

**ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE**

**A PROTOTYPE WEB BASED EXPERT SYSTEM FOR ROUTE FINDING TO THE
ROAD NETWORK OF ADDIS ABABA
(AARFExSys)**

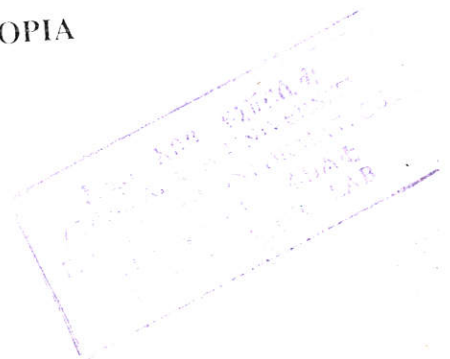
**BY
DANIEL ADMASSU**

**PRESENTED TO THE SCHOOL OF GRADUATE STUDIES OF ADDIS ABABA
UNIVERSITY IN PARTIAL FULFILMENT OF THE REQUIRMENTS FOR THE DEGREE
OF MASTER OF SCIENCE IN INFORMATION SCIENCE**

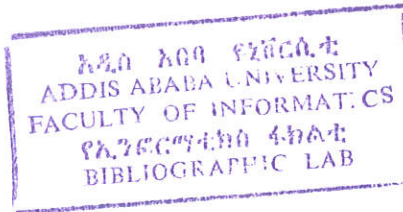
UNDER THE SUPERVISION OF DR. DANEAL FEKERSILLASSIE

**ADDIS ABABA, ETHIOPIA
JULY, 2005**

**ADDIS ABABA UNIVERS
LIBRARIES
P.O. BOX 1176
ADDIS ABABA ETHIOPIA**



ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
Faculty of Informatics
Department of Information Science



WEB BASED SYSTEM FOR ROUTE FINDING TO THE ROAD NETWORK OF
ADDIS ABABA

BY
DANIEL ADMASSU

Name and Signature of Members of the Examining Board

Dr B. L. Desai, Chairman, Examining Board

A blue ink signature of Dr. B. L. Desai, consisting of stylized initials and a surname.

Dr. Daneal Fekersellassie, Advisor

Daneal Fekersellassie

Dr. S. K. Sarma, examiner

A blue ink signature of Dr. S. K. Sarma, consisting of a stylized 'S' and 'K' followed by a surname.

Chairman, Faculty

Signature

Date

Chairman, Graduate Council

Signature

Date

A black ink signature of Daniel Admassu, written in a cursive style. To the right of the signature is the date '12/07/05'.

DEDICATION

*This work is dedicated to my mother
Etagegnehu Bekele*

DEDICATION

*This work is dedicated to my mother
Etagegnehu Bekele*

ACKNOWLEDGEMENT

I would like to express my sincere appreciation to the many individuals who contributes to this thesis. A special thanks is given to my advisor, Dr. Daneal Fekersillassie, for his assistance, supports and friendship offered throughout my degree program. It has been a pleasure working with him.

I would like to thank Yared Abebe, an expert in Ethiopian Mapping Authority, for his helpful suggestions on shaping the initial concept of the problem and contribution on specific comments. My sincere thanks go also to my family for their endless support and encouragement in completion of this work.

Many thanks are extended to Meseret Ayano, librarian in the department of information science, for her patience and helping hand. Especial thanks are also given to all the staffs in the department of information science and my classmates for their friendship and assistance.

The financial support of this study by the graduate study program is gratefully acknowledged. Finally, but the Greatest Thanks to **God**. Thanks **God** for your Endless Love, Grace and Splendid plan for me.

List of Acronyms

AI	Artificial Intelligence
GPS	Global Positioning Status
ES	Expert System
GIS	Geographical Information System
GUI	Graphical User Interface
CF	Certainty Factor
AAMP	Addis Ababa Master Plan
WWW	World Wide Web
KPWIN	Knowledge Pro for WINdows
ORAAMP	Office for the Revision of Addis Ababa Master Plan

List of Figures

Figure 2.1. Convergence of the three factors to create the modern rule based expert system.....	16
Figure 2.2. General structure of a modern rule based system.....	19
Fig. 3.1 Structure of a Geographic Information System (GIS).....	38
Figure 4.2. Concepts in Users' Location	57
Fig 4.4 Geo-referenced map of Addis Ababa in Arc View GIS software.	60
Fig.4.5 The three selected locations of Addis Ababa	61
Figure 6.1. showsWelcome to web Based Expert System For Route Finding to the Road Network of Addis Ababa (AARFExSys).....	73
Fig. 6.2. Sample international hotel in the city of Addis Ababa.....	74
(Fig.6.3): A Prototype web Based Expert System For Route Finding to the Road Network of Addis Ababa (AARFExSys).....	76
Fig.6. 4 How AARFExSys looks at run time	78
Figure 6.5. The first question in AARFExSys prototype system.	78
Fig.6.6. Below shows identifying the location of users.....	79
Fig.6.7.Shows the means of transport users plan to use.	80
Fig.6.9.Shows the final result in a new window	84
Fig. 6.10.Shows the result when the user prefers a contract taxi.....	85

List of Table

Table 1.Shows the input and output parameters.	56
---	----

Table of Contents

ACKNOWLEDGEMENT	i
LIST OF ACRONYMS	ii
LIST OF FIGURES	iii
LIST OF TABLE	iii
TABLE OF CONTENTS	IV
ABSTRACT	IV
CHAPTER ONE	vi
1.1. BACK GROUND OF THE STUDY	1
1.1.1 KNOWLEDGE-BASED SYSTEM IN TRANSPORTATION SECTOR	1
1.1.2. ADDIS ABABA ROUTE COMPLEXITY	3
1.2. STATEMENT OF THE PROBLEM AND JUSTIFICATION OF THE STUDY	4
1.3. OBJECTIVES	6
1.3.1. GENERAL OBJECTIVE	6
1.3.2. SPECIFIC OBJECTIVES	7
1.4. SCOPE AND LIMITATION OF THE STUDY	7
1.5. METHODS	8
1.5.1. Data collection Method	8
1.5.1.1. LITERATURE REVIEW	9
1.5.2. METHODS OF DEVELOPING THE SYSTEM	10
1.5.2.1. PROTOTYPING	10
1.5.2.2. PROGRAMMING TOOLS	11
1.6 ORGANIZATION OF THE THESIS	11
CHAPTER TWO	13
LITERATURE REVIEW	13
2.1 INTRODUCTION	13
2.2 DEFINITION OF EXPERT SYSTEMS	14
2.3. APPROPRIATE DOMAIN OF AN EXPERT SYSTEM	15
2.5. THE BASIC ADVANTAGE AND CHARACTERISTICS OF AN EXPERT SYSTEM	17
2.6. RULE BASED EXPERT SYSTEM	18
2.6.1. THE PRODUCTION RULE SYSTEM	20
2.6.2. THE INFERENCE ENGINE	20
2.6.3. UNCERTAINTY	22
2.6.4. DIFFERENT METHODS OF HANDLING UNCERTAINTY	22
2.7. THE APPLICATION OF EXPERT SYSTEM IN ROUTE FINDING TO TRANSPORT ROAD NETWORK	23
2.8. TOUR OPERATORS AND TRAVEL AGENCIES IN ETHIOPIA	31
2.9. THE ADVANTAGE OF WORLD WIDE WEB IN TOURISM PROMOTION	31

CHAPTER THREE	33
GENERAL OVERVIEW OF THE ADDIS ABABA CITY AND ROADS.....	33
3.1. ESTABLISHMENT OF THE CITY	33
3.2. STANDARD AND DEFINITION OF ROADS.....	34
3.3. ADDIS ABABA CITY ROAD NETWORK.....	35
3.4. STRUCTURE OF A GEOGRAPHIC INFORMATION SYSTEM (GIS).	37
3.4.1. GIS, DECISION SUPPORT, AND EXPERT SYSTEM	37
3.5. EXPERT SYSTEMS AND GIS TO THE CONTRIBUTION OF TOURISM INDUSTRY	41
3.6. TRANSPORT INFRASTRUCTURE IN THE COUNTRY	43
4.1 INTRODUCTION <i>→ Chapter Four</i>	46
4.2.KNOWLEDGE ACQUISITION TECHNIQUES	47
4.2.1. TRADITIONAL KNOWLEDGE ACQUISITION	47
4.2.2. AUTOMATED KNOWLEDGE ACQUISITION.....	47
4.3. ANALYSIS OF THE INTERVIEW.....	51
4.4. CONCEPTUAL MODELING.....	54
4.4.1. THE DIRECTION CHOICE	56
4.4.2. LOCATION OF USERS	56
4.4.2. LOCATION OF USERS	57
4.4.3. MEANS OF TRANSPORTATION.....	58
4.5.GEO-REFERENCED MAP OF ADDIS ABABA	60
CHAPTER FIVE	65
KNOWLEDGE REPRESENTATION	65
5.1 INTRODUCTION	65
5.2 THE KNOWLEDGE BASE	66
5.3 THE EXPLANATION FACILITY	68
5.5 MEASURING EXPERT SYSTEM PERFORMANCE.....	69
5.6 SYSTEM'S CONTROL STRUCTURE	71
CHAPTER SIX.....	73
APPLICATION OF KNOWLEDGE	73
6.1 WEB DESIGN	73
6.2. USER INTERFACE.....	75
6.3. TESTING AND EVALUATION.....	86
6.4. KNOWLEDGE BASE MAINTENANCE.....	87
6.5. DESIGN REQUIREMENTS	87
CONCLUSIONS AND RECOMMENDATIONS	88
7.1 CONCLUSIONS	88
7.2 RECOMMENDATIONS	90
BIBLIOGRAPHY	92
Annexes	92

CHAPTER THREE	33
GENERAL OVERVIEW OF THE ADDIS ABABA CITY AND ROADS.....	33
3.1. ESTABLISHMENT OF THE CITY	34
3.2. STANDARD AND DEFINITION OF ROADS.....	35
3.3. ADDIS ABABA CITY ROAD NETWORK.....	37
3.4. STRUCTURE OF A GEOGRAPHIC INFORMATION SYSTEM (GIS).	37
3.4.1. GIS, DECISION SUPPORT, AND EXPERT SYSTEM	41
3.5. EXPERT SYSTEMS AND GIS TO THE CONTRIBUTION OF TOURISM INDUSTRY	43
3.6. TRANSPORT INFRASTRUCTURE IN THE COUNTRY	46
4.1 INTRODUCTION <i>→ Chapter Four</i>	47
4.2.KNOWLEDGE ACQUISITION TECHNIQUES	47
4.2.1. TRADITIONAL KNOWLEDGE ACQUISITION	47
4.2.2. AUTOMATED KNOWLEDGE ACQUISITION.....	51
4.3. ANALYSIS OF THE INTERVIEW.....	54
4.4. CONCEPTUAL MODELING.....	56
4.4.1. THE DIRECTION CHOICE	56
4.4.2. LOCATION OF USERS	57
4.4.2. LOCATION OF USERS	58
4.4.3. MEANS OF TRANSPORTATION.....	60
4.5.GEO-REFERENCED MAP OF ADDIS ABABA	65
CHAPTER FIVE	65
KNOWLEDGE REPRESENTATION	65
5.1 INTRODUCTION	66
5.2 THE KNOWLEDGE BASE	68
5.3 THE EXPLANATION FACILITY	69
5.5 MEASURING EXPERT SYSTEM PERFORMANCE	71
5.6 SYSTEM'S CONTROL STRUCTURE	73
CHAPTER SIX.....	73
APPLICATION OF KNOWLEDGE	73
6.1 WEB DESIGN	75
6.2. USER INTERFACE.....	86
6.3. TESTING AND EVALUATION.....	87
6.4. KNOWLEDGE BASE MAINTENANCE.....	87
6.5. DESIGN REQUIREMENTS	88
CONCLUSIONS AND RECOMMENDATIONS	88
7.1 CONCLUSIONS.....	90
7.2 RECOMMENDATIONS	92
BIBLIOGRAPHY	
Annexes	

ABSTRACT

In Africa, though the potentials for tourism is said to be much higher, the number of tourists arrived at the continent and revenue generated from this sector is very low as compared with the other continents. This is because lack of technical person to provide relevant and sufficient information to the tourist is one of the main stumbling blocks for development of tourism in Africa. Ethiopia is no exception to this. Ethiopia having higher potential of tourist sites can promote its tourism industry using expert system technology.

Information makes tourism possible from the very onset of promoting places to making the experience worthwhile. It empowers tourists in their choices and makes them become more comfortable about the experience they are going to have. There lies a big responsibility for making that information accessible in a format that is desirable and easily accessible.

Our country fosters huge potentials in the tourism sector, which if taken advantage of can certainly provide enormous revenues to aid development. The tourism industry, however, lags behind with inadequate infrastructure facilities and manpower rendering the potential benefits unexploited. One of the major problems in this regard relates to access to information for tourists: when they need it and where they need it. Given the circumstance, the question is, Is/are there any areas or possible applications of expert systems that could make this problem easier to overcome?

In an attempt to address such problems, the present study, specifically, explores the potentiality of applying expert systems technology in advising tourist to find the direction of the route(s) in a

convenient and comfortable manner. A Web Based Expert System For Route Finding To The Road Network Of Addis Ababa (**AARFExSys**) is developed by using maps and textbooks in the area of transportation as source of knowledge.

The system is not only meant to replace the tour guides but also to assist them in providing sufficient information regarding routes of Addis Ababa with the users preferred mode of transportation. It helps to avoid the scarcity of manpower, the language barrier and the level of understanding that may face with the experts.

The knowledge acquired for this study is then modeled using hierarchical structure that represents concepts and parameters involved in finding direction. A GIS based software (Arc View) is used as an automated knowledge acquisition tool to extract quantitative data such as distance from the road map of the city. Based on this model, the knowledge is represented using production rules. These rules are then implemented in the Knowledge Pro Gold expert system shell. Back ward reasoning is used in inferring the rules and extracting recommendations.

The needs of users are assessed before the recommendation is extracted so that users are being able to select according to their interest in finding the direction.

The **AARFExSys** has been developed to assist users particularly of tourists, who are coming from different countries of the world. Since these users are scattered through out the world, they need to access this system published on the Internet so that they could be able to get the information wherever possible. That is why the necessity of this system being developed as a web based.

AARFExSys has been tested by users and experienced travel and tour agents at the selected Travel and Tour Agencies found in Addis Ababa. Based on an encouraging result obtained, they have established its acceptability and accuracy.

CHAPTER ONE

1.1. BACK GROUND OF THE STUDY

Experts systems represent computer applications that perform tasks that would be performed by human experts. As automated tools make access to information relatively easier, their application has gained momentum over the years and has also gone in to all arenas, from diagnosing human illnesses, making financial forecasts, to scheduling routes for delivery vehicles (Ignizio, 1991).

Some expert systems are designed to take the place of human experts, while others are designed to aid them. Obtaining knowledge from a human expert and coding it into a form that a computer may apply to similar problems construct expert systems. This reliance on the knowledge of a human domain expert for the system's problem solving strategies is a major feature of expert systems.

Artificial Intelligent in general and Expert system in particular, are receiving increased attention in the area of transport engineering. Knowledge Based systems (Expert systems) are suitable for application to specific problems that are amenable to treatment on the basis of rules and relationships rather than numerical calculations.

1.1.1 KNOWLEDGE-BASED SYSTEM IN TRANSPORTATION SECTOR

Expert system research in the transportation field is comparatively recent. However, significant research contributions have been made, and considerable progress has been achieved in this area. Applications of advanced information technologies in transportation have entered a new era world wide with the activation of several strategic initiatives affecting both road infrastructure and vehicle technologies. Development and application of expert systems to transport and traffic

engineering started around the mid-1980s (Leondes, 2000). Since then a number of contributions have been developed.

In Africa, though the potentials for tourism is said to be much higher, the number of tourists arrived at the continent and revenue generated from this sector is very low as compared with the other continents. Lack of technical person to provide relevant and sufficient information to the tourist in this area is one of the main stumbling blocks for development of tourism in Africa as stated in (Melaku, 1997).

Ethiopia is no exception to this. There is only one training institute called (CTTI)¹ which has so far trained 59 tour guides and travel operators; out of which 23% of them do not work in the tourism sector (Melaku, 1997). However, the number of tour operators and travel agencies in the country is increasing just after 1990 and now altogether they are 109. These tour operators and travel agencies are acting as a Tourism Information Center (TIC) in providing information to the tourists coming to the country.

Ethiopia having higher potential of tourist sites can promote its tourism industry using expert system technology.

In this study an expert system was developed to guide tourists who are planning to visit Addis Ababa in finding the required route(s) and its accompanied cost and time with the available mode of transport.

In light of this expert systems have been developed particularly to aid in the tourism industry. The potentials of expert systems in this sector are enormous given that they are properly designed and implemented. It can work miracles especially in conditions where relative availability of information and/or domain experts is scarce like the case in Ethiopia.

¹ CTTI Catering Tourism Training Institute, the only government owned Training institute accredited by WTO.

2. It allows the use of domain-specific knowledge.

The basic concepts necessary to understand and use expert system are (Yazdani, 1986).

1. The components of knowledge,
2. Methods of inference
3. Methodological guidelines.

Once they are explained in the right way, these concepts can be understood and utilized by the experts.

Expert systems have emerged as the most successful application oriented branches of AI. Over the past several years there have been many implementations of expert systems using various tools and various hardware platforms. The technology has left the confines of the academic world and has spread through many commercial institutions.

This chapter will give a concise account of pertinent concepts in Expert Systems, and in the later chapter application areas of Expert systems in route advisor will be discussed.

2.2 Definition of Expert systems

Various scholars give different definitions for expert systems. Some of these definitions given by scholars are reviewed:

An expert system is a computer program that represents and reasons with knowledge of some specialist subject with a view to solving problems or giving advice (Jackson, 1999).

Expert systems are computer programs that emulate the reasoning process of a human expert or perform in an expert manner in a domain for which no human expert exists (Kendal, 1991).

An expert system is a computer program that relies on Knowledge and reasoning to perform a difficult task usually undertaken only by a human expert (Parsaye & Chignel, 1988).

The appropriate domain of expert system as explained by (Giarratano & Riley, 1998) are stated below:

2.3. Appropriate Domain of an Expert System

Some of the appropriate domain areas of the expert system are indicated as follows:

- Can the problem be effectively solved by conventional programming? If yes Expert system is not the best choice because of ill structured problem.
- Is the Domain well bounded?
- Are there a need and a desire for an expert system?
- Is there at least one human Expert who is willing to cooperate?
- Can the expert explain the knowledge so that it is understandable by the Knowledge engineer? (Technical terms)
- Is the problem solving knowledge mainly heuristic and uncertain?

Expert system paradigm allows two levels of abstraction in a rule based expert system language that is facts (data abstraction) and rules (knowledge abstraction).

There are many ways of characterizing expert systems such as representation of knowledge, forward or backward chaining, support of uncertainty, hypothetical reasoning, explanation facilities, etc.

