85	27	40.30
Incidence of tuberculosis, per 100000 people, 2000	44	22.68	7	10.45
Prevalence of HIV, % of adults, 2001	44	22.68	7	10.45
Improve maternal health, Births attended by skilled health staff, % of total, 2000	80	41.24	13	19.40
Prevalence of undernourishment, % of population, 1998-2000	70	36.08	8	11.94
Population with sustainable access to affordable essential drugs, %, 1999	23	11.86	0	0.00
Total	712	28.23	118	13.55

Appendix A
Note:
Number (194): number of missing values in 194 records
Number (67): number of missing values in 67 records that have GDCs

Environment

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
Surface area thousand sq,km (2001)	0	0.00	0	0.00
Access to an improved water source, % population, 2000	63	32.47	15	22.39
Access to improved sanitation facilities, % of population, 2000	70	36.08	20	29.85
Land Area, Thousands sq. km, 2000	2	1.03	0	0.00
Land use, Arable land, % of land area, 2000	45	23.20	7	10.45
Land use, Permanent crop land, % of land area, 2000	48	24.74	9	13.43
Land use, Other, % of land area, 2000	48	24.74	9	13.43
Arable land, hectares per capita, 1998-2000	44	22.68	6	8.96
Irrigated land, % of cropland, 1998-2000	53	27.32	11	16.42
Crop production index, 1999-2001	52	26.80	8	11.94
Food production index, 1999-2001	52	26.80	8	11.94
Livestock production index, 1999-2001	52	26.80	8	11.94
Cereal yield, kilograms per hectare, 1999-2001	47	24.23	9	13.43
Agricultural productivity, Agriculture, value added per worker, 1995 $, 1999-2001	64	32.99	12	17.91
Forest area, % of total land area, 2000	46	23.71	8	11.94
Average annual deforestation, %, 1990-2000	45	23.20	8	11.94
Mammals, Threatened species, 2000	44	22.68	7	10.45
Nationally protected area, % of total land area, 2002	50	25.77	8	11.94
Freshwater resources, Internal flows billion cu. M, 2000	45	23.20	8	11.94
Freshwater resources, Total renewable resources per capita cu. M, 2000	45	23.20	8	11.94
Annual freshwater withdrawals, billion cu. m	49	25.26	9	13.43
Annual freshwater withdrawals, % of total renewable resources	55	28.35	10	14.93

Emissions of organic water pollutants, kilograms per day, 2000	80	41.24	13	19.40
Emissions of organic water pollutants, kilograms per day per worker, 2000	80	41.24	13	19.40
Commercial energy use, thousand metric tons of oil equivalent, 2000	72	37.11	8	11.94

Appendix A
Note:
Number (194): number of missing values in 194 records
Number (67): number of missing values in 67 records that have GDCs

Environment

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
Commercial energy use, average annual % growth, 1980-2000	91	46.91	11	16.42
Commercial energy use per capita, kg of oil equivalent, 2000	72	37.11	8	11.94
Commercial energy use per capita, average annual % growth, 1980-2000	92	47.42	12	17.91
Carbon dioxide emissions, Total million metric tons, 1999	12	6.19	0	0.00
Carbon dioxide emissions per capita metric tons, 1999	40	20.62	2	2.99
Carbon dioxide emissions damage, % of GNI, 2001	53	27.32	9	13.43
Continent	0	0.00	0	0.00
Year, Environmental strategy or action plan	45	23.20	7	10.45
Year, Treaty, Climate Change	43	22.16	6	8.96
Year, Treaty, Ozone layer	44	22.68	7	10.45
Year, Treaty, CFC control	44	22.68	7	10.45
Year, Treaty, Law of the Sea	43	22.16	6	8.96
Year, Treaty, Biological diversity	44	22.68	7	10.45
Total	1874	25.42	304	11.94

Institution

Name of Attribute	Number(194)	Percentage(194)	Number(67)	Percentage(67)
Composite International Country Risk Guide (ICRG) risk rating, September 2002 (Political, Financial and economic risk components)	65	33.51	7	10.45

Institutional Investor credit rating, September 2002	55	28.35	7	10.45
Type of Government, 2002	0	0.00	0	0.00
Federation	0	0.00	0	0.00
Year Independence, 2002	1	0.52	0	0.00
Year Current borders established, 2002	0	0.00	0	0.00
Total	121	10.40	14	3.48

Appendix B Weight of Attributes for MP Method

Name of Attribute	Condition1	weight	Condition2	weight
(meta)data accessibility	Metadata	2	Non_Standard	1
Year, first version of implementation on internet	>=1999	2	<1999	1
Number of Datasets	>=62	2	<62	1
Number of data suppliers	>=3	2	<3	1
Last update (days)	>=55	1	<55	2
Metadata Standard	Is CEN or FGDC	2	Not	1
ISO-project	Is ISO	2	Not	1
Language used	English	2	Not	1
Web mapping service	Yes	2	No	1
Maps used	Yes	2	No	1
Newest dataset	>=20	2	<20	1
Register	Yes	1	No	2
Web_AltaVista	>=20	2	<20	1
Web_Google	>=76	2	<76	1
FGDC-node	Yes	2	No	1
Stability of Funding	Yes	2	No	1

Appendix C Weight of Aspects

Name of Aspect	Weight
Data	0.2
People	0.3
Standards	0.07
Access	0.4
Policy	0.03

Appendix D Weight of Attributes for EK Method

Name of Attribute	Aspect	Condition1	Weight	Condition2	Weight	Condition3	Weight	Condition4	Weight	Condition5	Weight	Condition6	Weight
(meta)data accessibility	Data	Metadata + data	50	Metadata	31.25	Non-standardised	12.5						
Number of Datasets	Data	>=5000	50	>=1500 and <5000	37.5	>=500 and <1500	25	>=100 and <500	12.5	<100	0		
Number of data suppliers	People	>=51	15	>=16 and <51	12	>=6 and <16	8	>=2 and <16	4	<2	0		
Last update (days)	People	>=365	0	>=101 and <365	4	>=21 and <101	10	>=4 and <21	16	<4	20		
Metadata Standard	Standards	CEN, FGDC and AS	80	National	50	Other	20	No_standard	0				
ISO-project	Standards	Yes	20	No	0								
Language used	People	English	45	Mutli(no E)	30	Single (E)	15	Single	0				
Web mapping service	Access	Yes	30	No	0								
Maps used	Access	Yes (b+l)	30	Yes (l)	25	Yes(b)	20	No	0				
Register	Policy	Yes	0	Partly	5	No	20						
Web_AltaVista	People	>=1000	10	>=250 and <1000	8	>=101 and <250	6	>=31 and <101	4	>=6 and <31	2	<6	0
Web_Google	People	>=1000	10	>=250 and <1000	8	>=101 and <250	6	>=21 and <101	4	>=4 and <21	2	<4	0
Stability of Funding	Policy	Yes	80	No,piecemeal	10	No	0						
Status	Access	GDC	10	portal	0								
(De)centralisation	Access	Yes	30	No	0								

Appendix E Survey question and results

Survey question:
On which criteria is your classification based? Explain your choice of criteria.

Results

	(Percentage, normalized for one year)			
Criteria used to classify (attributes)				
	2000	**2001**	**2002**	**Total**
Number of datasets (availability)	5	6	12	9
Metadata Accessibility	9	9	7	8
Quality of Metadata description	2	4	4	3
Data Accessibility (Data Download)	14	9	6	8
Search mechanism	16	19	12	15
Web Mapping	5	2	3	3
Preview dataset		1	2	1
Speed	2	8	3	4
Interface (User-friendliness)	12	10	8	10
Update frequency	5	5	7	6
Metadata Standards	5	4	5	4
Language (English)	14	15	12	14
Payment	7	1	2	3
Linkage to other web-sites	2	1	5	3
International data	2	3	0	1
Number of themes		3	1	1
Number of visitors			2	1
Number of data suppliers			2	1
Contact Webmaster			4	2
Registration			2	1
Vision			1	0
Total (%)	100	100	100	100
# participants	21	32	41	94

Appendix F The attributes selected for the existence of GDC

Gender-related development index (GDI), 2001
Population (million), 2001
Rural population % (2001)
Average annual population growth rate(% 2001-15)
Adolescent fertility rate, births per 1000 women, ages 15-19, 2002
Gross domestic product, average annual % growth, 1990- 2001
Euromoney country credit-worthiness rating, September 2002
Taxes on international trade, % of total current revenue, 2000
Gross domestic savings, % of GDP, 2001
Purchasing power parity (PPP) conversion factor, local currency units to international $, 2001
Services value added, % of GDP, 2001
Household final consumption expenditure, $ millions, 2001
Military expenditure, of % of GDP
Food price index, average annual % growth, 1990-2001
Average years of schooling, Total, 2000
Adult illiteracy rate, Male, % ages 15 and above, 2001
Combined primary, secondary and tertiary gross enrolment ratopm (%), 2000-01
Adult literacy rate %age 15 and above 2001
Primary pupil-teacher ratio, pupils per teacher, 2000
Arable land, hectares per capita, 1998-2000
Land use, Permanent crop land, % of land area, 2000
Access to improved sanitation facilities, % of population, 2000
Agricultural productivity, Agriculture, value added per worker, 1995 $, 1999-2001
Nationally protected area, % of total land area, 2002
Emissions of organic water pollutants, kilograms per day, 2000
Surface area thousand sq,km (2001)
Prevalence of undernourishment, % of population, 1998-2000
Health Expenditure, Total % of GDP, 1997-2000
Health expenditure per capita, $, 1997-2000
Improve maternal health, Births attended by skilled health staff, % of total, 2000
Legal system, 2002
Patent applications, filed, residents, 2000
Patent applications filed, non-residents, 2000
Internet, Secure servers, 2001
Access to electricity, % of population, 2000
Internet, Monthly off-peak access charges, Service provider charge, $, 2001
Internet, Monthly off-peak access charges, Telephone usage charge, $, 2001
Institutional Investor credit rating, September 2002
Year Independence, 2002
Composite International Country Risk Guide (ICRG) risk rating, September 2002