Apart from the above listed ones there are also three important factors to create the modern rule based expert system. Convergence of three important factors are shown in the figure below

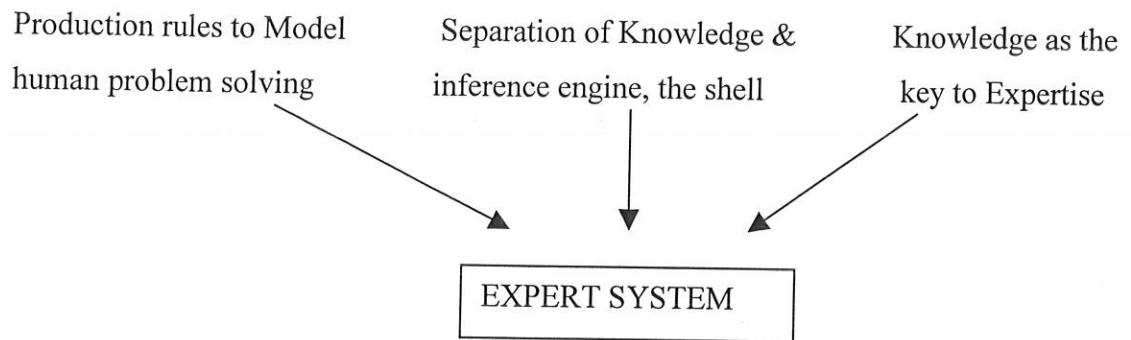


Figure 2.1. **Convergence of the three factors to create the modern rule based expert system**

2.4. Elements of Expert system

In a rule based system the knowledge base contains the domain knowledge needed to solve problems coded in the form of rules. (Giarratano and Riley, 1998) explains that the fundamental elements of expert systems include the following basic ideas.

1. User interface- where user and the expert system communicate.
2. Explanation facility –explain the reasoning of the system to a user
3. Working memory – a global data base of facts used by the rules
4. Inference engine – makes inferences by deciding which rules are satisfied by facts or objects, prioritizes the satisfied rules and executes rules.

5. Agenda – a prioritized list of rules created by the inference engine, whose patterns are satisfied by facts or objects in working memory.
6. Knowledge acquisition facility- an automatic for the user to enter knowledge in the system instead of having the knowledge engineer explicitly codes the knowledge.

2.5. The Basic Advantage and Characteristics of an Expert System

The basic advantages and characteristics of an expert system in general can be described as (Giarratano and Riley, 1998) in the following:

Advantage

- **Increased availability**-expertise is available on any suitable computer hardware.
- **Reduced cost**- the cost of providing expertise per user is greatly lowered.
- **Reduced danger** – expert system can be used in environment that can be hazardous for human.
- **Permanence**- the expertise is permanent. Unlike human experts, who may retire, quit or die, the expert system's knowledge will last indefinitely.
- **Multiple expertises** can be made available to work simultaneously and continuously on a problem at any time of day or night. The level of expertise combined from several experts may exceed that of a single human expert.
- **Increased reliability**-expert system increase confidence that the correct decision was made by providing a second opinion to a human opinion to a human expert.
- **Explanation** –it can explicitly explain in detail the reasoning that led to a conclusion.
- **Fast response** – depending on the software and hardware used some emergency situation may require responses faster than a human expert.
- **Steady, unemotional, and complete response** at all times when a human expert may not operate at peak efficiency because of stress or fatigue.

- **Intelligent database** – expert system can be used to access a database in an intelligent manner.

Characteristics

An expert system is usually designed to have the following general characteristics:

- **High performance** – the system must be capable of responding at a level of competency equal to or better than that of an expert in the field.
- **Adequate response time** – the system must also perform in a reasonable amount of time, comparable to or better than the time require by the expert.
- **Good reliability**- reliable and not prone to crashes or it will not be used.
- **Understandable** – the system should be able to explain the steps of reasoning while executing so that it is understandable. Rather than being just a black box that produces a miraculous answer, the system should have an explanation capability in the same way that human experts can explain their reasoning.

2.6. Rule Based Expert System

Today rule based production systems are a popular method of implementing expert systems. The individual rules that comprise a production system are the production rules.

In designing an expert system an important factor is the amount of knowledge or granularity of the rules as indicated by (Giarratano and Riley, 1998).

Structure of a rule based expert system

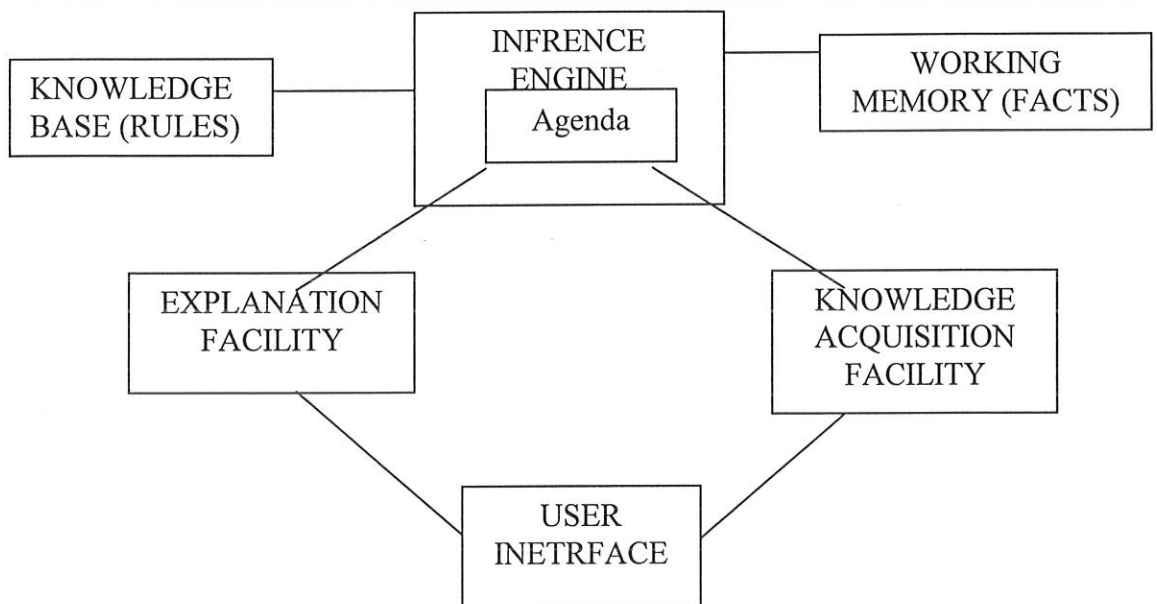


Figure 2.2. General structure of a modern rule based system.

2.6.1. The production rule system

A simple example of a rule-based system in a transportation area that can be represented according to (Giarratano and Riley, 1998) is shown as:

The production rule system

The light is red $\longrightarrow$ Stop

The light is green $\longrightarrow$ go

The equivalent pseudo code If ----THEN format is

Rule: Red_Light

IF

The light is red

THEN

Rule: Green_Light

IF

The light is green

THEN

Go

Stop

One of the major features of an expert system capability is its ability of inferring knowledge. The reasoning capability of an expert system is the most important one of the other features.

2.6.2. The Inference Engine

Inference is the process of deploying evidence in order to arrive at new conclusions. An inference mechanism consists of search and reasoning methods that enable the system to find solutions and,

if necessary, provide justifications for its answers. There are two overall reasoning strategies (Yazdani, 1986), forward chaining and backward chaining.

1. Forward chaining- reasoning from fact to the conclusion resulting from those facts or working forward from the evidence (symptoms) to the conclusions (diagnosis). In a rule based system it simply involves matching the IF condition to the facts, possibly in a Pre determined order. It is easy to computerize, and is suitable in cases where all the data is gathered anyway. (Yazdani, 1986)

E.g. If you see that it is raining before leaving home (the facts) then you should take an umbrella (the conclusion).

2. Backward chaining- reasoning in reverse from a hypothesis, a potential conclusion to be proved, to the fact that support hypothesis. The system picks a hypothesis and looks for data to support or refute it. It can be programmed in a recursive manner, and typically leads to a more natural kind of dialogue in consultation-style systems. (Yazdani, 1986)

E.g. If you are not look outside and someone enters with wet shoes and an umbrella your hypothesis is that it is raining.

Most systems use a mixture of forward and backward chaining. For example Naylor (1983) as quoted by Yazdani has developed the 'Rule Value' method, which decides what evidence to ask for next on the basis of how much it will reduce the overall uncertainty in the system.

Many expert systems allow for uncertainty in the facts obtained from the user and in the rule themselves.

This study uses a backward reasoning technique among the two.

2.6.3. Uncertainty

In most fields of science, like agriculture and medicine, knowledge is for the most part in exact and uncertain. People make decision in an environment where facts and rules make various shades of vagueness, imprecision and errors. A number of methods have been proposed to super impose uncertainty over the logical structure of expert systems (IPT, 2000).

- **Probability**- in terms of probability, probability is usually expressed as a chance of some particular state occurring as a percentage of population of all possible states.
- **Statistical uncertainty**- calculating the behavior of, and the certainty of various results statistically by examining past behavior (personal judgment).
- **Heuristic uncertainty** – uncertainty derived from the heuristic knowledge, which is less vigorous, more experimental and more judgmental knowledge of performance. Such knowledge is expressed by using rules of thumb or words like very likely, high, low, etc. Heuristic Knowledge underlies uncertainty, because it is expressed as “the art of good guessing”.

Representing uncertainty is one of the most controversial subjects in expert systems. Expert systems differ from conventional software because of their human like reasoning. There fore, it seems possible to expect the expert systems should be capable of handling uncertainty.

2.6.4. Different Methods of Handling Uncertainty

In handling uncertainty, there are a number of methods used. These methods and their differences are briefly discussed below:

Probability approaches

- Subjective (Bayesian) probability
- Objective probability

Dampster-Shafer (D-S) - a calculus for numerical degrees of beliefs.

Road Segment Weight = *Cost*

Total Attraction Value

Where *total Attraction Value* is calculated in the previous steps and *Cost* can be distance, time or other cost factors. With this weight attribute, the shortest path function is more likely to select the road segments that have higher attraction and lower cost value.

EVALUATION

Ideally, an evaluation of such a system should be carried out in a real tourist environment, in which the testers would be asked to follow several routes recommended by the software.

However, due to resource constraints, mainly the time required to follow the routes, a simulated test environment has been designed for use on computers.

In this simulated test environment, testers enter a weight for each attribute of the tourist attractions. The software then generates four tourist routes—one is generated based on the weights given by the tester, *i.e.*, the “original” preference, and the other three are generated based on the weights adjusted at different levels, *i.e.*, “noisy” preferences, from the weights given by the tester. The objective is to find out if the testers prefer the routes generated based on their “original” preferences to the others that were derived using “noisy” preferences.

In total, forty-five test cases were generated. Based on the analysis of these forty-five cases, it can be concluded that the preference of the routes by the testers is quite consistent with the weight adjustment given to each route by the software, *i.e.* that the bigger the difference of the weight adjustment is, the less likely the route is favored by the testers.

Another study From the Chinese University of Hong Kong, which is published in the Proceedings of the 38th Hawaii International Conference on System Sciences-2005, is A Multi-Modal Agent Based Mobile route Advisory System (MIRAS) for Public Transport Network.

The study emphasizes on route finding in a metropolis such as Hong Kong, which *is* massive, dynamic, and complicated, and therefore route finding is not an easy task, especially when routes comprising several modes of transport vehicles (such as bus, train, and ferry). This problem is even more important for e-tourism. Tourists and moving workforces, who may need to visit an unfamiliar part of the metropolis.

The study proposes several intelligent approaches in the search agents to enhance the performance of route finding, in particular a knowledge-basket approach. Backend information agents gather updates periodically from transportation companies. It built a route advisory system prototype with such features.

The contributions of this paper are as follows: (i) conceptual modeling for the multi-modal route finding problem; (ii) enhanced methodology including a variety of heuristics for route finding with tariff optimization; (iii) integrated support of public and private transportation advices; (iv) agent-based management of individual vehicle services.

The methodology used for tackling this problem is focusing on some effective design of agents. It starts with a route-based Dijkstra's algorithm and then, introduces various heuristics for speeding up, and in particular, Knowledge Basket approach. Thus introduce three main heuristics – search area limit, cost bound, and knowledge based finding.

Search Area Limit – A common heuristic is to limit the search area for route finding. When finding routes between two locations in a network, it does not make sense to search areas well outside an area enclosing the optimal route. In other words, the domain filter agent should restrict the search to an area within which it is likely that the optimal route will be found.

Cost Bound - The algorithm sets a cost bound in finding routes. If routes exceed the cost bound, the agent discards the routes. Initially, the cost bound is set to a small value. Then, it is progressively set larger and larger until answers are found.

Knowledge Based Finding - If there is no direct route between two points, the search agent attempts to find the nearest nodes step by step. It is possible to speed up the process if some unnecessary nodes, edges, or routes are pruned. Based on knowledge-based information, even if the search domain is reduced, the result found is the same as that in the original search domain.

Knowledge Basket

In order to speed up the search agent, it uses a knowledge-based approach by introducing a Knowledge Basket. In brief, the Knowledge Basket captures useful information for route finding beforehand. Whenever there is a similar search again, the agent can make use of the captured knowledge. Based on the knowledge, the agent prunes plenty of unnecessary nodes before and during the route finding algorithm.

The Implementation of MIRAS is users can either access MIRAS through Web interface of a browser or through WAP (Wireless Access Protocol) on a mobile phone.

The above two studies are concerned on an intelligent based technique mainly used search techniques of heuristic search and for optimum shortest path algorithm both methods use Dijkstra algorithm.

The advantage of using expert system over heuristic is that according to (Ignizio, 1991) expert system has imprecise relationships, transparent, relatively structured development process, revision is relatively easy, and conclusion is often multiple (with explanations)

So unlike heuristic search techniques this study is conducted using the generic expert system of rule based approach to tackle the problems specified earlier in the previous chapter of the study because it has the above stated advantage over other techniques.

2.8. Tour operators and Travel agencies in Ethiopia

The tourism industry is in general a smokeless industry that needs proper organization and enhancement in order to utilize its resources of various tourist attraction sites available. Basically tourism industry of Ethiopia composed of two major organizations, which are directly involved in the business activities of the sector (Ayalew, 1992)

1. Travel agency and Tour operators
2. Hotel/Catering and related hospitality

Travel agency and tour operation sector are part of the tourism industry activities that need expertise in guiding tourists. They are the one who have a day-to-day contact with tourists and provide relevant information timely and accurately.

In Ethiopia, both government and private organization run the business of travel agency and tour operation. Tourists' who came to visit needs sufficient information about the places they are planning to visit and needs to commute from one place to the other in a comfortable manner. So it is tour operators who are responsible for providing all this information. They have some limitations to provide efficient services and are not sufficient enough in number as compared to the number of tourists who plans to visit Ethiopia. An expert system designed to assist and replace and assist those tour operators (in places where they are not available) greatly help to promote the tourism industry.

2.9. The advantage of World Wide Web in Tourism promotion

One of the various methods of promoting the tourism industry is the use of web technology integration to the system. So the **www** is an important tool for this reason as explained by (Steiner, 1995):

The **www** is an Internet wide distributed hypertext or hyper media system that operates on a client /server mode. **WWW** based information servers provide the user with a multi media interface for access to text pictures and sounds.

A simple ASCII based language called HTML is used to construct files on the server, which display on the user's computer www client software usually called web browser. The main purpose of promotion is to make sure that the identity and the public image match. There are many things that make up identity. According to (Steiner, 1995) when promoting a tourist site, identity includes location, types of businesses in the destination, lodging and dining facilities, destination appearance, recreation activities, unique features, unique historical events, community hospitality, etc. It is impossible to promote effectively every characteristics of the identity one has to select which characteristics to promote. So a web based expert system towards achieving the goal of tourists' route finding or route advisory is an important issue for the countries tourism promotion activities.

CHAPTER THREE

GENERAL OVERVIEW OF THE ADDIS ABABA CITY AND ROADS

3.1. ESTABLISHMENT OF THE CITY

Ethiopia is situated in the Horn of Africa where it is bordered by Djibouti, Eritrea, Sudan, Kenya and Somalia. The population is estimated to be over 70 million. Most of the population resides in the cities as a result of day-to-day migration to wards the city from rural areas because of various economic, social and other related problems.

Addis Ababa, established in 1886 by emperor Menelik II and his wife Taitu, has experienced several planning changes that have influenced its physical and social growth. The city's early development centered around three nodal points: the Menelik palace (political and administrative center), the Arada st. George church (social and religion center), and the Arada area (business and market center) (Addis 21,2000).

The first plan prepared during the period of 1974 was that of Professor C.K. Polony, a Hungarian planner, Polony designed the "Meskel Square"(previously known as "Revolution square") area. His work was one of the most important spatial transformations implemented during the socialist regime according to the study by (ORAAMP, 2002).

3.2. STANDARD AND DEFINITION OF ROADS

Road are standardized according to their level of traffic volume; services required for different models of transport, length and width of the road, type of connection and land use (activity) that they are going to serve. Basically hierarchy refers to the level of importance of the street.

Accordingly, the roads treated in the revised Addis Ababa Master Plan (AAMP) are classified under the following broad categories as principal arterial streets, sub- arterial streets, collector streets and local streets (ORAAMP, 2002).

The city cultural centers should express the rich cultural diversity and coexistence of multi-nations and nationalities of the Ethiopian people. Higher order services of international standard (import/export, trade fares, transport, hotel and tourism, health, educational and recreational facilities, etc.) are envisaged in correspondence with the role of the city.

The current efforts aimed at developing high ways and airports should be encouraged to facilitate the interaction between regional capitals and Addis Ababa. The tourist values of Axum, Lalibela, Gondar, Bahirdar, the Semien Mountains, Sofomar etc., as well as the agro-industrial processing and export potentials from Oromia (Jimma) and Southern People Nations and Nationalities (SPNN) Awassa should be enhanced. Addis Ababa should also foster capital/technology intensive industries supportive to regional manufacturing industries.

The latest master plan for Addis Ababa, meant to serve the city for 20 years, was prepared by the Addis Ababa. master plan project through Ethio- Italian technical cooperation during 1984-1986. This plan was broader in scope and content than the previous ones, and was based on Polony's approach that gave considerable attention to the regional, metropolitan and urban hierarchy.

Although the technical aspects of this master plan were properly worked out, its inability to function as an effective development guide in the current socio-economic order has necessitated a comprehensive revision study of it as indicated in (ORAAMP, 2002). .

The Addis Ababa city roads at present are called by the name of the 54 African countries, which indicates that the city is not only the capital of Ethiopia but also of an African. The naming of the city roads is very important for this study, though not comprehensive, to locate the exact route from one location to the other in the process of commuting activity.

3.3. ADDIS ABABA CITY ROAD NETWORK

One important factor, for a city to have efficient transportation activities, is its road network. The present road network in Addis Ababa in principle couldn't satisfy the needs of the city for some years to come. Moreover, road capacity is badly constrained by pedestrians walking in the car lanes, unorganized driving of taxis. According to the information found on the publication of the Addis Ababa City Government (Addis 21, 2000):

The efficiency of the present road network is hindered by:

- Missing links in the inner city,
- Increasing pressure on the southern corridor (The city's link to the southern part of the country, access line to most industries as well as ports);
- Expansion areas are not equipped with necessary roads, resulting in poor accessibility
- Lack of continuity and proper maintenance of existing roads.

Another study (ORAAMP, 2002) on the problem of the Addis Ababa city network system indicated that most of the challenges are coupled with old and narrow bridges, inadequate

- Manual digitizing most widely used method for entering spatial data from maps. It uses a digitizing table and mouse with a cursor to trace and record points, lines and polygons needed for a particular data set. It is used when existing maps are available.
- Scanning solutions to draw backs of manual digitizing, used for entering map and photo information.
- Remote sensing in map production.
- Global Positioning System (GPS) for data input and uses satellites, which transmit signals that can be decoded by specially designed receiver to determine positions precisely (Antennuci, 1990). Data for GPS's can be used in association with existing spatial database or GIS, both for update and for a range of applications in spatial decision making (Malczewski, 1999).

Data outputs

Components of GIS provide a way to see the data or information in the form of a map, tables, diagrams, and so on.

Data storage and management

It incorporate those functions concerned with storage and retrieval of data from the database.

- Nature of geographic data has four major components: its geographic position, its attributes, its spatial relationships, and time (Arnoff, 1989).
- Geographic position basically a form of spatial data. Each feature has a location that must be specified in a unique way. Locations are recorded in geographic data in terms of a coordinate system. Like the Latitude/Longitude, state plane coordinate systems. In some cases the coordinates of one system can be mathematically transformed in to the coordinates of the other.

- Attributes also called non-spatial attribute to indicate that they do not themselves represent location information. Representing character, qualities or relationship of map features and geographic location.
- Spatial relationships are numerous or may be complex and are important. For a computer based GIS; relationship must be expressed in a computer usable manner.
- Time GIS is referenced to a point in time or a period of time.

3.5. EXPERT SYSTEMS AND GIS TO THE CONTRIBUTION OF TOURISM INDUSTRY

The Geographic Information System and expert system technologies extend the possibilities of managing the road network in the provision of digital information, which is a computer-based tool to capture, manipulate process and display spatial or geo-referenced data. They contain both geometry data (co-ordinates and topological information) and attribute data, (information describing the properties of geometrical objects such as points, lines, and polygons). Expert systems are one of the technologies that enable to make use of expert knowledge to solve problems. The knowledge is encoded in a computer program in the form of IF...THEN rules. The new generation expert system combines: rule-based and object-oriented knowledge representation, message passing, GUI and the ability to integrate with other systems. Usually, GIS enables us to analyze quantitative spatial information, while expert systems enable us to operate qualitative information. In qualitative reasoning, a situation is described by parameters that have a rather small-predefined set of values, and by rules that use these symbolic values instead of numeric values. Linking an expert system with a GIS enables these both system to

perform new tasks, and opens the way for more flexible and complex analysis of spatial data, based on rules and logical inference.

It is proposed to use rule-based knowledge for more intelligent visualization of maps. According to (Arnoff,1989), an intelligent GIS should assist the user in the analysis, and this presupposes the following capabilities: the capability to understand user's information-seeking goals, the capability to select and visualize appropriate data in a way productive for achieving these goals, and the capability to support the user's analytical activity with the use of the generated presentations.

Depending on the problem addressed, on uses and users, various levels of integration are possible, ranging from the simple exchange of data files to a complete integration within one common environment, framework, and user interface.

Recent GIS implementations have begun to support the use of client-server techniques as an effective means of transferring data while retaining data structure. This has presented the possibility of linking other systems with GIS by using these techniques.

The expert system shell makes inferences using knowledge stored in knowledge bases. This architecture enables to implement specialized GIS applications that use an appropriate part of GIS services and the expert system shell with appropriate knowledge bases. Filling the expert system shell with knowledge makes it an expert system in a certain field. By replacing knowledge bases, we get another expert system, which uses the same knowledge representation methods and inference engine.

In this study the application of GIS is used in the process of knowledge acquisition process from the city map in order to georeference the map and find out the quantitative data automatically using Arc View GIS software such as the distance from one place to the other. This has been discussed in the next chapter in detail.

3.6. TRANSPORT INFRASTRUCTURE IN THE COUNTRY

Now a day the application of GIS and expert system technology greatly helps to provide digital information to the transport sector. As indicated in the previous chapter some European countries and Japan use such an integrated system in this sector to help the development of a tourism industry.

Transport infrastructure is one of the most important infrastructures for the development of a country in general and the tourism sector in particular. When tourists plan to visit some attraction sites, they need to know whether they comfortably commute with in the country or the city so that they can find the exact location of those sites. Tourism is absolutely dependent upon the infrastructure (road, hotel, etc) a country has, meaning these activities are highly dependent on the means of transportation. Transport, as mentioned above, is an integral part of tourism, which seeks high amount of investment and technologies. Inadequate transport facilities in a country will certainly affect the development of tourism even with amazing natural aesthetic value the country has.

Ethiopia, like most of the developing countries, faces big problems in developing its infrastructure especially in the transportation sector.

Even though all means of transport exist, namely, land, air and water (lake) transport they are inadequate and not up to the standard. For this study the focus is on land transport.

Land Transport

It is the most suitable means of transport for tourists because it provides a chance of on-way touring opportunity, if possible by stopping the vehicle easily.

One of the measures of the sufficiency of roads in a particular country is its density in relation to the inhabitants and area. In Ethiopia, the available density of roads is only about 0.9 km per 100 inhabitants and the density per 100km² is 31 km for all types of roads. Both figures are below the

average for Africa, where the weighted means are 2.6 km and 58 km respectively (WTO, 1990, 122).

There are also roads which connect Ethiopia with other African countries there by promotes an inter-regional tourism. For example the Moyale high way connects Addis Ababa with Nairobi, dry weather roads link Addis Ababa with Djibouti, Khartoum and Hargessa in Somalia. Gabrone-Cairo high way connects Ethiopia to the trans African high way through Gondar, Bahir Dar, Addis Ababa or in the south through Awassa to Moyale and finally to Kenya (Ayalew, 1992,21).

Similarly the Addis Ababa city road link is constructed in various directions from one location to the other using the ring road by the Chinese construction Engineers. The overall density with in the built up boundary of the city is 9% and density of roads in densely built up areas that covered 1800 hectare is 0.1522 that is 15.22%of the total land area of the city is used for street network (ORAAMP, 2002). The general road network of the city should have necessary hierarchy in responding to the future traffic pressure that is necessity to develop high-speed (free way) roads and important boulevards. In this respect, the areas on the other side of the ring road (phase I and II) are provided by important lines of high-speed ways being connected to the ring road according to the information found from ORAAMP study document on 2002.

Next to the high-speed lines, there are important boulevards with width 40 m and more. These are as indicated in (ORAAMP, 2002):

- The East –west axis (Ayat mender to Army hospital)
- Lideta to Mercato regional bus station
- Regional bus station to Piassa and Sidist kilo
- Meskel square to Shiro meda
- Meskel square to bole Airport
- Piassa to Legare (Churchill street)

- Gotera to Akaki (Tulu Dimtu)
- Kara Alo to Bole Lemi
- Jemo ring road (round about through Buri to Jimma road).

CHAPTER FOUR

KNOWLEDGE ACQUISITION

4.1 INTRODUCTION

Knowledge acquisition is the process of capturing knowledge and expertise. Expertise is fairly easy to recognize, although it may be difficult to define. First, expertise is inherently task-specific. Most of the expert's knowledge cannot be transferred between different domains, and the expert must learn about each new domain almost from scratch. Second, expertise is relatively inflexible so that the experts may not be able to cope with new situations that do not conform to their past experience. Third, expertise tends to focus the reasoning of the expert so that only one type of solution is considered. (Parsaye and Chignell, 1988).

It's the most important step in the development of a Knowledge Based System but at the same time the most painstaking job.

(Parsaye and Chignell, 1988) explains that knowledge acquisition is the bottleneck in expert system construction. Since expert systems rely heavily on knowledge, the quality of knowledge acquired for an expert system will often be the determining factor in its success.

So building an expert system is not, however, simply a matter of selecting an expert system shell, and obtaining a set of facts, rules and frames and transforming them into the syntax of the knowledge representation language. In most cases the facts, rules, and frame structures are not initially available or well identified, and must be sought out.

The process of seeking out the knowledge required by an expert system is referred to as Knowledge acquisition. The knowledge must be acquired either through a process of knowledge

transfer from one or more agents or experts, who already possess the expertise, or through some type of learning process that observes examples and distills knowledge.

Since expert systems rely heavily on knowledge. The quality of knowledge often determines the success of an expert system. Thus the ratio of effort expended to results achieved for the expert system as a whole is often decided by the knowledge acquisition process and the level of success achieved there.

In this chapter the study discuss the manual approach such as interviewing and then focus on automated knowledge acquisition using GIS software.

4.2.KNOWLEDGE ACQUISITION TECHNIQUES

There are generally two methods of knowledge acquisition process in modern expert system.

4.2.1. TRADITIONAL KNOWLEDGE ACQUISITION

An interview question is designed to gather information which can assist to select the most frequently visited locations of Addis Ababa and questions for the most experienced tour agents to gather information regarding the location of the selected routes of Addis Ababa as well as tourists. Structured type of interview is prepared and asked for six travel agents in six travel agencies randomly selected from government and non-government organizations. The result is analyzed and used in the representation of rules in the next chapter.

4.2.2. AUTOMATED KNOWLEDGE ACQUISITION

One disadvantage of manual methods such as interviewing is that they consume large portions of the time of people who are in short supply, namely, experts and experienced developers. For this an automated knowledge acquisition method is an important part of an expert system study. According to (Parsaye and Chignell, 1988) to automate knowledge acquisition process, one need to provide tools that:

- Helps experts without a knowledge engineer capture their own expertise
- Allow knowledge engineer to capture knowledge more effectively in very complex applications
- Provide methods of automatic induction that capture knowledge usually unobtainable without computer assistance.

In this study an automated knowledge acquisition is used apart from the traditional method known as interviewing. The automated tool used to capture knowledge as explained earlier in chapter three, is one of the most popular GIS software called Arc View 3.1 version. This tool used to convert the existing map of the Addis Ababa road map in to the database. Then the database will have the information about the selected sample locations from the road map, since this is a prototype study, the source and destination of the location, the means of transportation, the costs, etc. This way captured knowledge can easily be represented using expert system shell.

The design and prototype will demonstrate the application of the expert system by involving a limited number of initial points, destinations and transportation options. It will not provide a full-fledged system nor will it attempt to make generalizations that such systems are applicable to all such conditions. The assumptions on which the prototype is developed are:

Users of the Expert System are assuming guests (Tourists) in one of the following six hotels:

- Axum Hotel / Classic Hotel
- Ethiopia Hotel / Harambe Hotel
- Meridian Hotel
- Sheraton Hotel
- Hilton Hotel
- Wabe Shebelle
- The destination that the System can give direction on are:

- To the Bole Airport
- To Meskel Square
- To Megenagna (Adwa Square)
- The possible means of transportation to either of these destinations are:
 - Bus
 - Drive
 - Taxi
 - Walking

The study will be developing a direction/route finder with transport options that will be able to tell tourists how to get to specific locations through their preferred means of transport.

Having identified the domain, the next process was the knowledge acquisition stage where the study tries to gain as much insight into the conditions surrounding the development and functionality of the expert system. The study tries to build up on the available road map of Addis Ababa using GIS software, and in addition to this a good number of travel agents knowledge and my knowledge of the city routes and available travel options. The knowledge that needs to be captured by the system was, therefore, taken also from my personal experiences as residents of the city. By doing so, the study try to adopt the possible knowledge acquisition option where the knowledge engineers serve as the domain experts as well and to acquire the Knowledge of the experts for the domain, interview and planned discussion will be conducted with the expert in tourism commission and different Travel and Tour Agents.

Inclusion of an explanation facility was deemed necessary in this particular case though the questions being asked of the user more or less represented clear indications as to why the questions were being asked. Similarly, the questions are very few in number that is they are 3 (three) hence given the objective of the user in using the system.

The general queries the user has to answer to be directed from a particular hotel to a specific location are:

1. Where would you like to go?
2. Where are you now?
3. What means of transportation are you planning to use to get there?

If, for instance, the user chooses to travel by 'Taxi' among the various modes of transportation, then he/she is required to specify the following:

- Regular mini bus
- Contract

Should they choose to travel with a contract taxi, users will be further prompted to select their options from:

- Regular (Blue cabs)
- Hotel (Yellow cabs)

Having identified the domain, the next process was the knowledge acquisition stage where the study tried to gain as much insight into the conditions surrounding the development and functionality of the expert system. I tried to build up on my knowledge and the available road map of the city routes and available travel options. The knowledge that needs to be captured by the system was, therefore, taken from the map through GIS software and my personal experiences as residents of the city (that is, partially I am the expert in this area in capturing the required knowledge) and the expertise of different experienced tour agents. By doing so, the study tried to adopt the possible knowledge acquisition option where the knowledge engineers serve as the domain experts as well. The study also further consulted information brochures

regarding bus routes, distance factors surrounding the transport means, and available price standards. One of the steps followed in capturing somewhat complete information on each direction include the typical approach of listing all possible replies to a given query.

4.3. ANALYSIS OF THE INTERVIEW

The interview is conducted with the selected Travel and Tour agents found in the city and some six tourists who came to visit the city of Addis Ababa.

- **Interview conducted with Travel and Tour Agents**

The interview questions were conducted with the following tour and travel agencies selected randomly found in the city. Purposive sampling technique is used to select these agencies in Addis Ababa because of their experienced staff and willingness to give response:

1. Galaxy Express Travel and Tour Specialists

2. Ghion Travel and Tour

3. Experience Ethiopia Travel

4. Lalibela Travel and Tour

5. Rainbow Travel

6. Queen of Sheba Travel and Tour

7. Blen Ethiopia Travel Agency

8. Abyssinian Travel and Tour

The results found from the interview questions asked to eight travel and tour agencies tour guides shows the following key points for this study.

For the questions asked to explain for how long they guides tourists in this specific aspect describes that most of them 60% guides for 7-10 years and the rest 40% are 3-7 years. So almost all of them selected for this study have a good number of experiences in tour guide and knows the interests of tourists since they have sufficient experience in this area.

Another question asked to assess how tourists are assisted to get the location they require by tour guides explains that most tour guides use their experience about the city roads, some use guides and map of Addis Ababa roads. But they explain that they face many difficulties to tell the exact location through the possible modes of transportation because of the cities complex road network and master plan.

Further more they have few information about the various city roads, the distance from one location to the other, the cost and time it takes. Since most of them depend on their personal experience they also lack sufficient information about the roads and its surrounding environments such as recreational centers and other service centers.

According to the question asked on if they need any computer assisted tool to help them find the location required with related description they respond that 75% of them were positive and the rest 25% was negative because their ability to use computers and have less understanding.

To find out the most frequently visited hotels and the along side roads by tourists to develop a prototype system, tour guides were asked similar question related to this and describe that most frequently visited roads and hotels are from Bole Airport to Meskel Square, from Meskel Square to Megenagna (Adwa square) and the hotels in this road and in addition to this hotels some surrounding hotels like Sheraton Addis, Hilton hotel, Ethiopia hotel and Wabe shebelle hotel are most visited ones.

To assess the interests of tourists in finding the routes they require, a question was asked and several responses were found. That is some of them need the shortest route that takes less time

and others require the longest route with a lot of opportunities to get several hotels, recreational sites, etc. While others need the route that has good historical value regardless of the distance and the time it takes. Others rely on the less traffic route than the other factors mentioned above and some of them on the comfort ability of the road to drive and walk. So this shows that tourists have diverse interest in finding the location in their commuting activity with in Addis Ababa and they need such diverse information to satisfy their needs.

Besides their diverse interest they also need effective communication with guides so in order to identify the communication ability of the guides, guides were asked a question on whether they have barriers to communicate effectively with tourists coming to Ethiopia .The reply was since tourists coming from different countries, their knowledge of the English language and their accent also varies so this create much problem to communicate effectively and provide sufficient information considering also the limited knowledge of guides.

To assess the English Language ability of tour guides another interview question were asked to them and the reply shows that most of them have moderate language skills on the basis of Very good listening ability but good speaking (Expressing) ability. Almost all of them are a graduate from the tourism and catering institute as a tour guide.

- **Interview conducted with tourists**

The results found from the interview questions asked how they are assisted to get the location they require by tour guides as explained by tourists that most tour guides use their experience about the city roads, some use guides and map of Addis Ababa roads. But they explain that they face many difficulties to tell the exact location through the possible modes of transportation because of the cities complex road network and master plan.

Further more tourists clarifies that tour guides have few information about the various city roads, the distance from one location to the other, the cost and time it takes since they mostly depend on their personal experience.

To assess their interests in finding the routes they require, tourists were asked a question about their interest and several responses were found. That is some of them need the shortest route that takes less time and others require the longest route with a lot of opportunities to get several hotels, recreational sites, etc. While others need the route that has good historical value regardless of the distance and the time it takes. Others rely on the less traffic route than the other factors mentioned above and some of them on the comfort ability of the road to drive and walk. So this shows that tourists have diverse interest in finding the location in their commuting activity with in Addis Ababa and they need such diverse information to satisfy their needs.

Tourists also identified that the language skills of the tour guides they found are in a moderate level. That is, most of them are better in listening and understanding and some of them are good in speaking or expressing ideas.

According to the question asked on if they need any computer assisted tool to find the location required with related description that the response was 100%positive and they explain they have such experience in other countries in using computer assisted tool in this respect.

4.4. CONCEPTUAL MODELING

In order to formalize the knowledge acquired, The study then progressed to modeling trees using a top-down approach of hierarchical structure. The developed model is used as a general framework to design the knowledge base. In effect, set of interrelated tree structures were built to assist the task of representing the knowledge base.

The attributes needed to solve the problem were then determined and strategies were devised to acquire the values of these attributes as inputs from the user. The attribute values used in defining the rules are:

- ‘Destination’: value of the place where the user wishes to go

- 'Location': value of the place where the user wishes to travel from
- 'Transportation': value of the means of transportation the user intends to use
- 'Class': value of the type of taxis (contract or regular)
- 'Type': value of the type of contract taxis (yellow hotel taxi or regular blue taxi).

Taxis are available in cities and major towns. These include the blue-and-white mini buses, pickups with a closed Canopy or Saloon Cars. They normally operate on a fixed rate and standard route. Personalized and specific trips should be negotiated with the driver in advance. In the Capital, there are Luxury Mercedes Benz taxis operated by National Tour Operation (NTO). They are stationed outside major hotels and at the airport. They do not have meters. In addition, there are yellow taxis stationed at the airport for special hire, which necessitates a prior negotiation with the driver for the price.

Towards this end, simple input lines of three attributes were determined. They are:

1. The end user would be asked where s/he wants to go.
2. The end user would be prompted where he is.
3. The end user will be required to input his/her preferred mode of transportation to that specific location.

After having generated decision trees, formalizing the knowledge into a knowledge representation scheme was the next task. Considering the kind of system one aim to develop and representation ease by the rule, the researcher found rule-based mode of knowledge representation more appropriate for this purpose. Accordingly, 76 If-Then rules were extracted based on the decision trees. (See Annex 1)

4.4.2. LOCATION OF USERS

The location of users can be taken as some of the selected areas or hotel sites that can be considered as a sample to this study. These places are selected because of their convenience to various tourists according to the information found from the tourism commission.