Appendix G The sum of weight for every country by two methods

Country name	MP method	EK method	Country name	MP method	EK method
Austria	23	46	Colombia	24	46
Czech Rep.	26	51	Costa Rica	22	33
Denmark	28	57	Dom. Rep.	21	33
Finland	25	72	Dominica	22	49
France	23	29	El Salvador	23	38
Germany	24	62	Guatemala	22	27
Hungary	19	35	Honduras	21	35
Iceland	24	37	Mexico	21	46
Ireland	21	46	Nicaragua	22	37
Netherlands	24	42	Peru	21	31
Portugal	24	48	Trinidad & tobago	19	31
Slovenia	23	45	Uruguay	25	50
Sweden	22	33	USA	30	97
Switzerland	23	36	Venezuela	23	41
UK	26	70	Belgium	21	28
Ethiopia	22	35	Croatia	21	37
Ghana	21	35	Ecuador	17	13
Senegal	19	24	Estonia	22	41
South-Africa	27	64	Greece	17	8
Brunei	21	33	Guyana	17	18
China	24	44	Iran	19	28
Indonesia	24	62	Italy	18	10
Japan	23	43	Luxemburg	17	12
Malaysia	23	61	New Zealand	24	63
Philippines	22	36	Norway	22	45
Qatar	21	27	Panama	15	9
Singapore	21	44	Russia	19	12
Australia	28	83	Slovak Rep.	20	22
Argentina	25	43	South-Korea	19	17
Barbados	19	31	Spain	18	34
Bolivia	22	34	Turkey	16	14
Brazil	23	37	Uganda	18	18
Canada	27	94	United Arab Emirates	20	17
Chile	26	56			

Appendix H The attributes selected for the existence of GDC by two methods

Aspect of society	Attributes	MP method	EK method
demography	Average annual population growth rate(% 1980-2001)	X	
	Crude death rate(/1000people), 2001	X	
	Life expectancy at birth, years, 2001	X	
	Median Age, years, 2002	X	
	Rural population % (2001)	X	
	Population (million), 2001	X	
	Labor force gender parity index, 2001	X	
	Population (million), 2001		X
	Life expectancy at birth, Female years, 2001		X
	Labor force, total millions, 2001		X
economy	Net national savings, % of GNI, 2001	X	
	Gross domestic product, % growth, 2000-01	X	
	GDP implicit deflator, average annual % growth, 1990-2001	X	
	Total external debt, $ millions, 2001	X	
	Household final consumption expenditure per capita, average annual % growth, 1990-2001	X	
	General government final consumption expenditure, % of GDP, 2001	X	
	GDP index	X	
	Net barter terms of trade, 2000	X	
	Gross national income, $ billions, 2001	X	X
	Net private capital flows, $ millions, 2001		X
	Domestic credit to private sector, % of GDP, 2001		X
	Foreign direct investment, $ millions, 2001		X
	Purchasing power parity (PPP) conversion factor, local currency units to international $, 2001		X
	Taxes on Income, profits, and capital gains, % of total current revenue, 2000		X
	Central government finances, Overall budget balance (including grants), % of GDP, 2000		X
	Average annual change in Consumer price index, %, 2000-2001		X
	Gross capital formation, average annual % growth, 1990-2001		X
education	Public expenditure on education % of GDP 2000	X	
	Adult literacy rate %age 15 and above 2001	X	
	Gross enrollment ratio, Primary % of relevant age group, 2000	X	X
	Combined primary, secondary and tertiary gross enrolment ratopm (%), 2000-01		X
	Adult illiteracy rate, Male, % ages 15 and above, 2001		X
environment	Nationally protected area, % of total land area, 2002	X	
	Carbon dioxide emissions, Total million metric tons, 1999	X	
	Access to improved sanitation facilities, % of population, 2000	X	
	Land use, Arable land, % of land area, 2000	X	
	Continent	X	X
	Mammals, Threatened species, 2000		X
	Agricultural productivity, Agriculture, value added per worker, 1995 $, 1999-2001		X
	Food production index, 1999-2001		X

Appendix H The attributes that were selected for the existence of GDC by two methods

Aspect of society	Attributes	MP method	EK method
health	Health Expenditure, Total % of GDP, 1997-2000	X	
	Child immunization rate, % of childern under age one, Measles, 2001	X	
	Public expenditure on health, % of GDP, 2000	X	
	Incidence of tuberculosis, per 100000 people, 2000		X
	Tuberculosis treatment success rate, % of registered cases, 1999		X
	Population with sustainable access to affordable essential drugs, %, 1999		X
legal	Patent applications, filed, residents, 2000	X	X
	Patent applications filed, non-residents, 2000	X	X
	Convention on the Elimination of All Forms of Discrimination Against women 1979	X	
	International Covenant on Civil and Political Rights 1966	X	
network reading	Social Capital, 2002	X	
	General Infrastructure, 2002	X	
	Info Infrastructure, 2002	X	
	Networked Readiness Index, 2002	X	
	Hardware, Software, and Support, 2002		X
	E-Commerce, 2002		X
	Network Policy, 2002		X
	Network Use, 2002		X
culture	Power_Distance	X	X
	Uncertainty Avoidance	X	X
	Individualism/Collectivism	X	X
	Masculinity/Feminity	X	
technology	Roads, Total road network km, 1995-2000	X	
	Internet, Hosts Total, 2002	X	
	Main telephone lines, Annual growth %, 1997-2002	X	
	Fuel prices, Diesel $ per liter, 2002	X	X
	Main telephone lines per 100 inhabitants, 2002		X
	Internet, hosts per 10000 inhab., 2002		X
	Internet, Monthly off-peak access charges, Telephone usage charge, $, 2001		X
	Electricity production, billion kwh, 2000		X
institutional	Year Independence, 2002	X	
	Composite International Country Risk Guide (ICRG) risk rating, September 2002	X	X
	Institutional Investor credit rating, September 2002	X	X
	Type of Government, 2002	X	X
	Year Current borders established, 2002	X	X
	Federation		X