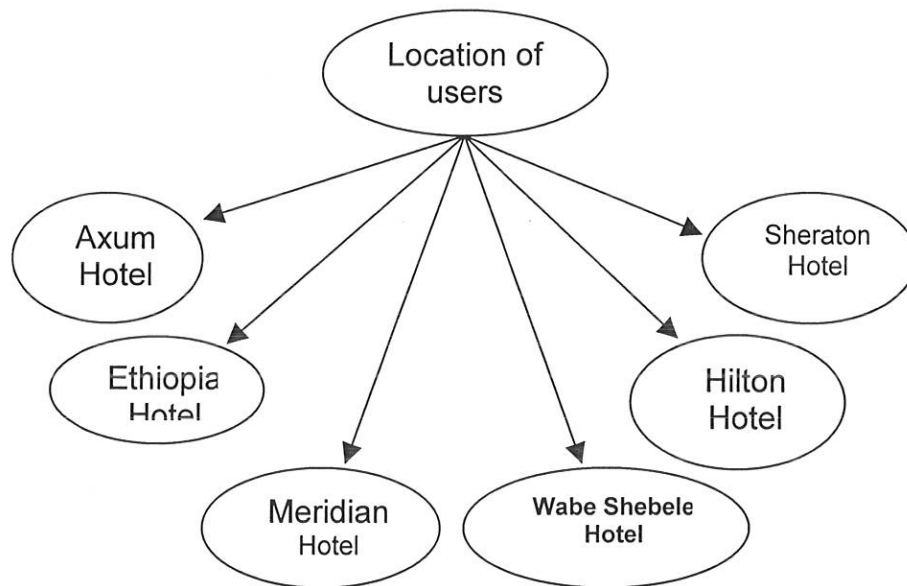


Figure 4.2. Concepts in Users' Location

4.4.3. MEANS OF TRANSPORTATION

The means of transportation which the user prefers is described in the figure below shows all the available transport options of the city is considered except the city buses because it is believed that this study is meant for tourists, it is not convenient for them to use the city bus to commute with in the city.

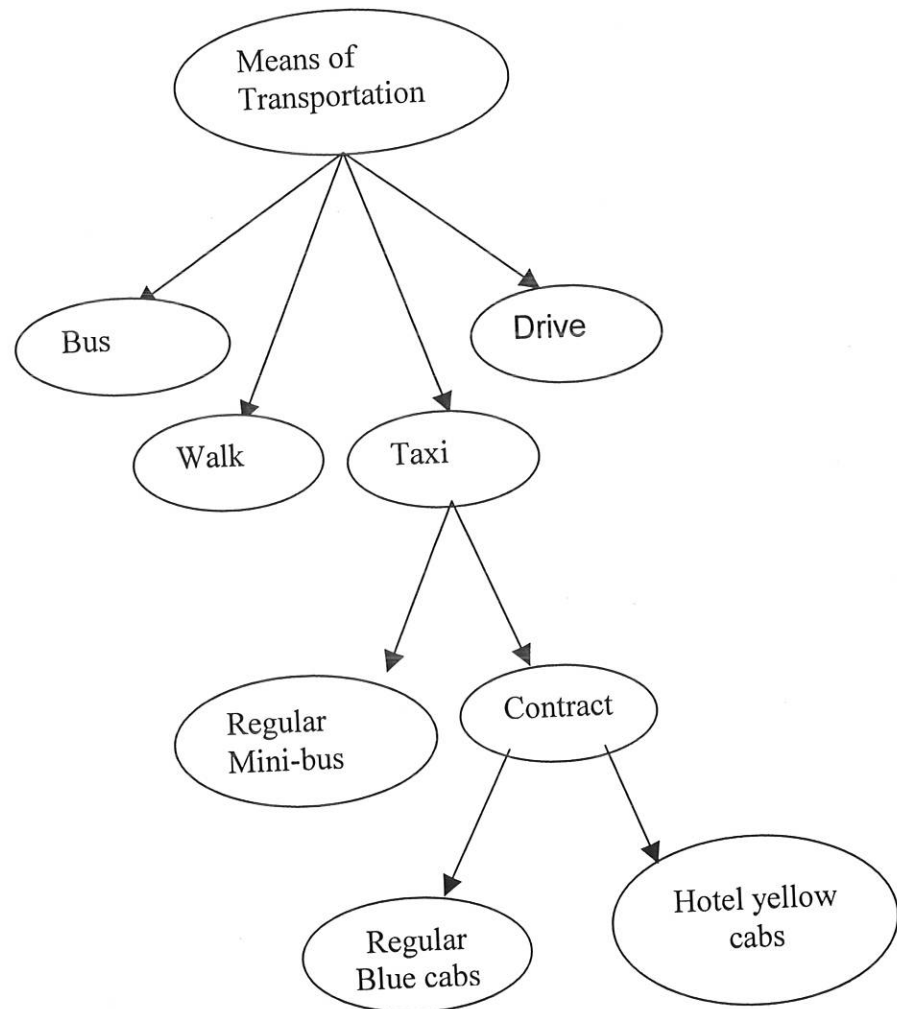


Figure 4.3. Concepts in Means of Transportation

The general approach to be taken in mapping the rules to the inference engine would be a backward chaining. In such an approach, one can obtain optimal standards of being able to combine the two approaches with emphasis given to forward chaining in this case. The researcher also understand that these algorithms by themselves don't guarantee a good system and the fact that any results the study likely to obtain will be functions of the standard of rules developed in the context.

Following the inference engine, another important concern becomes the design of the interface through which users will be able to interact with the system. For this case the researcher have opted to use a simple text based interface, as the details to be presented are simple for understanding and requiring minimal effort. If one had more complex representation and interaction one would have opted for a graphical interface. The details of the user interface are further discussed later.

During this stage of implementation, the formalized knowledge as a rule based inference mechanism is mapped or coded into the framework of the development tool, namely Knowledge-Pro Gold version 3.0 to build a working prototype. The contents of knowledge structures and inference rules established in the previous stages are organized into suitable format.

4.5.GEO-REFERENCED MAP OF ADDIS ABABA

The picture below shows that part of the Addis Ababa map selected for the purpose of this study.



Fig 4.4 Geo-referenced map of Addis Ababa in Arc View GIS software.

Selected Addis Ababa Routes

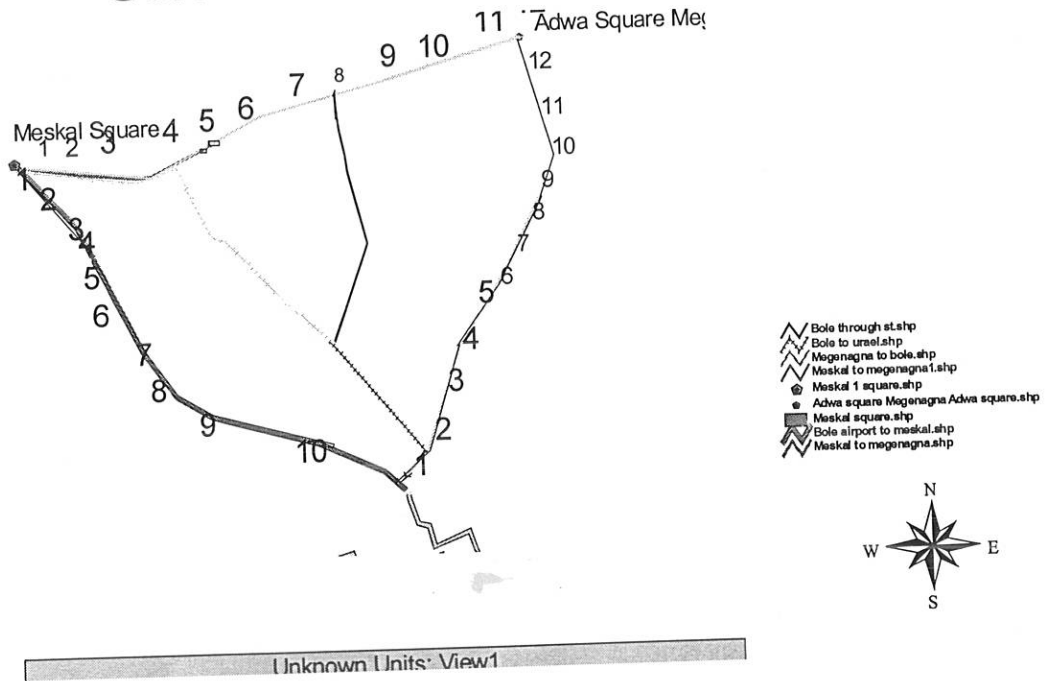


Fig.4.5 The three selected locations of Addis Ababa

This is the selected site from the overall map of Addis Ababa for purpose of this study using special GIS software called **Arc View GIS 3.1** version. The figure shows that the three selected locations of Addis Ababa, which are the International Bole Airport, Meskal Square (Abyot) and Adwa square (Megenagna).

These sites were selected according to the interview conducted with travel agents in Addis that are mostly the tourist attraction sites and a good number of hotels are found near to each other around this three roads and the transportation facility is relatively comfortable than the others. Most of the hotels that are found in Addis Ababa are found in this three location therefore it is a good representative of the Addis Ababa roads.

The overall Roads in the selected locations are:

1.Meskel square to Adwa Square (Megenagna) through Haile G/Sillassie Street.

This road composed of a good number of hotels, shops, offices, etc. and it is a distance of around 4560.18 meters or 4.5 km .In this road some hotels, shops and offices are:

- i. St. Estifanos
- ii. Yordanos Hotel
- iii. Ministry of Justice
- iv. Bambis Supermarket
- v. St.Urael Church
- vi. Plaza Hotel
- vii. Nigist Saba Hotel (Queen of Sheba Hotel)
- viii. Holiday Hotel
- ix. Ministry of Water Resources
- x. Axum Hotel
- xi. Debre Damo Hotel
- xii. Duty Free Shop
- xiii. Central Shoa Hotel
- xiv. Lem Hotel

2. International Bole Airport to Meskel Square.

This road is the main road, which takes from Bole to Meskel square and vice versa. It is called by street name Bole road. It is a distance of around 5033.16 meter or 5.03 km long. Some of the recreational center and service centers are:

- i. Bole International Airport
- ii. Awash Insurance
- iii. Africa Insurance Head Quarter
- iv. Peacock park
- v. Nib Bank and Head Quarter
- vi. Wegagen Bank and Head Quarter
- vii. Exhibition Center
- viii. Addis Ababa Museum
- ix. St. Joseph School
- x. Meskel Square (Abyot)

3. International Bole Airport to Meskel Square through Megenagna alternate 1.

This is the alternative route and known by street name called Bole road and Haile G/Selassie road. It has a distance of about 7871.63 meters or 7.87 km long the farthest route. Consists of the following:

- i. Bole International Airport
- ii. Brass MC Hospital
- iii. Airport Motel

- iv. Hayat Hospital
- v. Moenco
- vi. Imperial Hotel
- vii. Unity University College
- viii. Yeka Anbesa Garage
- ix. Ethio Plastic No. 2
- x. Quality and Standard Authority of Ethiopia (QSAE)
- xi. AMCE
- xii. Adwa Square Megenagna

4. International Bole Airport to Meskel Square through St. Urael alternate 2.

This is the second alternate route and known by street name called **Mandela Street**. It has a distance of about 5522.19 meters or 5.52 km long. Consists of the following:

- i. Bole International Airport
- ii. Bole Medhanealem Church
- iii. Desalegn Hotel
- iv. Atlas Hotel
- v. St.Urael Meskel Square

5. International Bole Airport to Meskel Square through St. Gabriel alternate 3.

This is the third alternate route and known by street name called **Djibouti Street**. It has a distance of about 7060.27 meters or 7.06 km long the next farthest route. Consists of the following:

- i. Bole International Airport
- ii. Brass MC Hospital
- iii. Bole Medhanealem Church
- iv. Awraris Hotel & St. Gabriel Church

- v. Axum Hotel and others listed earlier in Megenagna road from Axum Hotel to Meskel Square
- vi. Meskel Square

CHAPTER FIVE

KNOWLEDGE REPRESENTATION

5.1 INTRODUCTION

To use knowledge on machines, one must bring that knowledge out in the open and represent it formally. The process of representing knowledge formally is referred to as Knowledge representation (Parsaye and Chignell, 1988).

The model developed is used as a general framework to design the knowledge base. The type of knowledge representation that is appropriate in a given situation depends on what sort of knowledge is being represented and how it is to be applied. Various methods of knowledge representation have been proposed, each attempting to determine how knowledge should be represented by identifying the essential components of knowledge that allows us to reason about interesting and practical problems. The knowledge model in this study has been developed using the Top-Down approach of hierarchical structure.

This chapter deals with the knowledge representation technique used in the prototype development, i.e. the rule based system, and how system is tested and the development and integration aspects of **AARFExSys**, the prototype Web Based Expert System For Route Finding To The Road Network Of Addis Ababa.

The rule based system uses rules in the form of IF-THEN. The IF part needs to be satisfied by the facts for the goal, i.e. the THEN part, to be fired.

5.2 THE KNOWLEDGE BASE

The knowledge base is a collection of knowledge in the domain area. Rules in the form of if – then have been used in the research hence we can also refer to the Knowledge Base as the Rule base.

How the knowledge has been represented in the particular rule base using the syntax of the Knowledge Pro Gold Version 3.0 expert system shell is shown in the samples below:

Sample rules for route finding:

Topic result.

Set_number_of_values (result, 1).

IF ?'destination' is 'Airport'

AND ?'location' is 'Ethiopia Hotel'

AND ?'transportation' is 'Bus'

THEN result is 'Take bus no. 48 and get off after 4 stations you are there at the left side'.

IF ?'destination' is 'Airport'

AND ?'location' is 'Ethiopia Hotel'

AND ?'transportation' is 'Drive'

THEN result is 'Drive east past Ghandi hospital and Ghion hotel. After reaching the huge junction at Meskel Square, leave the road that turns right and go straight. Then at the junction, take the road to the right, leaving the one that goes straight. Drive for about 3.5 kms till you see the airport'.

IF ?'destination' is 'Airport'

AND ?'location' is 'Ethiopia Hotel'

AND ?'transportation' is 'Walk'

THEN result is 'Walk towards Meskel Square. Cross the big junction towards the gate of the Exhibition Center. Leave the exhibition center to your right and at the junction take the road to the right. Keep walking straight for another 3.5 k.ms and you'll see the airport'.

As Shown in the representation of the rules both modules in the knowledge base use the same format of representation. The whole set of terms preceded by a question mark (?) stand for the value of those terms. For example:

'? Destination' has the same meaning as 'the value of the variable destination', though we can't use the later for syntactic reasons.

Just like the question mark operator, the 'is' operator is found in each rule. It is like the equal sign used in more conventional programming languages. It assigns the right hand value to the left hand variable. For example:

IF? 'Destination' is 'Airport' is the same as the value of the variable destination is equal to 'Airport'. Similarly? 'And' is implied to additional value of another variable which is related to the first variable.

The then part is the final decision part of the rules that is the condition of 'IF' and 'AND' part satisfied then the 'Then' part is going to be executed. For Example:

THEN result is 'Drive east past Ghandi hospital and Ghion hotel. After reaching the huge junction at Meskel Square, leave the road that turns right and go straight. Then at the junction, take the road to the right, leaving the one that goes straight. Drive for about 3.5 kms till you see the airport'.

In the sample rules, it has been shown that the variables take only one value. For instance, Topic result.

Set_number_of_values (result, 1)

This will necessitate a quotation mark to bound terms with more than a word.

5.3 THE EXPLANATION FACILITY

The WHY feature of Expert systems is incorporated in this prototype. This feature gives explanation on why a certain question is being asked. WHY Buttons are available for each question number and the user is expected to click on the respective button when he/she needs to know why that particular question is being asked. An example follows:

One of the questions in the First Visit module of AARFExSys is “3rd Question: **Which means of transport are you planning to use?**”. In case the user was curious to know why the system is asking that question (i.e. 3rd Question), he/she can click on the button, ‘? Why’, which is supposed to mean, “Why this question?” At the click of the button, a separate window pops up with the following information:

Means of transportation is a key factor to determine the time to reach from one location to the other location, that is, drive is faster than any type of taxi, which is much faster than walk. Besides in case the user selects bus as means of transportation, there are some roads where there is no city bus assigned yet so it is important to ask users their preference mode of transportation

The **WHY** facility is built for the 'Means of Transportation' module.

5.5 MEASURING EXPERT SYSTEM PERFORMANCE

After knowledge is acquired and built an expert system how do we know either that it works or how well it works? Obviously, one way would be to ask an expert. Intuitively, the aim of a validation effort is to ensure that the expert system behaves “correctly”. But what does this mean for systems that produce in exact results? Can we ever be 100% sure that the knowledge is correct? From a philosophical viewpoint, the answer is no. From different perspectives, it is important to realize that expert systems will inevitably make mistakes. Rather than aiming to verify the behavior of an expert system, one should try to measure its performance.

Validation can be done either by judging the ultimate performance of the expert system as a black box or by focusing on the quality of the expertise itself.

To a similar extent one can understand that an effective **validation procedure** is critical to the success and acceptance of the system. Since this stage of the design is mainly concerned with:

1. Correctness, consistency and completeness of the rules;

2. Ability of the control strategy to consider information in the order that corresponds to the problem solving process;
3. Appropriateness of information to how conclusions are reached and why certain information is required; and most critically,
4. Agreement of the computer program output with the domain expert's corresponding solutions.

The study tried to maintain a simple and clear approach to end up with rules that integrate very easily without conflict. However, it should be noted that validation is an ongoing process requiring the output recommendations to be accurate for a specific user's case. Validation is enhanced by allowing others to review the system critically and recommend improvements. A formal user based evaluation would be necessary to establish whether the system meets the intended original goal.

Validation of expert systems is a crucial issue for applications in which one has to "get it right the first time"(Parsaye and Chignell, 1988).

Automated expert system validation as the need for using knowledge-based systems increases two fundamental questions need to be answered mentioned by"(Parsaye and Chignell, 1988):

- How do we acquire knowledge?
- How good is the knowledge we have acquired?

On a similar note, conditions that relate to uncertainty in expert system implementation always come up. The major causes of uncertainty usually stem from the validity of the knowledge base rule or with regards to the validity of the user response or both. The study tried at best to avoid uncertainty arising from the validity of the knowledge base rules as they are simple rules that are exclusive of each other and similarly each doesn't represent generalizations about conditions of the world that might be misleading.

With regards to uncertainty from users response, the conditions of the input requirements are absolute and hence require no subjective representations nor do they require any assumption to be made. However, further testing and implementation of the system is expected to yield results showing if conditions exist where the above might occur.

Testing the system has been done by checking the results of the system with the results of the experts, in this case experts are travel and tour agents and in addition to the experts the result has also been cross checked with the city map. Accordingly the result found shows that the system is satisfactory. Response taken from six experts four of them 66.66% yield result similar to the AARFExSys while one 16.66% gave more detail and accurate result than the system and one 16.66% less accurate than the system.