Appendix I The results of prediction by the two methods

	MP method	EK	Country name	MP	EK
Botswana	F	F	Myanmar	F	F
Cyprus	F	F	Nepal	F	F
India	F	F	Niger	F	F
Israel	S	S	Oman	F	F
Latvia	F	F	Pakistan	F	F
Lesotho	F	F	Papua New Guinea	F	S
Lithuania	F	F	Puerto Rico	F	F
Madagascar	F	F	Rwanda	F	F
Namibia	S	F	Saudi Arabia	F	F
Nigeria	F	F	Sierra Leone	F	F
Paraguay	F	F	Somalia	F	F
Poland	F	F	Sri Lanka	F	F
Zambia	F	F	Sudan	F	F
Afghanistan	F	F	Swaziland	F	F
Albania	F	F	Syrian Arab Republic	F	F
Algeria	F	S	Tajikistan	F	F
Angola	F	F	Tanzania	F	F
Armenia	S	F	Thailand	F	F
Azerbaijan	F	F	Togo	F	F
Bangladesh	F	F	Tunisia	F	S
Belarus	F	F	Turkmenistan	F	F
Benin	F	F	Ukraine	F	F
Bosnia and Herzegovina	F	F	Uzbekistan	F	F
Bulgaria	F	F	Vietnam	F	F
Burkina Faso	F	F	Yemen, Rep.	F	F
Burundi	F	F	Yugoslavia	F	F
Cambodia	S	F	Zimbabwe	S	F
Cameroon	F	F	Andorra	F	F
Central African Republic	F	F	Antigua and Barbuda	F	F
Chad	F	F	Bahamas	F	F
Congo, Dem. Rep.	F	F	Bahrain	F	F
Congo, Rep.	F	F	Belize	F	F
Cote d'Ivoire	F	F	Bhutan	F	F
Cuba	F	F	Cape Verde	F	F
Egypt, Arab Rep.	F	S	Comoros	F	F
Eritrea	F	F	Djibouti	F	F
Gabon	F	F	East Timor (Timor-Leste)	F	F
Gambia	F	F	Equatorial Guinea	F	F
Georgia	S	F	Fiji	F	F
Guinea	F	F	Grenada	F	F
Guinea-Bisseau	F	F	Kiribati	F	F
Haiti	F	F	Liechtenstein	F	F
Iraq	F	F	Maldives	F	F
Jamaica	F	F	Malta	F	F
Jordan	F	F	Marshall Islands	F	F
Kazakhstan	F	F	Micronesia, Fed. Sts.	F	F
Kenya	S	F	Monaco	F	F
Korea, Dem. Rep.	F	F	Palau	F	F
Kuwait	F	F	Saint Kitts and Nevis	F	F
Kyrgyz Republic	F	F	Saint Lucia	F	F
Lao PDR	F	F	St. Vincent and the Grenadines	F	F
Lebanon	S	S	Samoa	F	F
Liberia	F	F	San Marino	F	F
Libya	F	F	Sao Tome and Principe	F	F
Macedonia, FYR	F	F	Seychelles	F	F
Malawi	S	F	Solomon Islands	F	F
Mali	F	F	Suriname	F	F
Mauritania	F	F	Tonga	F	F
Mauritius	F	S	Vanatu	F	F
Moldava	F	F	Holy See	F	F
Mongolia	F	F	Nauru	F	F
Morocco	F	F	Tuvalu	F	F
Mozambique	F	F	Taiwan	S	F

Appendix J The Example of the 40 decision trees to process the msising values (Mean method for technology aspect)

Node View

Node ID 5
Non-leaf Node

Node Size

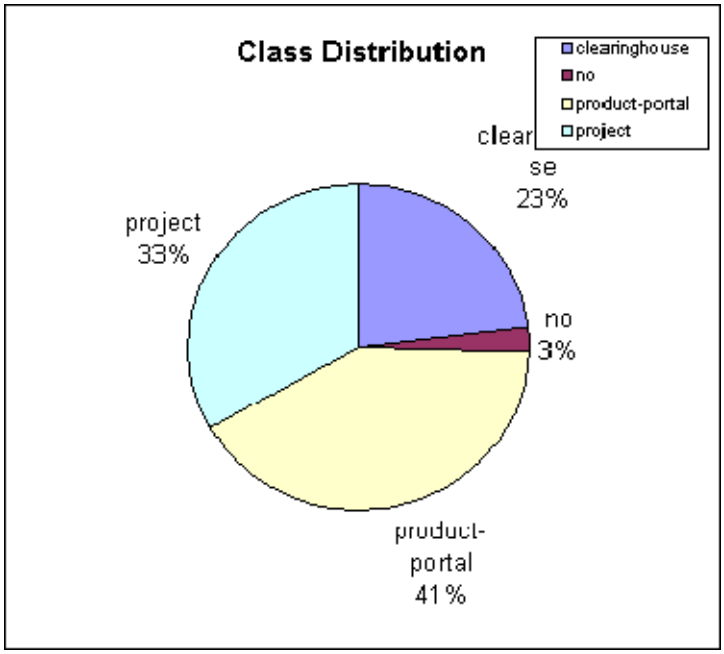
Number of Records	39
% of total Records	20.21%

Majority Class 3

% MissClassified 58.97%

Class Distribution

Class	Label	Proportion
1	clearinghouse	23.08%
2	no	2.56%
3	product-portal	41.03%
4	project	33.33%