5.6 SYSTEM'S CONTROL STRUCTURE

The basic operation of AARFExSys is controlled by a method of backward reasoning rule interpreter. During operation the inference engine attempts to establish values for parameters by backward chaining so that it starts from the solutions at hand, which is from the rules generated, it will search and extract the required problem.

An actual AARFExSys rule in its English-like translation is as follows:

RULE 7

IF

Destination is Airport AND location is Ethiopia Hotel AND transportation is Taxi AND class is Contract AND type is Blue.

THEN

Take any blue and white taxi with four passenger seats that are usually found parked by the side of a road or flag down one that is just passing by. Negotiate the price to take you to the Airport. It shouldn't normally cost more than 20 birr to get there from your hotel and the time it takes is approximately 7 to 8 minutes.

To establish the value of the parameter Result, the rule interpreter follows the following steps:

1. Rule 7 is retrieved from the knowledge base, since its right hand side deals with Result.
2. The rule interpreter now has 5 goals to prove, namely the five goals on the left hand side of rule 7. These can prove by back ward reasoning because this ways the system could find out the solution.

The system interacts with the user during inference if it encounters parameters whose values are not known.

When the user wants to find the destination Airport from his hotel (Ethiopia hotel as shown in the rule) the search he use is a backward reasoning and at some other time he wants to start from the solution then he is also using a backward reasoning technique.

CHAPTER SIX

APPLICATION OF KNOWLEDGE

In this chapter, the actual working of the system **AARFExSys** and the design of the web site is discussed in detail.

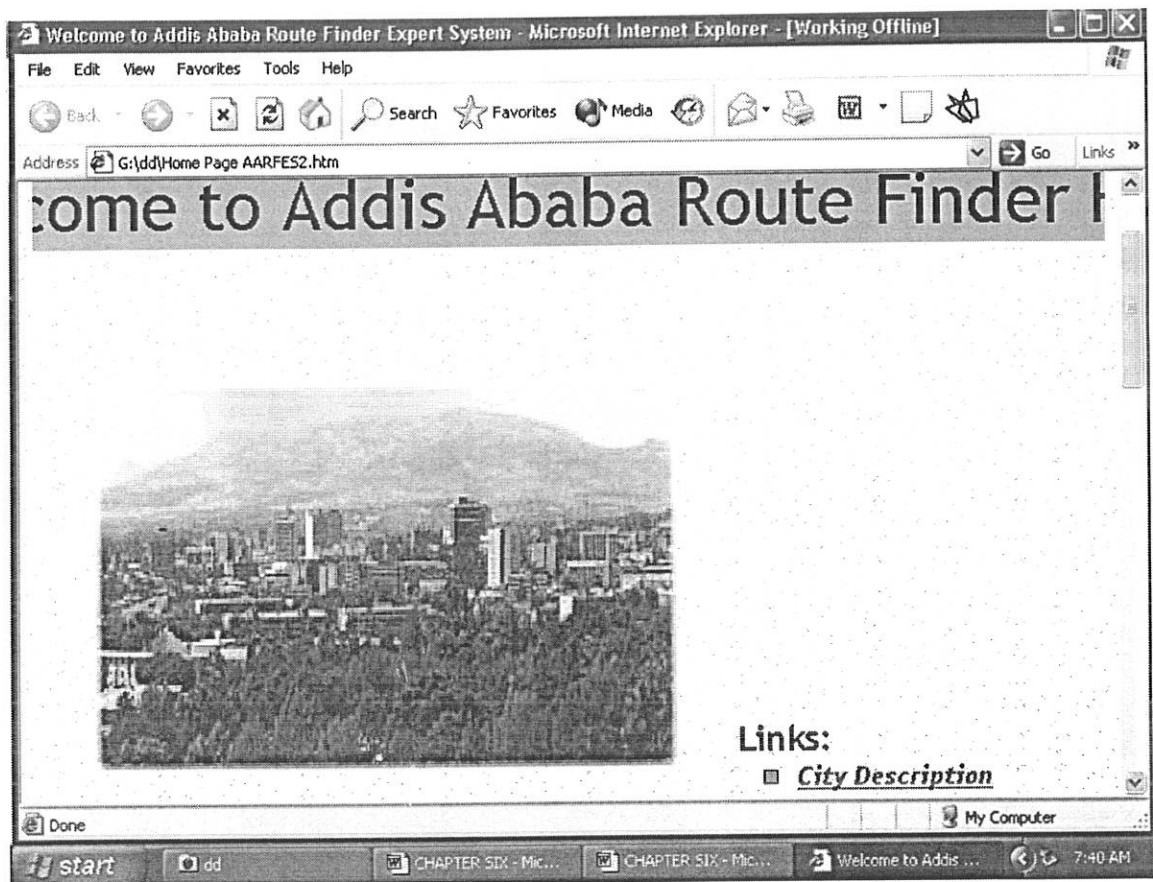
6.1 WEB DESIGN

The web is designed using Hyper Text Markup Language (HTML) and Microsoft Front Page to help users access the system anywhere in the world. Since users of this system are mainly Tourists coming to visit our country from different countries, **AARFExSys** helps them by giving online access regarding to the route finding information. Even though the system is not dynamic and work offline for the purpose of this prototype system.

Home

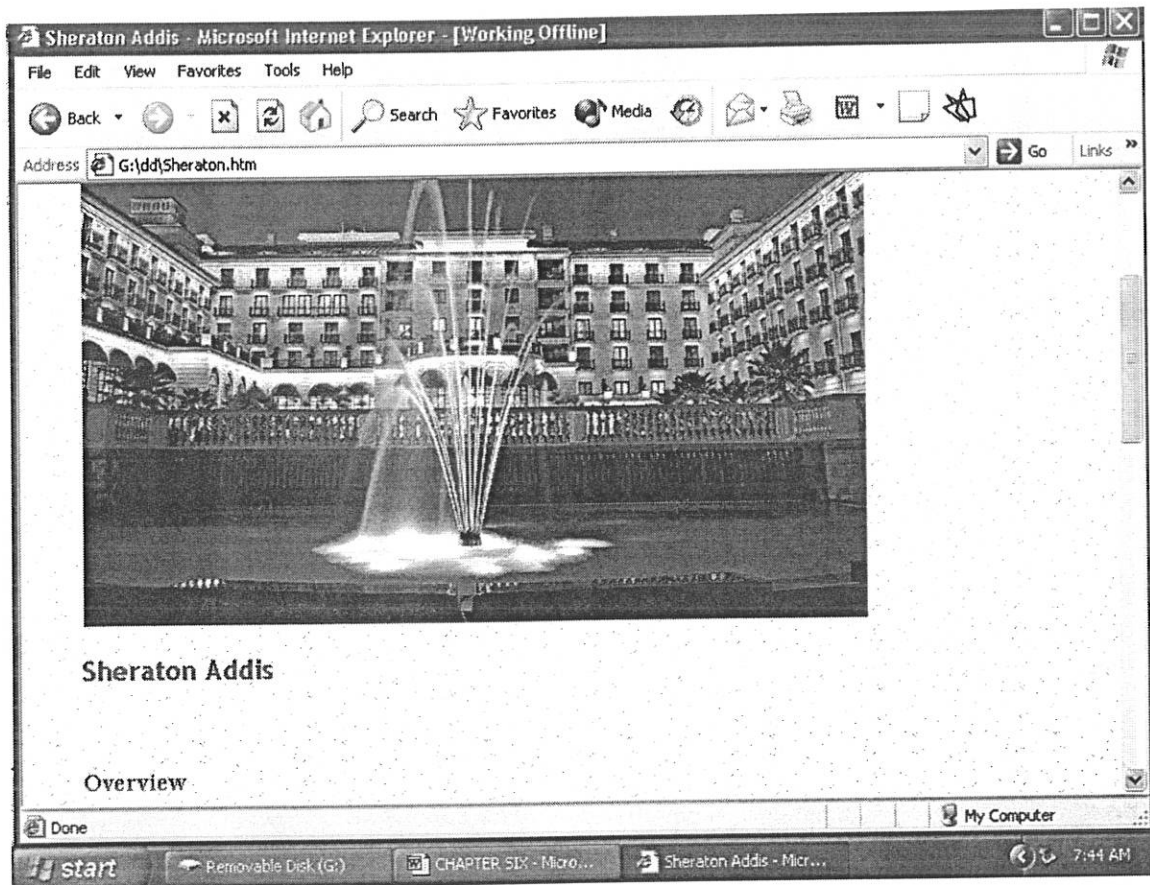
The home page consists of the description of the city of Addis Ababa and the recent photo that shows part of the city. It consists of various links including Sample roads and sample hotels found in Addis Ababa. It also includes links on the history of the city in general; the contact addresses of different important city service centers and most importantly the **AARFExSys** system.

Figure 6.1. shows Welcome to web Based Expert System For Route Finding to the Road Network of Addis Ababa (AARFExSys).



As indicated above the home page consists of important links and one of the links among the various others is a sample international hotel in the city of Addis Ababa, Sheraton Addis with its description is shown below.

Fig. 6.2. Sample international hotel in the city of Addis Ababa.

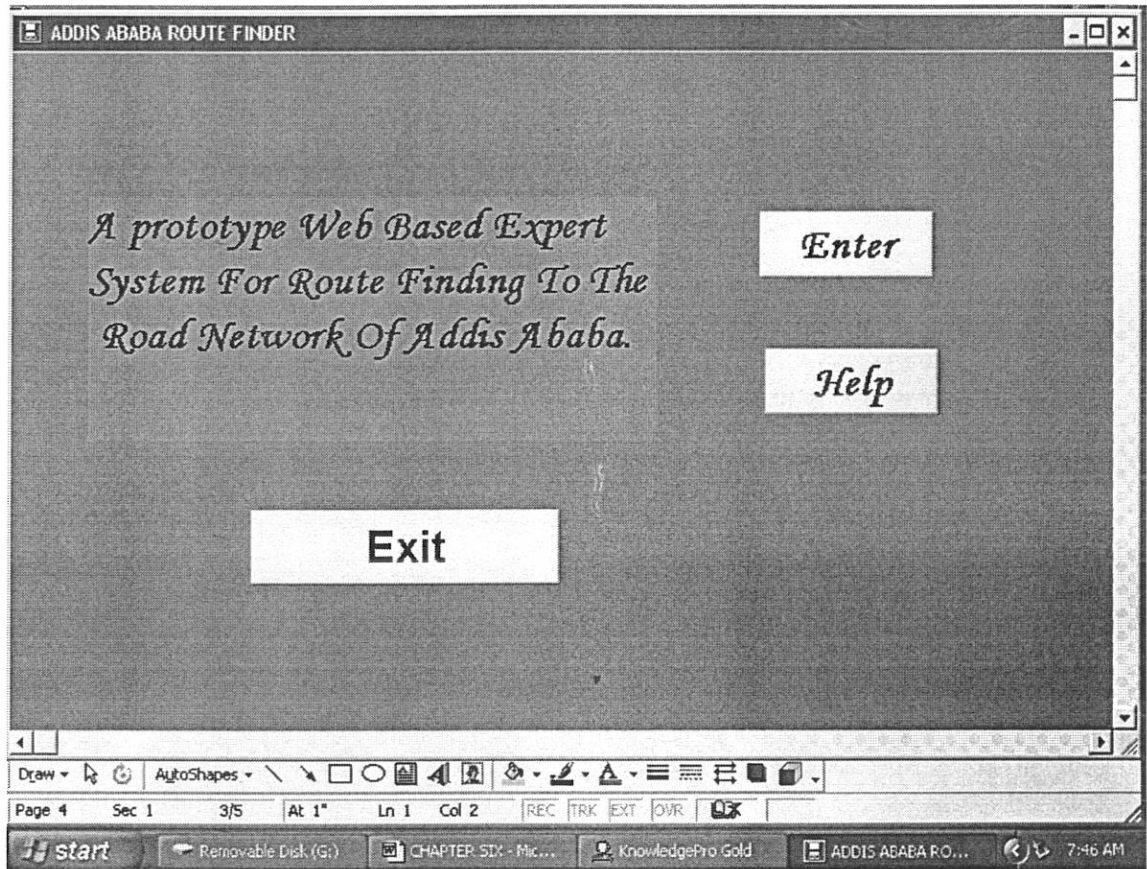


6.2. USER INTERFACE

The system maintains a user interaction through interface samples shown below. The user is initially prompted to a welcoming screen along with command buttons indicating available services given by the system. In the initial prompt users are greeted by the interface that tells about the system, and gives choices of proceeding further.

The user can enter the search system through the “enter” button, or click the “Help” button to find out more about the system, or simply leave the system through “exit” button.

(Fig.6.3): A Prototype web Based Expert System For Route Finding to the Road Network of Addis Ababa (AARFExSys).

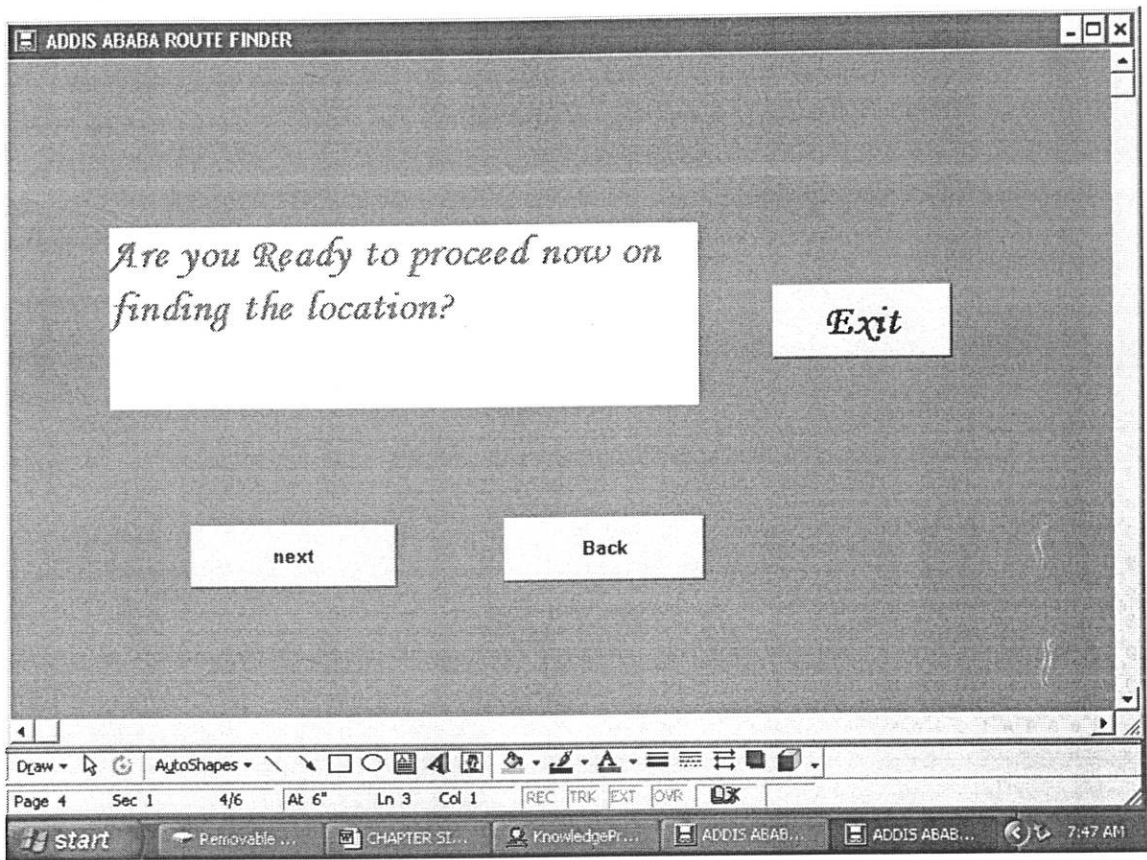


Once a user has entered the system for expert information, the next screens would pop with three major questions along with a predefined drop down list of options about user

preferences. Finally, once the user requirements have been defined one can view the system recommendation through “next” button to arrive at the results of the system.

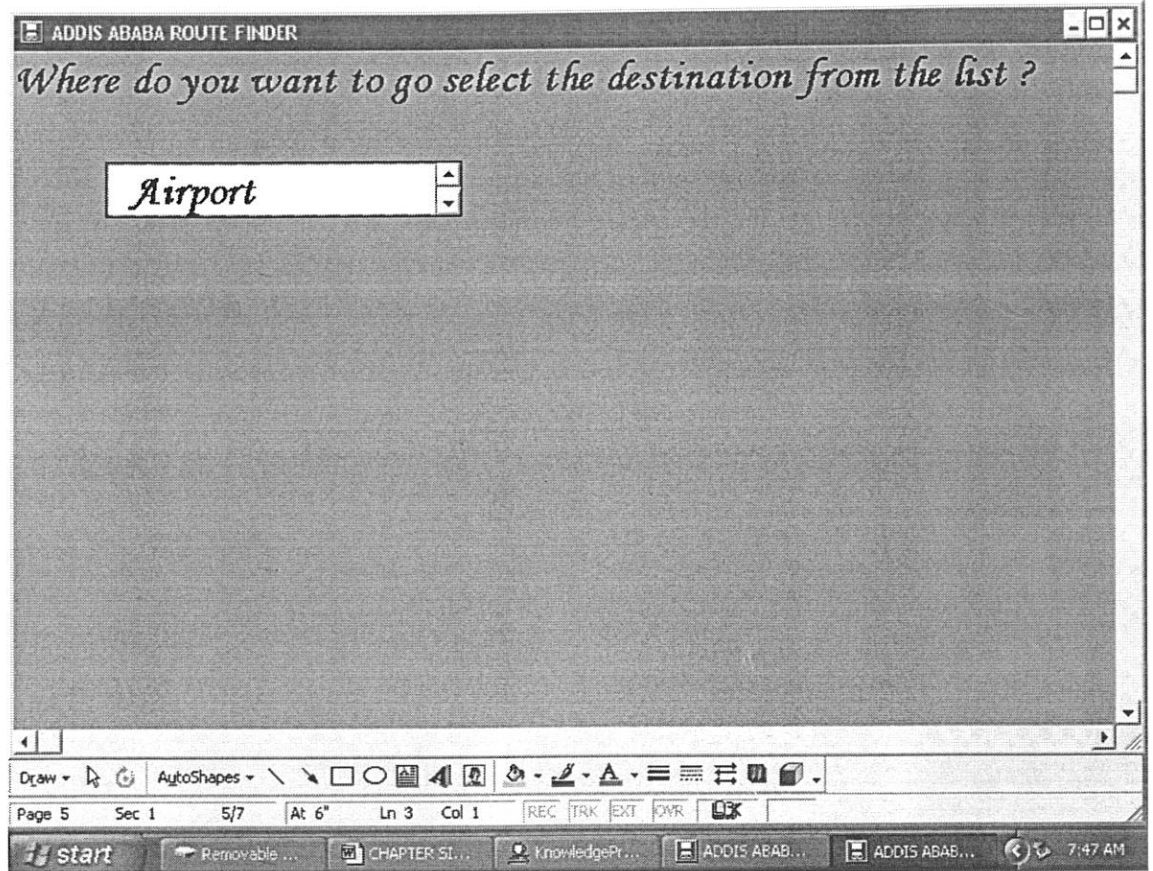
Though the system has been able to design the interface functions through which user interaction may be achieved, the system has not been able to integrate it with the inference engine due to difficulties in the designing software and at the same time continuous crashing of the shell used in this case.