Rule

Rule Text

Rule0		Status_1 = no
Rule1	IF	Personal computers in education, 2001 < 58491
	THEN	Status_1 = no
Rule2	IF	Internet, Secure servers, 2001 >= 56
	THEN	Status_1 = clearinghouse
Rule3	IF	ICT-expenditures, % of GDP, 2001 >= 7.5
	THEN	Status_1 = clearinghouse
Rule4	IF	ICT-expenditures per capita, $, 2001 >= 325
	AND	ICT-expenditures, % of GDP, 2001 < 7.5
	THEN	Status_1 = product-portal
Rule5	IF	ICT-expenditures per capita, $, 2001 >= 325
	AND	Telecommunication $million, 2001 >= 1443.3
	THEN	Status_1 = clearinghouse
Rule6	IF	Personal computers in education, 2001 >= 58491
	AND	Roads, Goods hauled million ton-km, 1995-2000 < 8474
	THEN	Status_1 = project
Rule7	IF	Roads, Goods hauled million ton-km, 1995-2000 >= 8474
	THEN	Status_1 = clearinghouse

Rule

Rule Summary Table

Rules 7

Rule ID	Class	Length	Support	Confidence	Capture
0	no	0	100.0%	58.5%	100.0%
1	no	1	59.6%	97.4%	99.1%
2	clearinghouse	1	28.5%	65.5%	75.0%
3	clearinghouse	1	20.7%	92.5%	77.1%
4	product-portal	2	9.3%	77.8%	73.7%
5	clearinghouse	2	15.5%	93.3%	58.3%
6	project	2	11.4%	54.5%	92.3%
7	clearinghouse	1	32.6%	66.7%	87.5%

Results

Classification

Number of Training	193
Number of Test	0
Number of	44
Class Variable	Status_1
Number of Classes	4
Majority Class	no
% MissClassified if is used as Predicted	41%

Tree

Total Number of	18
Number of Leaf	10
Number of	6
%	
On Training	3.11%
On Test Data	0.00%

Time Taken

Data	2 Sec
Tree Growing	3 Min :
Tree Pruning	0 Sec
Tree Drawing	0 Sec
Classification using Rule	2 Sec
	0 Sec
Total	3 Min :

Confusion Matrix

Training Data

Test

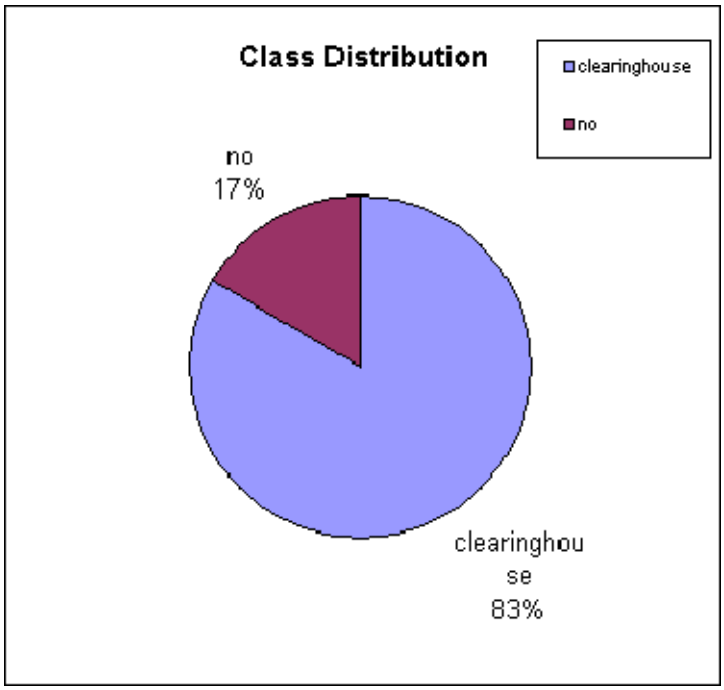
True \ Predicted	clearingho	n	product	project	
clearingho	48				48
no		1			113
product-	4	1	14		19
project			1	12	13
	52	1	15	12	193

Appendix K The Example of the 10 decision trees to select the pretictor variables for the decision tree about the existence of GDCs
(For technology aspect)

Node view

Node ID	**4**
Non-leaf Node	
Node Size	
Number of Records	6
% of total Records	3.11%
Majority Class	1
% MissClassified	16.67%
Class Distribution	

Class	Label	Proportion
1	clearinghouse	83.33%
2	no	16.67%



Rule

Rule Text

Rule0		Status_2 = no
Rule1	IF	ICT-expenditures per capita, $, 2001 >= 196
	THEN	Status_2 = clearinghouse
Rule2	IF	ICT-expenditures per capita, $, 2001 < 196
	THEN	Status_2 = no
Rule3	IF	Access to electricity, % of population, 2000 < 45
	THEN	Status_2 = no
Rule4	IF	Telecommunication $million, 2001 >= 3770.68
	THEN	Status_2 = clearinghouse

Rule Summary Table

Rules 4

Rule ID	Class	Length	Support	Confidence	Capture
0	no	0	100.0%	65.3%	100.0%
1	clearinghouse	1	32.6%	98.4%	92.5%
2	no	1	67.4%	96.2%	99.2%
3	no	1	16.6%	90.6%	23.0%
4	clearinghouse	1	25.4%	95.9%	70.1%

Results

Classification Tree Model

Number of Training observations	193
Number of Test observations	0
Number of Predictors	44
Class Variable	Status_2
Number of Classes	2
Majority Class	no
% MissClassified if Majority Class is used as Predicted Class	35%