Fig.6. 4 How AARFExSys
looksatruntime



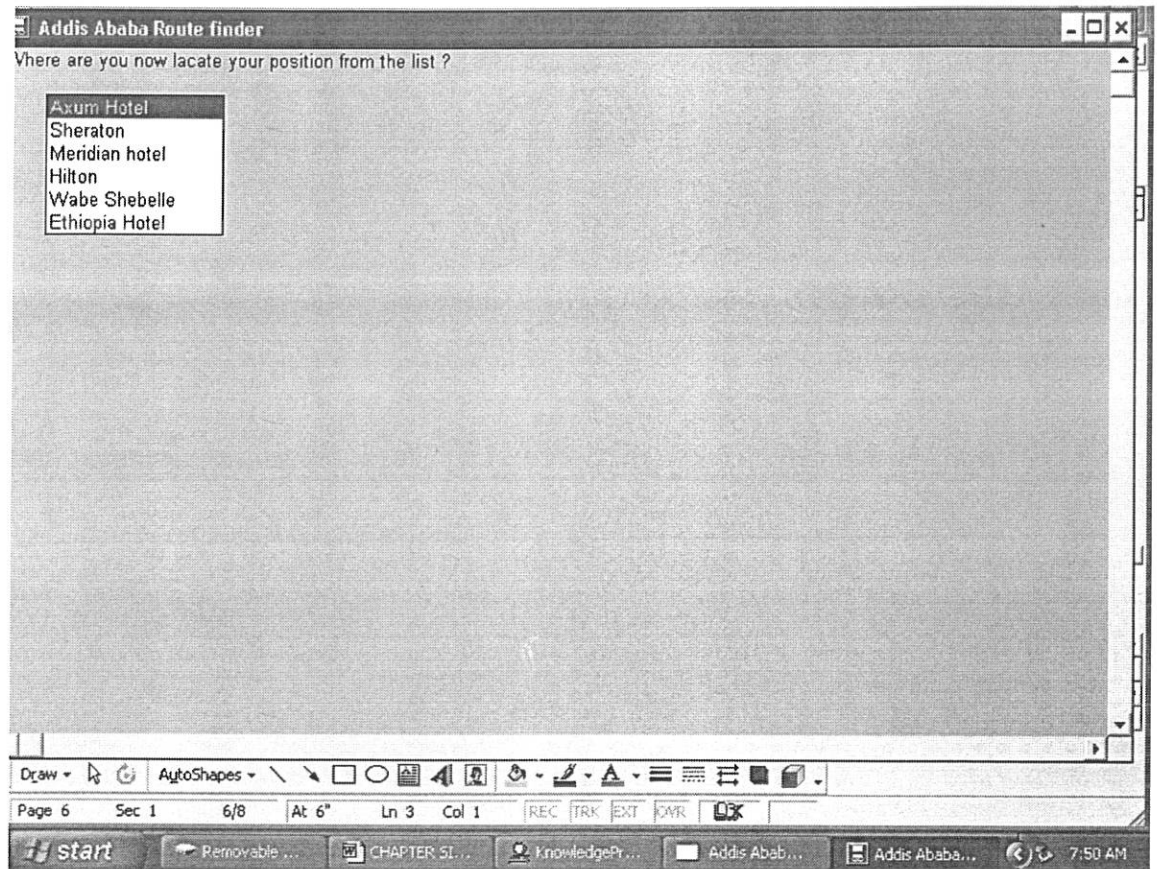
After the user confirms that he is ready to continue by clicking the “Next” button then the system asks questions, and the user is expected to answer them by double clicking on the right options from the list.

Figure 6.5. The first question in AARFExSys prototype system.



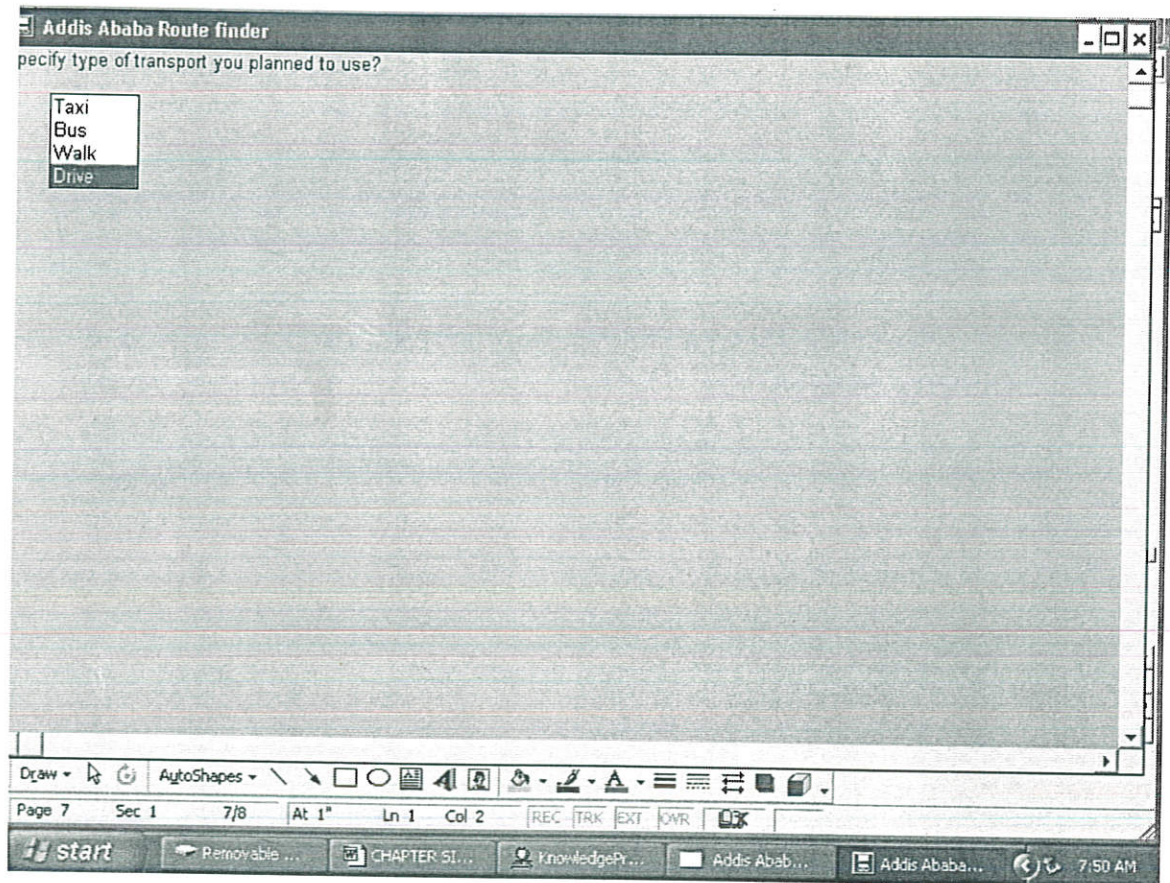
The user is expected to identify the destination from the list shown in the above screen so that the system will go to the next question that is identifying the location of users.

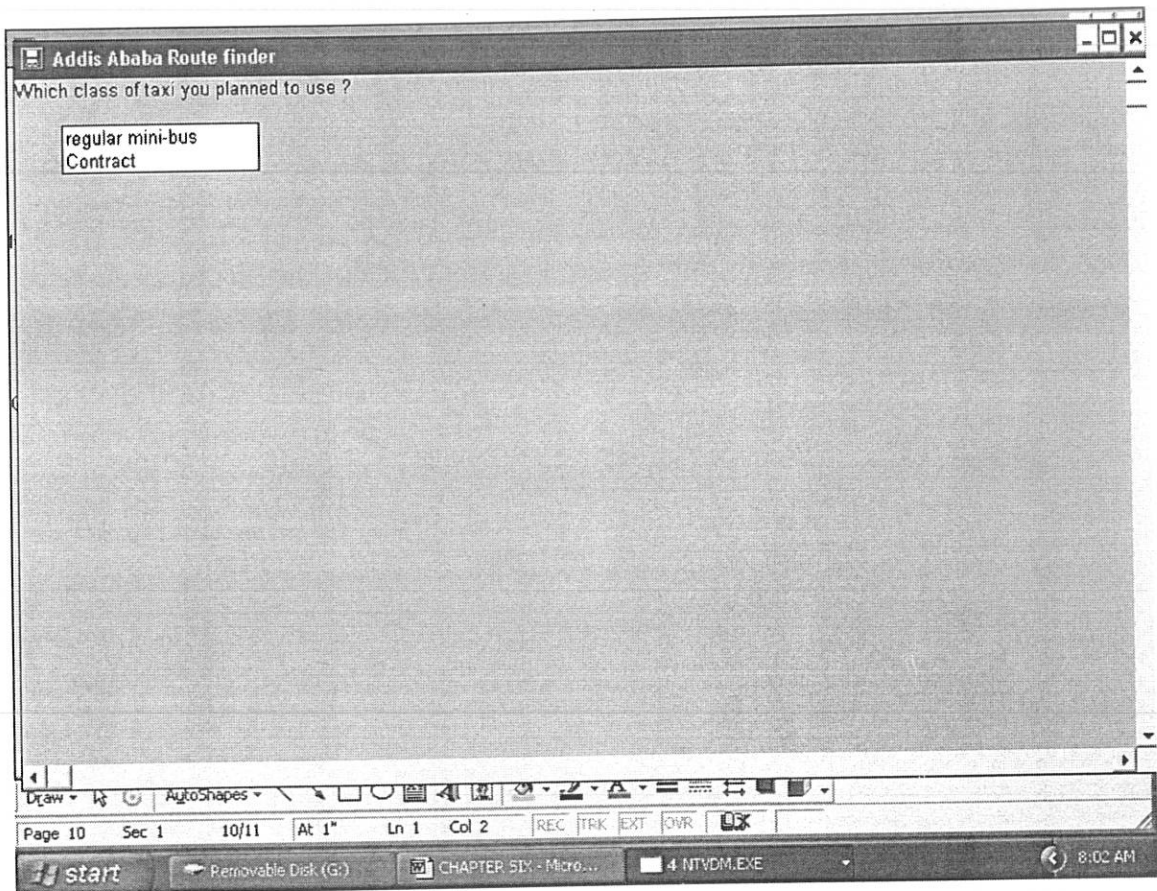
Fig.6.6. Below shows identifying the location of users.

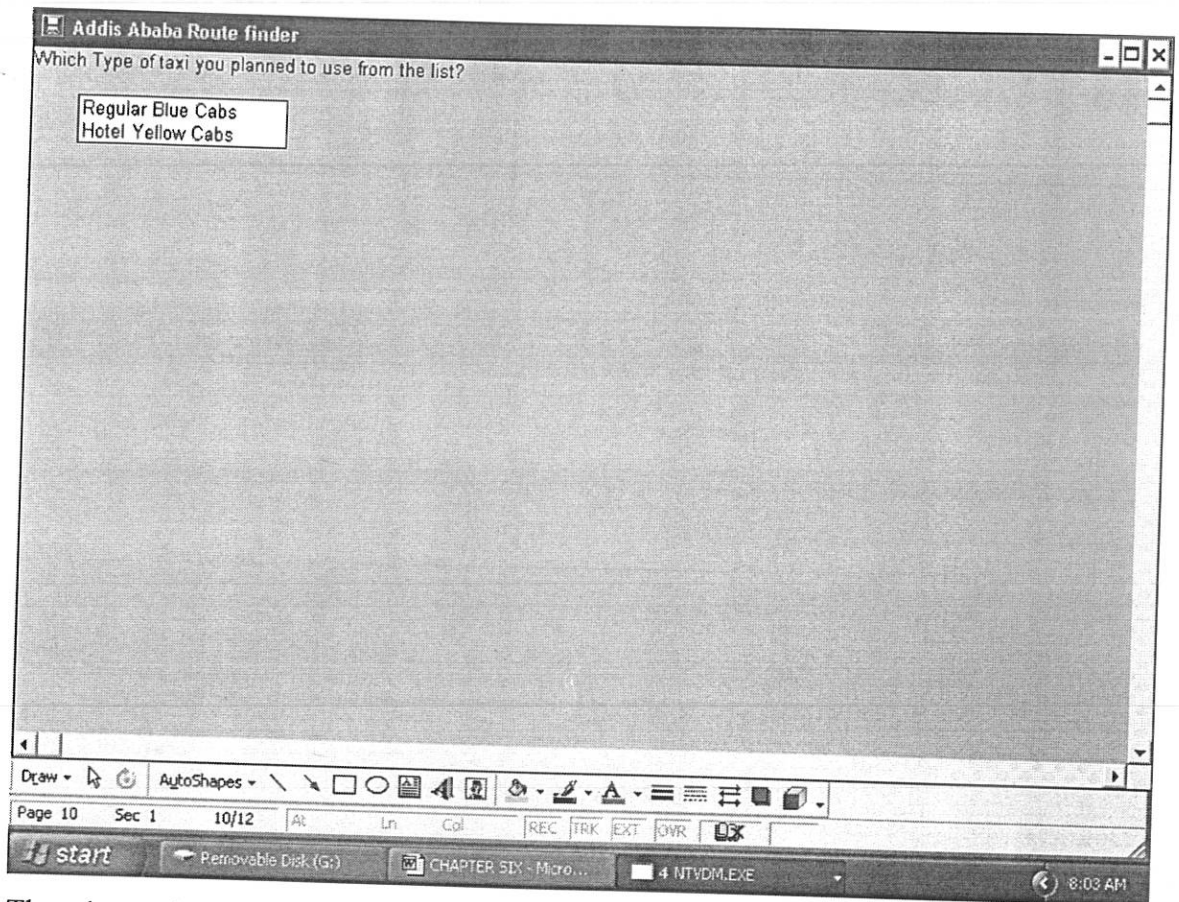


In the next screen below users are asked to specify the means of transport they are planned to use to find the location they require.

Fig.6.7.Shows the means of transport users plan to use.

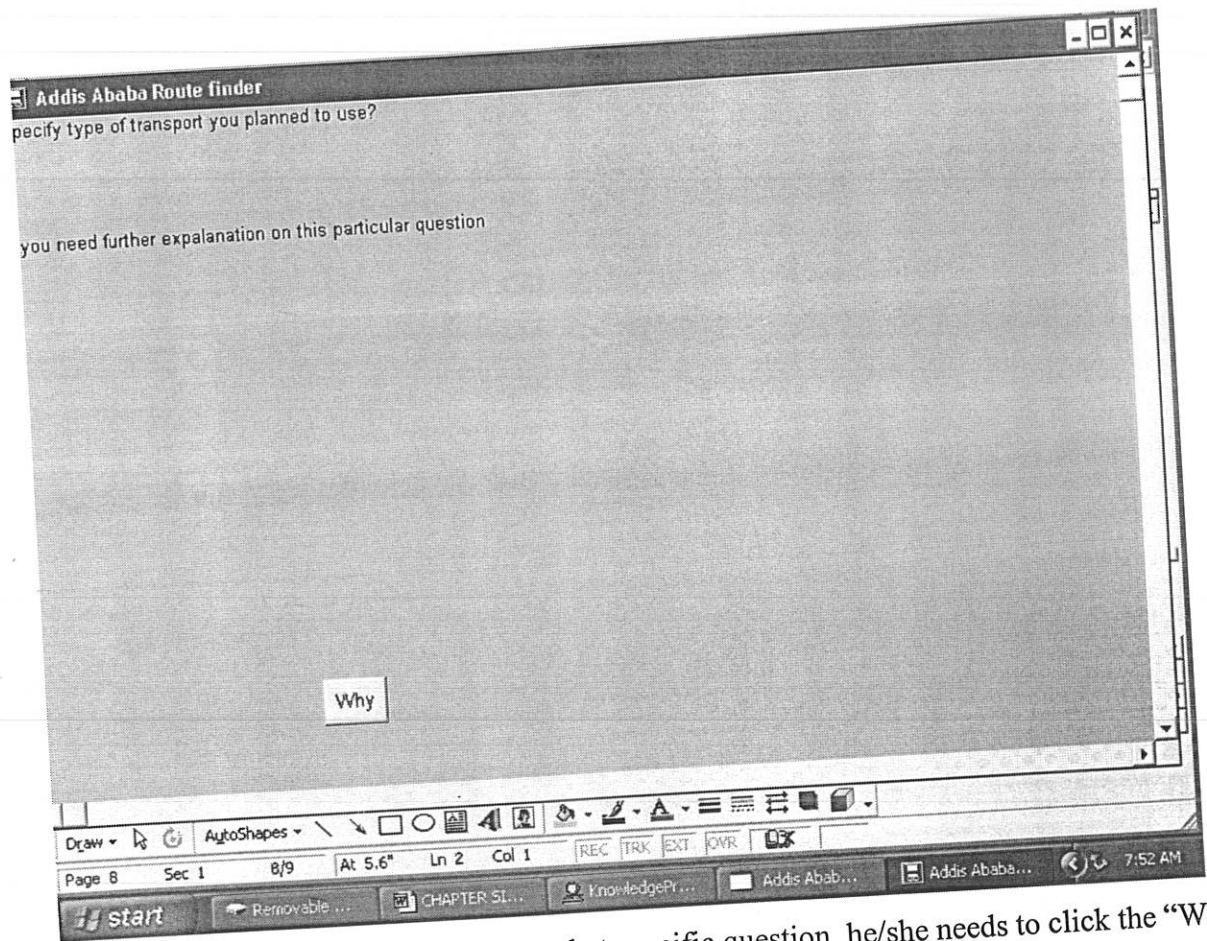






The why explanation Facility is included in the system for instance if the user wants to get further explanation the “why” button will describe about that specific question why that question is specifically asked. If the user wants to know further explanation about why that question is asked he/she can click to the why button and see the description in new window.