Tree Information

Total Number of Nodes	6
Number of Leaf Nodes	4
Number of Levels	4

% Missclasssified

On Training Data	0.52%
On Test Data	0.00%

Time Taken

Data Processing	12 Sec
Tree Growing	2 Min : 33 Sec
Tree Pruning	0 Sec
Tree Drawing	0 Sec
Classification using final tree	9 Sec
Rule Generation	1 Sec
Total	2 Min : 55 Sec

Confusion Matrix

Training Data

Test Data

Predicted Class

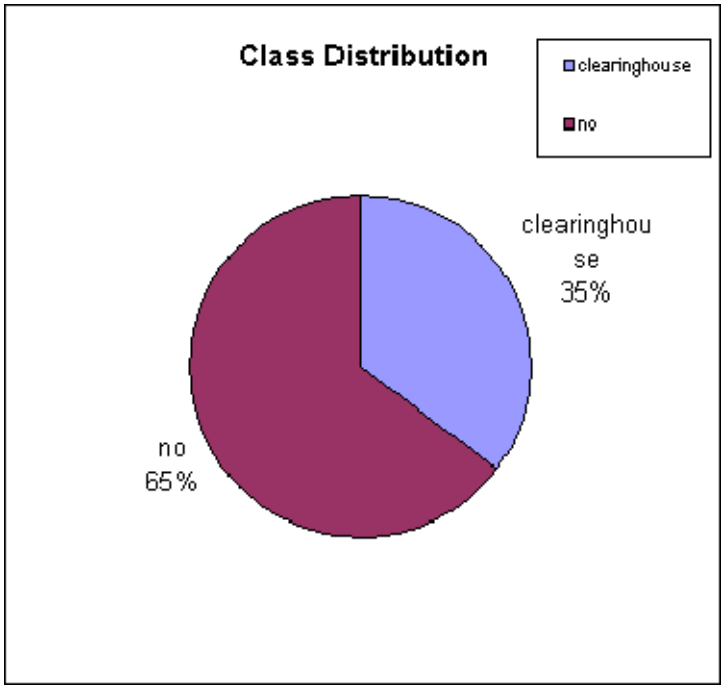
True Class	clearinghouse	no	
clearinghouse	67		67
no	1	125	126
	68	125	193

Appendix L The final decision tree about the existence of GDCs

Node view

Node ID	**0**
Non-leaf Node	
Node Size	
Number of Records	142
% of total Records	100.00%
Majority Class	2
% MissClassified	35.21%
Class Distribution	

Class	Label	Proportion
1	clearinghouse	35.21%
2	no	64.79%



Rule

Rule Text

Rule0		status_2 = no
Rule1	IF	Average years of schooling, Total, 2000 < 4.6
	THEN	status_2 = no
Rule2	IF	Agricultural productivity, Agriculture, value added per worker, 1995 $, 1999-2001 >= 3590
	THEN	status_2 = clearinghouse
Rule3	IF	Average years of schooling, Total, 2000 < 7.2
	AND	Average years of schooling, Total, 2000 >= 4.6
	THEN	status_2 = no
Rule4	IF	Average years of schooling, Total, 2000 >= 7.2
	THEN	status_2 = clearinghouse
Rule5	IF	Agricultural productivity, Agriculture, value added per worker, 1995 $, 1999-2001 < 3590
	AND	Internet, Monthly off-peak access charges, Telephone usage charge, $, 2001 >= 0.61
	THEN	status_2 = no

Rule

Rule Summary Table

Rules 5

Rule ID	Class	Length	Support	Confidence	Capture
0	no	0	100.0%	64.8%	100.0%
1	no	1	52.1%	98.6%	79.3%
2	clearinghouse	1	30.3%	88.4%	76.0%
3	no	2	21.1%	56.7%	18.5%
4	clearinghouse	1	26.8%	94.7%	72.0%
5	no	2	16.9%	58.3%	15.2%

Results

Classification Tree

Number of Training	142
Number of Test observations	51
Number of Predictors	40
Class Variable	status_2
Number of Classes	2
Majority Class	no
% MissClassified if Majority is used as Predicted Class	26%

Tree

Total Number of	10
Number of Leaf	6
Number of	5

%

On Training	1.41%
On Test Data	9.80%

Time Taken

Data	1 Sec
Tree Growing	1 Min :
Tree Pruning	0 Sec
Tree Drawing	0 Sec
Classification using Rule	2 Sec
Total	0 Sec 1 Min :

Confusion Matrix

Training Data

True Class	Predicted Class: clearinghouse	Predicted Class: no	
clearinghouse	49	1	50
no	1	91	92
	50	92	142

Test

True	Predicted: clearingh	Predicted: no	
clearingh	13	4	17
no	1	33	34
	14	37	51

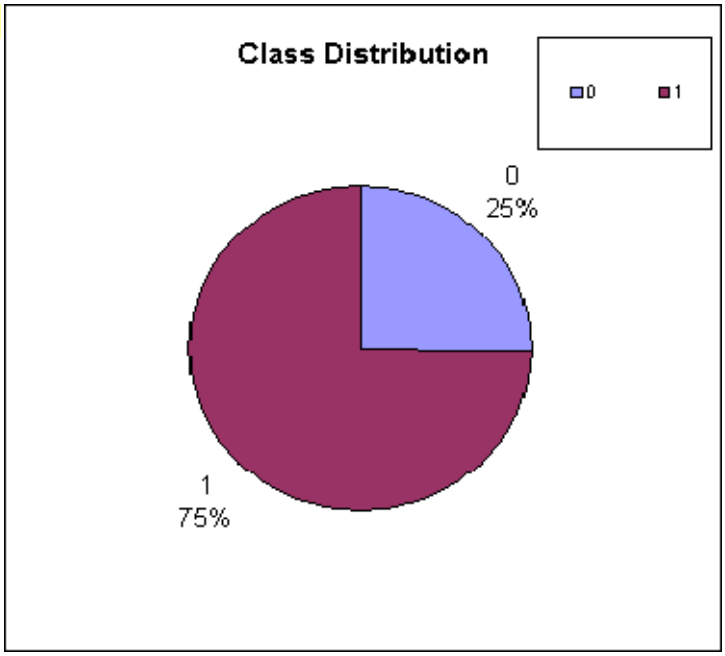
Appendix M The Example of the 60 decision trees to select the pretictor variables for the decision trees about the success of GDCs

(For Technology Aspect)

Node view

Node ID	**0**
Non-leaf Node	
Node Size	
Number of Records	67
% of total Records	100.00%
Majority Class	2
% MissClassified	25.37%
Class Distribution	

Class	Label	Proportion
1	0	25.37%
2	1	74.63%



Rule

Rule Text

Rule0		uneven_25 = 1
Rule1	IF	Personal computers per 1000 people, 2001 >= 275.7
	THEN	uneven_25 = 1
Rule2	IF	Electric power, consumption per capita kwh 2000 >= 4075
	AND	Personal computers per 1000 people, 2001 < 275.7
	THEN	uneven_25 = 0
Rule3	IF	Internet, Hosts Total, 2002 >= 7725
	THEN	uneven_25 = 1
Rule4	IF	Personal Computers, thousands, 2002 < 110
	THEN	uneven_25 = 1
Rule5	IF	Roads, % Paved roads, 1995-2000 < 20.1
	THEN	uneven_25 = 1

Rule

Rule Summary Table

Rules 5

Rule ID	Class	Length	Support	Confidence	Capture
0	1	0	100.0%	74.6%	100.0%
1	1	1	26.9%	100.0%	36.0%
2	0	2	17.9%	83.3%	58.8%
3	1	1	76.1%	80.4%	82.0%
4	1	1	13.4%	77.8%	14.0%
5	1	1	14.9%	80.0%	16.0%

Results

Classification Tree Model

Number of Training observations	67
Number of Test observations	0
Number of Predictors	33
Class Variable	uneven_25
Number of Classes	2
Majority Class	1
% MissClassified if Majority Class is used as Predicted Class	25%

Tree

Total Number of Nodes	12
Number of Leaf Nodes	7
Number of	7

%

On Training	2.99%
On Test Data	0.00%

Time Taken

Data	2 Sec
Tree Growing	1 Min : 3
Tree Pruning	0 Sec
Tree Drawing	0 Sec
Classification using Rule	1 Sec
	1 Sec
Total	1 Min : 8

Confusion Matrix

Training Data

Test

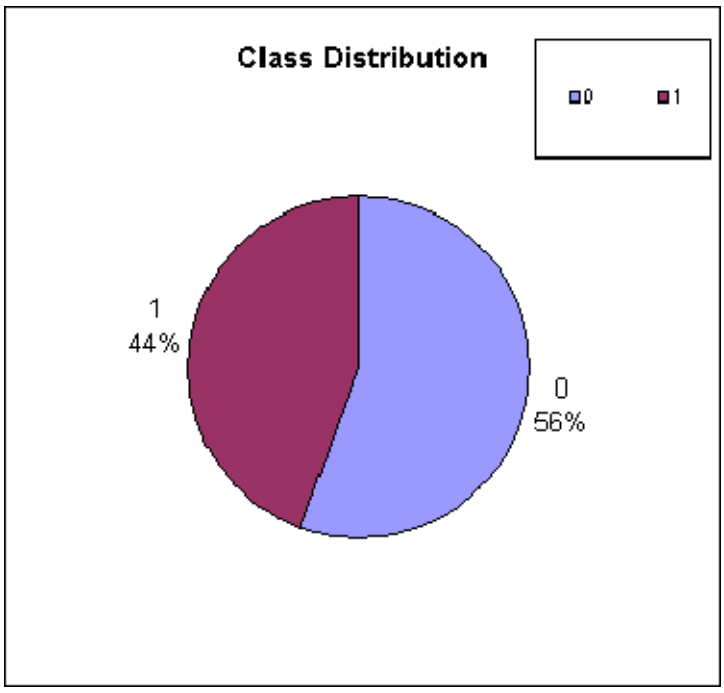
True Class	Predicted 0	Predicted 1	
0	17		17
1	2	48	50
	19	48	67

Appendix N The Example of the 3 decision trees about the success of GDCs using MP method

Node View

Node ID	**0**
Non-leaf Node	
Node Size	
Number of Records	54
% of total Records	100.00%
Majority Class	1
% MissClassified	44.44%
Class Distribution	

Class	Label	Proportion
1	0	55.56%
2	1	44.44%



Rule

Rule Text

Rule0		status_1 = 0
Rule1	IF	Internet, Hosts Total, 2002 < 24138
	THEN	status_1 = 0
Rule2	IF	Land use, Arable land, % of land area, 2000 >= 15.2
	THEN	status_1 = 0
Rule3	IF	Crude death rate(/1000people), 2001 < 6
	THEN	status_1 = 0
Rule4	IF	Health Expenditure, Total % of GDP, 1997-2000 >= 7.2
	THEN	status_1 = 1
Rule5	IF	Gross domestic product, % growth, 2000-01 >= 1.1
	THEN	status_1 = 0