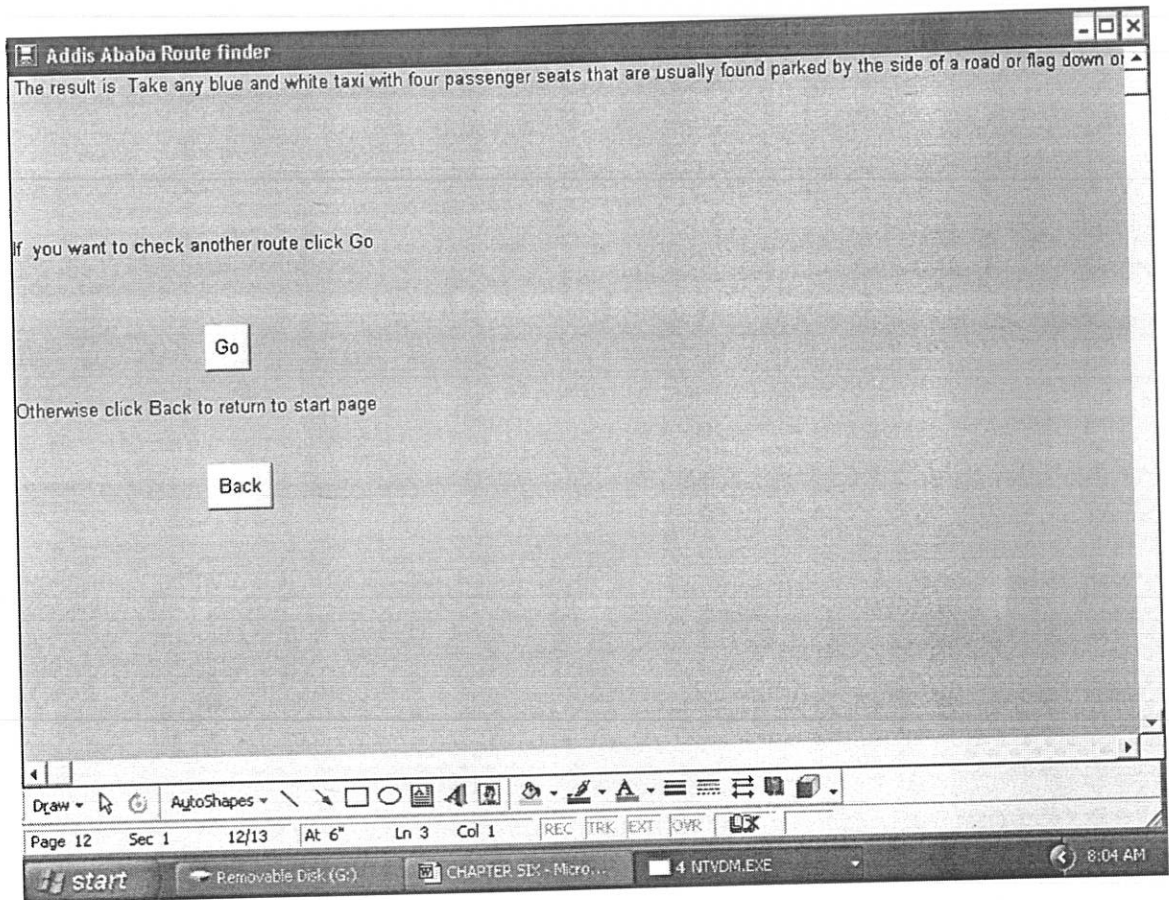
Fig.6.8.shows the why explanation facility



If the user wants further explanation about that specific question he/she needs to click the “Why” button as shown above then a new window pop up will appear and give the explanation about that question. After that the user come back to the question and select the choice and get the result.

The final result will appear in a new window that suggests the direction with the distance, the cost and time it takes to reach the destination.

Fig.6.9.Shows the final result in a new window



6.3. TESTING AND EVALUATION

The prototype was verified to demonstrate that it has been built in conformance with a number of well-defined properties; primarily freedom from logical conflict, redundancy and deficiency. Conflict refers to the ability of the system to arrive at logically inconsistent conclusions from consistent input.

6.4. KNOWLEDGE BASE MAINTENANCE

The quality of knowledge in a knowledge base is the ultimate factor that determines the worth of decisions reached by an expert system. The care taken during knowledge acquisition and representation should be extended further into the knowledge base maintenance.

6.5. DESIGN REQUIREMENTS

HARDWARE REQUIREMENT

Though, AARFExSys is developed using Pentium III PC with 128 MB of system memory and 644 MHz processor speed, according to KPWIN (1991), it is suggested that Knowledge-Pro for Windows (KPWIN) could be used in PCs having at least 4MB system memory and processor speed of 133 MHz. However, 8 MB memory at least for use with Visual C++ environment is recommended.

SOFTWARE REQUIREMENT

The Knowledge-Pro for Windows (KPWIN) shell is version 3.0. To work on this version of KPWIN, there is a need for at least the latest versions of Microsoft C/C++ (V7.0) or Borland C++ (V3.1), or the Windows version of C++, i.e. Microsoft Visual C++.

In the network environment, the new system requires a network operating system, preferably Windows NT 4.0 and at least Windows 95 operating system on client PCs.

CHAPTER SEVEN

CONCLUSIONS AND RECOMMENDATIONS

7.1 CONCLUSIONS

Tourism in Ethiopia is not a well treated and it is an ignored sector. Whereas tourism is ranking at the top head of all categories of international trade, evident from the statistical information compiled from all over the world by WTO (World Trade Organization) as discussed in the previous chapters.

Tourists come for various reasons, such as visiting historical places, museums, and churches, and cultural and social heritages. These tourists expect reasonable transportation and hotel services during their stay in the country. Particularly their stay in the country is greatly depend upon the transportation facility since they mostly commute to visit those places mentioned above.

The importance of making information available to users when and where needed seeks no justification. Addis Route Finder (AARFExSys) is an expert system prototype designed to assist tourists in having a full-fledged information on how to move about the town without being limited by language barrier, inexact rendering of direction by humans and exposure to spending unnecessary time and money in getting experienced tour guides available in the city. According to the findings the interests of tourists in their commuting activity to visit places are diverse in type and needs to be fulfilled to promote tourism in Ethiopia. The AARFExSys tries to fill this gap to fulfill various needs of users.

The system can be made available at various hotels so that the visitors can get access to this guide before they leave their hotels. The system can also be available in the web site, which can be integrated with Ethiopian Tourism Commission web site as a link to provide tourists this particular information wherever they are regardless of the location.

This study is made in an effort to explore applicability of Expert system in Addis Ababa city Route finder and develop a prototype, which assists tour agents in the choice of location with transport option and avoid the language barrier.

The knowledge acquisition is done using manual (from interview) and automated technique (using a GIS software called Arc View GIS 3.1 version) as well as standard texts in the domain area.

Hierarchical structuring was used to model knowledge acquired from the documents. Rules are used to represent the knowledge in accordance with the model developed. The Knowledge Pro Gold expert system shell was used to develop the knowledge base and ultimately run the prototype, AARFExSys. KPWIN 3.0 uses backward chaining, which was found to be appropriate with rules and the ultimate goal.

Microsoft Front Page 2000 and the Hyper Text Mark up Language (HTML) tools used to develop the web (home page and the different links) to make accessible to users.

The user interface is a Graphical User Interface, and clicking on the respective buttons can access the two modules of the system. Users had to go through a series of questions at the end of which they will be presented the location with alternatives, the distance, the cost and time it takes to reach.

keeping with the view that knowledge is never complete, a self-learning system should be developed. That way if a rule is not there, it would be when it's needed.

Wider application areas can also be identified in similar domains. Such systems should be developed with user-focused demands and approaches should be highly scalable through continuous prototyping to arrive at an applicable system that is fully functional.

BIBLIOGRAPHY

1. Addis 21: planning the future of our city Addis Ababa revised master plan proposals draft summary. Addis Ababa city government. December 2000.
2. Ayalew Sisay, (1992). Development of Tourism In Ethiopia. Addis Ababa: Ethiopia Tourism Commission.
3. Chiu, Dickson K.W et.al (2004). A Multi-Modal Agent Based Mobile route Advisory System (MIRAS) for Public Transport Network. Proceedings of the 38th Hawaii International Conference on System Sciences-2005. Chinese University of Hong Kong, Hong Kong.
4. Chuenpichai, Kanjana et.al (1998). GIS and Expert System Integration for Land Development Village. Thailand. Paper No: 755
5. Copeland, Jack (1993). Artificial Intelligence: a Philosophical introduction. Blackwell, Massachusetts.
6. Gardner, H. (1986). The Mind's New Science. New York: Basic Books, 1986.
7. Giarratano, J. and Riley. (1998). Expert Systems: Principles and Programming. 3rd ed. California: PWS Publishing Company.
8. Ignizio, James P. (1991). Introduction to expert systems: The Development and Implementation of Rule-Based Expert Systems. San Diego: McGraw-Hill Inc.
9. IPT (2000). Artificial Intelligence: An Overview.
Available at: www.sunnybank.qid.edu.au/ipt/ai/ai3.htm

10. Jackson, Peter (1999). Introduction to Expert System. New York: Addison-Wesley Publishing Company.
11. Jackson, Peter (1986). Introduction to Expert System. New York: Addison-Wesley
12. Kendal, A. [1991]. Fuzzy Expert Systems. London: CRC Press.
Publishing Company.
13. JHS Travel Expert (2001). Japan Home Search. Japan.
Available at: http://www.japanhomesearch.com/travelexpert/te_main.asp.
14. Kefelegn Getahun (1994). An Overview Geographical Analysis of Tourism In Ethiopia.
Senior Essay. Unpublished. Addis Ababa, AAU.
15. KPWIWN, 1991. Knowledge Pro For Windows: A user manual, New York: Knowledge Garden Inc.
16. Leondes, Cornelius T. (2000). Knowledge Based Systems: Techniques and Applications.
Vol.4. San Diego: Academic Press.
17. Liu, B.J Tay. (2000). Using knowledge about the road network for route finding. National
Singapore: University of Singapore.
Available at: <http://www.cs.nps.navy.mil/people/faculty/rowe/c4i00c.htm>
18. Masuch, M.(ed.).(1990). Organization, Management and Expert Systems. Berlin: Walter De Gruiter.
19. Melaku Kebede. (1997). An Approach to the integration of Expert Systems and Hypermedia Technology in Developing A Prototype Tourist Information System For Ethiopia. Unpublished. School of Information Studies for Africa. Addis Ababa: AAU.

- 20.ORAAMP (2002). Office for the Revision of the Addis Ababa Master Plan. urban development and work bureau. Addis Ababa city government. Addis Ababa.
- 21.Parsaye, Kamaran and Chignell, Mark (1988). Expert Systems for Experts. New York: John Wiley & sons, Inc.
- 22.Pederson, K. (1989). Expert Systems Programming: Practical Techniques for Rule Based Systems. New York: John Wiley and sons.
- 23.Route Finder - UK & Ireland (1998). Third Edition. Maps & Route Planners In Computer Software.
Available at: www.reviewcentre.com/reviews54317.html
- 24.Ring road in Addis Ababa construction by Chinese company inaugurated. 2003 Publication: Xinhua News Agency; newspapers.
http://www.ethiopians.com/photoessay/photo_essay_june2301.htm
- 25.Stefik, M. [1998]. Introduction to Knowledge Systems. California: Morgan Kaufmann Publishers.
26. Steiner, T. (1995). Information Technologies and Destination Management in Tourism: A conceptual Framework. Germany. MBI Thesis. University Lausanne.
- 27.Sun, Yuxian and Lee, Lyndon (2000). Agent-based personalized tourist route advice system. Department of Geo-Information Science Ipswich .UK.

28. Winston, Patrick Henry (1992). Artificial Intelligence. 3rd ed. New York: Addison Wesley Publishing Company.
29. World Tourism Organization (1998). Tourism Development Report, Policy and Trends. Madrid.
30. Yazdani, Masoud (ed.) (1986). Artificial Intelligence: Principles and Applications. London: Chapman and Hall Computing.

Annexes

Annex1

Sample rules

Topic result.

set_number_of_values (result,1).

if ? 'destination' is 'Airport'
and ? 'location' is 'Axum Hotel'
and ? 'transportation' is 'Bus'
then result is 'Take bus no 48 and get off after 4 stations at Meskal square and take another bus number 74 and get off after 5 stations'.

if ? 'destination' is 'Airport'
and ? 'location' is 'Axum Hotel'
and ? 'transportation' is 'Drive'
then result is 'Drive east to Megenagna square and then take the ring road straight until you see the Airport'.

if ? 'destination' is 'Airport'
and ? 'location' is 'Axum Hotel'
and ? 'transportation' is 'Walk'
then result is 'Take the road perpendicular to the hotel (straight, facing the hotel) and take a left turn when you reach Atlas hotel. Keep on going for about 2.5 kms, passing Kaldis Pastry and St. Gabriel Church on the way until you reach Bras MCH hospital. You can see the Airport after that'.

if ? 'destination' is 'Airport'
and ? 'location' is 'Axum Hotel'
and ? 'transportation' is 'Taxi'
and ? 'class' is 'regular mini-bus'
then result is 'Take a taxi to Megenagna on the other side of the hotel. Get off at Megenagna and take another taxi to Bole'.

if ? 'destination' is 'Airport'
and ? 'location' is 'Axum Hotel'
and ? 'transportation' is 'Taxi'
and ? 'class' is 'Contract'
and ? 'type' is 'Regular Blue Cabs'
then result is 'Take any blue and white taxi with four passenger seats that are usually found parked by the side of a road or flag down one that is just passing by. Negotiate the price to take you to the Airport. It shouldn't normally cost more than 20 birr to get there from your hotel'.

if ? 'destination' is 'Airport'
and ? 'location' is 'Axum Hotel'
and ? 'transportation' is 'Taxi'
and ? 'class' is 'Contract'

and ?'type' is 'Hotel Yellow Cabs'
then result is 'Take the yellow hotel taxis straight to the Airport. Each hotel would have its own fare which you can find out from the reception'.

IF ?'destination' is 'Airport'
AND ?'location' is 'Ethiopia Hotel'
AND ?'transportation' is 'Bus'
THEN result is 'Take bus no. 48, or 42, or 52 and get off after 4 stations and turn left'.

IF ?'destination' is 'Airport'
AND ?'location' is 'Ethiopia Hotel'
AND ?'transportation' is 'Drive'
THEN result is 'Drive east past Ghandi hospital and Ghion hotel. After reaching the huge junction at Meskel square, leave the road that turns right and go straight. Then at the junction, take the road to the right, leaving the one that goes straight. Drive for about 3.5 k.ms till you see the airport'.

IF ?'destination' is 'Airport'
AND ?'location' is 'Ethiopia Hotel'
AND ?'transportation' is 'Walk'
THEN result is 'Walk towards Meskel Square. Cross the big junction towards the gate of the Exhibition Center. Leave the exhibition center to your right and at the junction take the road to the right. Keep walking straight for another 3.5 k.ms and you'll see the airport'.

IF ?'destination' is 'Airport'
AND ?'location' is 'Meridian Hotel'
AND ?'transportation' is 'Walk'
THEN result is 'Walk east for about 1.5 kms and you're there!'.

IF ?'destination' is 'Airport'
AND ?'location' is 'Ethiopia Hotel'
AND ?'transportation' is 'Taxi'
AND ?'class' is 'regular mini-bus'
THEN result is 'Walk east on the side of the hotel sticking to the road at the right at the junction. Get into a blue and white mini bus taxi to Bole. The final destination of Bole taxis is usually the airport itself.'.

IF ?'destination' is 'Airport'
AND ?'location' is 'Ethiopia Hotel'
AND ?'transportation' is 'Taxi'
AND ?'class' is 'Contract'
AND ?'type' is 'Regular Blue Cabs'
THEN result is 'Take any blue and white taxi with four passenger seats that are usually found parked by the side of a road or flag down one that is just passing by. Negotiate the

price to take you to the Airport. It shouldn't normally cost more than 20 birr to get there from your hotel.'

IF '?destination' is 'Airport'
AND '?location' is 'Ethiopia Hotel'
AND '?transportation' is 'Taxi'
AND '?class' is 'Contract'
AND '?type' is 'Hotel Yellow Cabs'
THEN result is 'Take the yellow hotel taxis straight to the Airport. Each hotel would have its own fare which you can find out from the reception.'

IF '?destination' is 'Airport'
AND '?location' is 'Meridian Hotel'
AND '?transportation' is 'Bus'
THEN result is 'Take bus no. 42 or 48 or 52 and get off after 3 stations.'

IF '?destination' is 'Airport'
AND '?location' is 'Meridian Hotel'
AND '?transportation' is 'Drive'
THEN result is 'Take the road that turns left when you drive out of the hotel. Keep driving straight for about 1.5 kms.'

IF '?destination' is 'Airport'
AND '?location' is 'Meridian Hotel'
AND '?transportation' is 'Taxi'
AND '?class' is 'regular mini-bus'
THEN result is 'Cross to the road opposite your hotel and get into a taxi that goes to Bole.'

IF '?destination' is 'Airport'
AND '?location' is 'Meridian Hotel'
AND '?transportation' is 'Taxi'
AND '?class' is 'Contract'
AND '?type' is 'Regular Blue Cabs'
THEN result is 'Take any blue and white taxi with four passenger seats that are usually found parked by the side of a road or flag down one that is just passing by. Negotiate the price to take you to the Airport. It shouldn't normally cost more than 20 birr to get there from your hotel.'

IF '?destination' is 'Airport'
AND '?location' is 'Meridian Hotel'
AND '?transportation' is 'Taxi'
AND '?class' is 'Contract'
AND '?type' is 'Hotel Yellow Cabs'
THEN result is 'Take the yellow hotel taxi straight to the Airport. Each hotel would have its own fare which you can find out from the reception.'

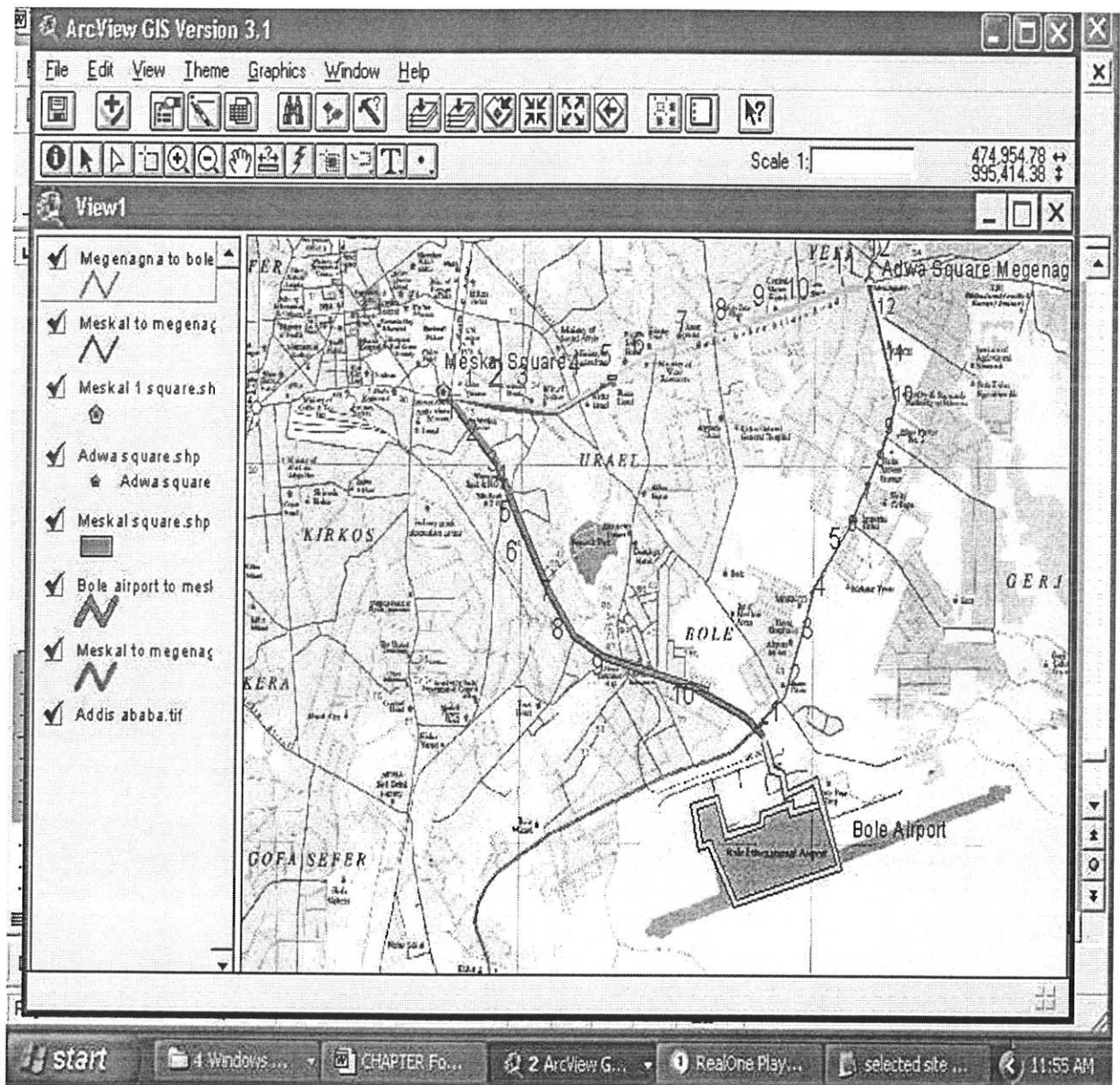
Annex 3

Interview Questions prepared for Tourists

1. How do you get information to find the location in Addis Ababa in your commuting activity?
2. Do you think that you can get sufficient information regarding the Addis Ababa route?
3. Do you need any computer-assisted tool to provide such information Efficiently?
4. Can you communicate with the tour guides in finding the direction such as finding hotels, preferred mode of transportation, etc?
5. What interests you in finding the location you want? Is it the shortest path? Or the less traffic crowd? Or the comfortability of the road to drive? Or the historical value of the street? Or any other?
6. How do you find the language ability and understanding of tour guides in Addis Ababa?