Rule

Rule Summary Table

Rules 5

Rule ID	Class	Length	Support	Confidence	Capture
0	0	0	100.0%	55.6%	100.0%
1	0	1	33.3%	94.4%	56.7%
2	0	1	35.2%	57.9%	36.7%
3	0	1	13.0%	71.4%	16.7%
4	1	1	42.6%	78.3%	75.0%
5	0	1	70.4%	57.9%	73.3%

Results

Classification Tree

Number of Training	54
Number of Test observations	13
Number of Predictors	48
Class Variable	status_1
Number of Classes	2
Majority Class	0
% MissClassified if Majority is used as Predicted Class	36%

Tree

Total Number of	14
Number of Leaf	8
Number of	5

%

On Training	1.85%
On Test Data	23.08%

Time Taken

Data	7 Sec
Tree Growing	52 Sec
Tree Pruning	0 Sec
Tree Drawing	0 Sec
Classification using Rule	1 Sec
	0 Sec
Total	1 Min : 0

Confusion Matrix

Training Data

True Class	Predicted 0	Predicted 1	
0	30		30
1	1	23	24
	31	23	54

Test

True	Predicted 0	Predicted 1	
0	7	2	9
1	1	3	4
	8	5	13

Appendix O The Example of the 3 decision trees about the success of GDCs using EK method

Node view

Node ID **0**

Non-leaf Node

Node Size

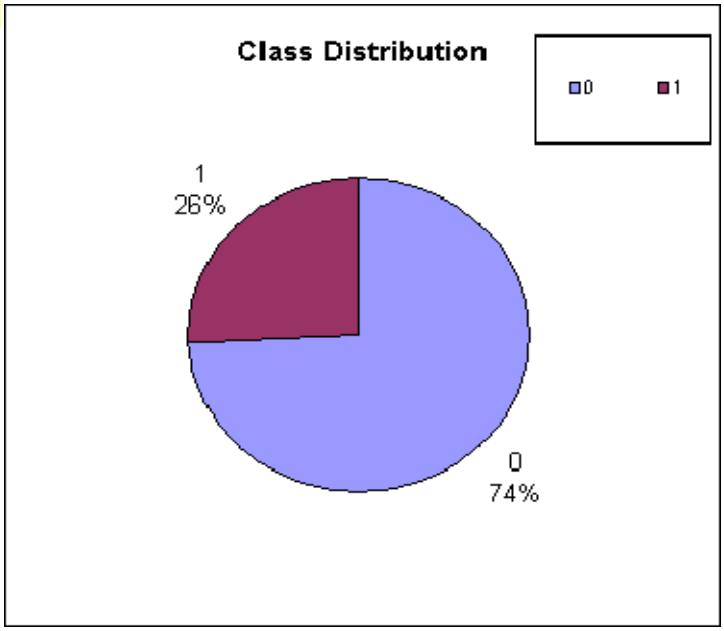
Number of Records	54
% of total Records	100.00%

Majority Class 1

% MissClassified 25.93%

Class Distribution

Class	Label	Proportion
1	0	74.07%
2	1	25.93%



Rule

Rule Text

Rule0		status_1 = 0
Rule1	IF	Taxes on Income, profits, and capital gains, % of total current revenue, 2000 >= 40
	THEN	status_1 = 1
Rule2	IF	Domestic credit to private sector, % of GDP, 2001 < 44.4
	THEN	status_1 = 0
Rule3	IF	Type of Government, 2002 = absolute monarchy
	THEN	status_1 = 0
Rule4	IF	Type of Government, 2002 = constitutional monarchy
	THEN	status_1 = 0
Rule5	IF	Type of Government, 2002 = other
	THEN	status_1 = 0
Rule6	IF	Tuberculosis treatment success rate, % of registered cases, 1999 < 78
	THEN	status_1 = 0
Rule7	IF	Uncertainty Avoidance < 36
	THEN	status_1 = 0
Rule8	IF	Average annual change in Consumer price index, %, 2000-2001 < 1.8
	THEN	status_1 = 0

Rule

Rule Summary Table

Rules 8

Rule ID	Class	Length	Support	Confidence	Capture
0	0	0	100.0%	74.1%	100.0%
1	1	1	11.1%	100.0%	42.9%
2	0	1	42.6%	95.7%	55.0%
3	0	1	5.6%	100.0%	7.5%
4	0	1	9.3%	80.0%	10.0%
5	0	1	3.7%	100.0%	5.0%
6	0	1	35.2%	89.5%	42.5%
7	0	1	11.1%	83.3%	12.5%
8	0	1	18.5%	90.0%	22.5%

Results

Classification Tree Model

Number of Training observations	54
Number of Test observations	13
Number of Predictors	41
Class Variable	status_1
Number of Classes	2
Majority Class	0
% MissClassified if Majority Class is used as Predicted Class	21%

Tree

Total Number of Nodes	15
Number of Leaf Nodes	10
Number of	7

%

On Training	1.85%
On Test Data	23.08%

Time Taken

Data	1 Sec
Tree Growing	43 Sec
Tree Pruning	0 Sec
Tree Drawing	0 Sec
Classification using	1 Sec
Rule	0 Sec
Total	45 Sec

Confusion Matrix

Training Data

True Class \ Predicted	0	1	
0	40		40
1	1	13	14
	41	13	54

Test

True \ Predicted Class	0	1	
0	9	2	11
1	1	1	2
	10	3	13