Annex 4

The map of part of Addis Ababa road map that is geo referenced using GIS software showing the selected area for this study.



DECLARATION

I undersigned declare that this thesis is my original work and has not been presented for a degree in any other university, and that all materials used for the thesis have been duly acknowledged



Daniel Admassu

June 2005

The thesis has been submitted for examination with my approval as a university advisor

Dr.Daneal Fekresellassie

June 2005

22

**ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
FACULTY OF INFORMATICS
DEPARTMENT OF INFORMATION SCIENCE**

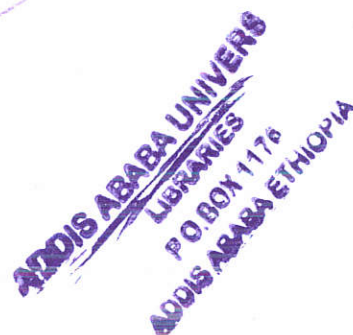
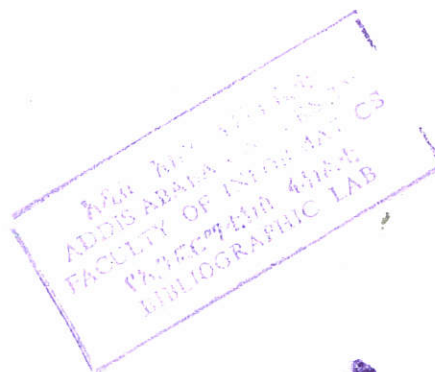
**Predictive modeling using Data Mining Technology in Support of
Cement Quality Assessment**

By

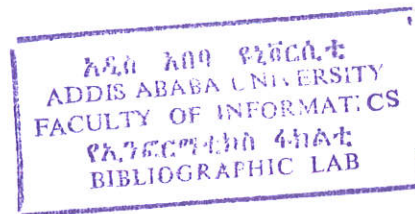
Abenet Tale Workenh

*A thesis submitted to the School of Graduate Studies of
Addis Ababa University in partial fulfilment of the requirements
for the Degree of Master of Science in Information Science*

July 2005



ADDIS ABABA UNIVERSITY
SCHOOL OF GRADUATE STUDIES
Faculty of Informatics
Department of Information Science



PREDICTIVE MODELING USING DATA MINING TECHNOLOGY IN SUPPORT OF
CEMENT QUALITY ASSESSMENT

BY
ABINET TALE

Name and Signature of Members of the Examining Board

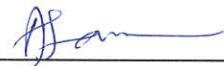
Dr. Gashaw Kebede, Chairman, Examining Board



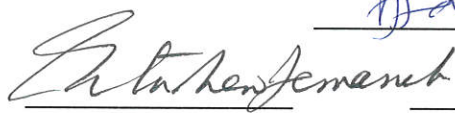
Dr. Lars Asker, Advisor



Prof. A. Santanam, Examiner



Chairman, Faculty



Signature

Date

12/07/05

Chairman, Graduate Council

Signature

Date

DEDICATION

I would like to dedicate this paper to my parents, Ato Tale Workenh and W/ro Birke Mitiku, my bothers and sister, and my dearest friends, Getenet Aklilu and Sosina Beryhune.

ACKNOWLEDGEMENT

I would like to express my deep sense of gratitude to my advisor, Dr Lars Asker, for his expert guidance and encouragement through out this thesis work.

Dr Lars Asker, besides recommending the Rule Discovery System software (RDSTM), he also facilitated the issuance of free temporary Academic license by the licensor (Compumine AB) to the Department of Information Science so that I can use it free of charge. I thank him very much.

Special thanks go to Ato Mesfin Getachew, for his comments and suggestions.

I am greatly indebted to my brother, Girma Tale, for his finical support during my two years of study.

I thank my class mates, Seifu Tatek and Maryie Yigzaw, for their help and support.

I am also grateful to the staffs of Mughher Cement Enterprise, Almaz Shite, Sendu Getachew, Meaza Zewdie, and Fanuel Cherinet, for their love and support.

TABLE OF CONTENTS

DEDICATION.....	i
ACKNOWLEDGEMENT.....	ii
TABLE OF CONTENTS.....	iii
LIST OF TABLES.....	v
LIST OF FIGURES.....	vi
LIST OF APPENDICES.....	vii
ABSTRACT.....	viii
CHAPTER ONE	1
INTRODUCTION	1
1.1 Background.....	3
1.2 Statement of the Problem.....	5
1.3 Justification of the study.....	5
1.4 Objectives.....	5
1.4.1 General Objective.....	6
1.4.2 Specific Objectives.....	6
1.5 Methodology Adopted.....	6
1.5.1 Review of Related Literature.....	6
1.5.2 Data collection.....	7
1.5.3 Data preprocessing.....	7
1.5.4 Training and building the models.....	7
1.5.5 Evaluation of the models.....	7
1.6 Scope and Limitation.....	8
1.7 Thesis Organization.....	9
CHAPTER TWO	9
LITERATURE REVIEW	9
2.1 DATA MINING TECHNOLOGY	11
2.1.1 Data Mining and Knowledge Discovery in Databases.....	11
2.1.2 Data mining process.....	13
2.1.3 Data Warehouse technology for Data mining.....	14
2.1.4 Data Mining Functionalities.....	16
2.1.5 Data Mining Applications.....	17
2.1.5.1 Data Mining Applications in Manufacturing Industry Sector.....	19
2.1.6 Overview of Data Mining technique.....	19
2.1.6.1 Predictive Modeling.....	24
2.2 CEMENT PRODUCTION	25
2.2.1 Overview of Cement.....	25
2.2.2 Cement Production Process.....	26
2.2.2.1 Raw materials.....	27
2.2.2.2 Proportioning, mixing, and Raw Milling.....	29
2.2.2.3 Burning of the Raw mix in kiln to form Clinker.....	30
2.2.2.4 Cement Clinker Composition.....	31
2.2.2.5 Finished milling of Clinker.....	31
2.2.2.6 Special Cements.....	31
2.2.2.7 Supplementary Cementitious materials.....	31

2.3 CEMENT TESTING AND QUALITY ASSESSMENT	32
2.3.1 Overview of Quality	32
2.3.2 Chemical tests	33
2.3.3 Physical tests	34
2.4 CEMENT STANDARDS	37
2.4.1 Overview	37
2.4.2 The British Standard(BS:12)	39
2.4.3 The European Pre-standard (ENV197-1)	41
2.4.4 The American standard for Testing and Materials (ASTM)	46
CHAPTER THREE	46
SURVEY OF CEMENT PRODUCTION AND QUALITY ASSESSMENT AT MCE	46
3.1 Overview	46
3.2 Mughher cement Enterprise	47
3.2.1 Mughher Cement Factory	48
3.2.1.1 Cement Production Process At Mughher Cement Factory	51
3.2.2 Addis Ababa Cement Plant	52
3.2.2.1 Cement Production Process At Addis Ababa Cement Plant	54
3.3 Quality Assessment Process at MCE	56
3.4 Cement Quality Approval Process at MCE	58
3.5 Finding of the survey	60
CHAPTER FOUR	60
EXPERIMENTATION	60
4.1 Overview	60
4.2 Data Mining Goal	61
4.2.1 Data Mining Tool Selection	62
4.3 Data Understanding	64
4.3.1 Initial Data Collection	65
4.3.2 Description of data collected	66
4.3.3 Data quality verification	66
4.4 Data Preprocessing	67
4.4.1 Data Cleaning	68
4.4.2 Data Selection	68
4.4.3 Data Transformation	71
4.5 Modeling	71
4.5.1 Selection of modeling technique	74
4.5.2 Test design	76
4.5.3 Model building	87
4.5.4 Evaluation of models	93
CHAPTER FIVE	93
CONCLUSION AND RECOMMENDATION	93
5.1 Conclusion	94
5.2 Recommendations	96
References	98
Appendices	100
Declaration	

LIST OF TABLES

Table 1: Chemical test variables.....	33
Table 2: Physical test variables.....	34
Table 3: Chemical requirement of BS:12 Standard for OPC.....	38
Table 4: Strength requirements of BS:12 Standard for OPC.....	38
Table 5: Setting time and expansion specification of BS: 12 for OPC.....	39
Table 6: Common cement types of the European Pre-Standard ENV 197-1.....	39
Table 7: The ENV 197-1 notation and designation of constituents of cement.....	40
Table 8: Common cement types and composition defined in the ENV 197-1 standard.....	41
Table 9: Physical requirements of the ENV 197-1 Standard.....	41
Table 10: Chemical requirements of the ENV 197-1 Standard.....	42
Table 11: Ordinary Portland Cement (OPC) types defined in ASTM C-150.....	42
Table 12: Chemical specifications of ASTM C-150 for Ordinary Portland Cements.....	43
Table 13: Physical specifications of ASTM C-150 for Ordinary Portland Cements.....	44
Table 14: Chemical specifications of ASTM C-595 for Blended cements.....	45
Table 15: Physical specifications of ASTM C-595 for Blended cements.....	47
Table 16: Annual cement production history of MCF.....	52
Table 17: Annual cement production history of AACP.....	70
Table 18: OPC data variables used for model building.....	71
Table 19: PPC data variables used for model building.....	78
Table 20: model statistics for tree model of experiment one.....	79
Table 21: Model statistics for ensemble model tree of experiment one.....	80
Table 22: Over all result of experiment one.....	80
Table 23: Performance comparison of models of experiment one.....	81
Table 24: Model statistics for tree model of experiment two.....	81
Table 25: Model statistics for ensemble model tree of experiment two.....	82
Table 26: Over all result of experiment two.....	82
Table 27: Performance comparison of models of experiment two.....	83
Table 28: Model statistics for tree model of experiment three.....	84
Table 29: Model statistics for ensemble model tree of experiment three.....	84
Table 30: Over all result of experiment three.....	84
Table 31: Performance comparison of models of experiment three.....	85
Table 32: Model statistics for tree model of experiment four.....	86
Table 33: Model statistics for ensemble model tree of experiment four.....	86
Table 34: Over all result of experiment four.....	86
Table 35: Performance comparison of models of experiment four.....	88
Table 36: Performance comparison of models of experiment one and two.....	91
Table 37: Performance comparison of models of experiment three and four.....	

LIST OF FIGURES

Figure 1: Sample decision tree.....	21
Figure 2: Clinker production process at MCF.....	48
Figure 3: OPC production process at MCF.....	49
Figure 4: PPC production processes at MCF.....	49
Figure 5: OPC production processes at AACP.....	52
Figure 6 PPC and SRPPC production processes at AACP.....	53

LIST OF APPENDICES

Annex 1: Academic license agreement	98
Annex 2: Physical test result of milled cement form.....	98
Annex 3: Chemical analysis form.....	99
Annex 4: OPC and PPC test certificate forms	99

ABSTRACT

The application of data mining technology has increasingly become very popular and proved to be relevant for many sectors. Proving this, a number of researches have been conducted in the past on the possible application of data mining technology for various economic sectors of Ethiopia, such as the airline, bank, telecommunication, insurance, and health sectors.

Meanwhile, the existing practice of assessing the quality of cement produced at Muger Cement Enterprise involves manual analysis of test results that complete with a day or so, and then predicting the quality. This method is prone to errors, subjective, and demands personnel with many years of experience.

This paper reports on the finding of a research that had the objective to build classification and regression models that provide predictive assistance in cement quality assessment and approval decision making process.

Cement quality test results data were collected for two types of cements: Ordinary Portland Cement (OPC) and Portland Pozzolana Cement (PPC). The OPC and PPC data sets were first preprocessed before the predictive models were built and evaluated. The Rule Discovery System™ Modeling Edition version 2.0 was used to build classification and regression models using tree model and ensemble modeling techniques. The predictive accuracies of the models built were found very encouraging for the models to be adapted in real problem.

The outcome of this research has paved the way for further researches to be conducted in the future to explore the potential applicability of data mining technology for product quality assessment task in other manufacturing industry sectors.

CHAPTER ONE

INTRODUCTION

1.1 BACKGROUND

Of the many sectors that are considered to have great contribution to the development of a country's economic infrastructures, the cement industry is perhaps the leading one, since the development of other important infrastructures, such as roads, water supply, electricity, etc, depend partly on cement. Cement is used to build a wide variety of structures, such as highways and streets, dams, bridges, large buildings, airport runways, and irrigation structures. Strengthening this idea, Hewlett (1998) states that, "In modern times no other material has had so great an influence upon the construction world. In fact, no other building material has made so lasting an impression on the general public as cement-based structures."

Cement is finely powdered substance that hardens and adheres after being mixed with water. The hardening of the cement is caused by hydration (reaction with water). The hydration of cement proceeds slowly and the ultimate strength is attained over a period of weeks. A fundamental property of cements is that they set and generate strong materials when stored under water. Compressive strength is used as a means of categorizing cement quality (Hewlett, 1998).

Categorizing cement quality can take the form of assessments at early (3 days) or late (28 days) ages or both. However, most cement standards specify the 28 days strength as major criteria for cement quality. The European Pre-standard ENV 197-1, for example, places primary emphasis upon the 28 day strength and for this purpose introduces three classes: 32.5, 42.5, and 52.5; these numbers represent the minimum characteristic strength, in N/mm^2 , which the cement is required to achieve at 28 days.

The strength-giving properties of cement need to be carefully controlled and a number of physical and chemical tests are used to ensure the quality of cement. However, the most important strength test requirements are those at 28 days, and great deal of cement (perhaps 500,000 ton at very large factories) can be produced before a 28 days strength results becomes available. A well-equipped cement factory laboratory and trained staff are necessary in order to control production and to make realistic estimate of the 28 days strength at the time the product is manufactured (Hewlett, 1998).

Mugher Cement Enterprise (MCE) is a state owned enterprise that has been engaged in the production and supply of cement. The total design capacity of cement plants of the enterprise is 700,000 tons of cement per year. However, the production capacity of MCE has been increasing over the past years mainly due to the strategic decision made to utilize Pumice (a naturally occurring cement additive) in the production process. According to the annual report of 2004, the Enterprise has produced over one million tons of cement.

While MCE has been successful in increasing its cement production capacity, at the same time there is a very high demand for cement in the country. This situation has created a conflict of interests. Firstly, due to success in increasing its production capacity, the enterprise sometimes produces cement equal to the full capacity of its cement silos can handle, thus preventing further production process from taking place because tests, that verify quality of cement in silos, require a number of days to complete and the silos would have to hold the cement until the results are made available. Secondly, there is a need to deliver cement produced daily to customers to satisfy the high demand, but testing and verifying quality cement require a number of days.

With effective analysis of cement quality variables, it is possible to predict results of tests that require long time to complete based on test results that complete within a day or so. But, analyzing test results and predicting cement quality manually is difficult for humans due to the size of the variables and the complex relationships. Hewlett (1998) claims that cement is a multi-component system, its hydration is rather complex process consisting of a series of individual chemical reactions that take place both in parallel and successively.

However, data mining technology can be applied as an automatic process of discovering patterns in cement quality test data, by analyzing a number of cement quality variables. If sufficient data regarding cement quality test records are collected, data mining may provide information on the 28 days compressive strength, and, hence, quality of cement by analyzing and identifying different patterns.

Data mining is the process of digging through and analyzing enormous sets of data, finding undiscovered patterns, and then deriving some meaning from them (Seidman, 2001). According to him data mining is a decision support tool that stands on its own when it comes to analyzing large databases. It has its own unique features, which are designed to address unique decision support problems that cannot be solved by other data analysis tools.

Businesses generate vast amount of data through their daily operations, such as production, marketing, etc., that has the potential, when combined with historical data, to tell a company where it is and where it is going; yet, a lot of businesses are "sitting on this data" (Henock, 2002). Intelligently analyzed business data, according to Witten & Frank (2000), is a valuable resource that can lead to new insights and, in commercial settings, to competitive advantage. In today's highly competitive, customer centered, and service oriented economy, data is the raw material that fuels business growth - if only it can be mined.

This research, based on the arguments stated above, attempted to apply the data mining technology on historical cement quality test records of MCE and experimented on the possibility of building models capable of providing predictive information about quality of cement produced daily.

1.2 STATMENT OF THE PROBLEM

The availability of sufficient quantity of cement is a prerequisite for infrastructure development. However, quality of cement is more critical than the quantity since low quality cement can result in failure or collapse of buildings, bridges, etc., which may take out lives of many, and huge loss of investment. Therefore, every batch of manufactured cement must be tested to assess its conformity to the quality requirements of standard specifications that have been developed to enable manufacturers control their production.

The overall quality of cement is governed by many physical and chemical variables. The physical variables include setting time, fineness and 3, 7, and 28 days compressive strength. Chemical variables include percent Pozzolana, Insoluble residue, Loss on ignition, Sulfur trioxide, Silicon dioxide, Magnesium oxide, etc. However, most of the widely adopted cement standards place primary emphasis upon the 28 days strength as a means of categorizing cement quality and great deal of cement can be produced before a 28 day strength test results become available.

To avoid the lengthy time requirements of the compressive strength tests being a potential bottleneck in the production and sales operations, cement factories often have redundant silos to store batch of cement until tests are predicted or completed. However, this redundancy in number of silos is a solution that can be implemented when the factory is erected and will not work once the factory, as in the case of MCE, has reached a production capacity more than its design capacity.

To further complicate the problem, currently and in foreseeable future there is a very high demand for cement in Ethiopia that every batch of cement produced daily need to be delivered to customers in order to satisfy the demand, but the lengthy time required to test and verify quality of cement becomes inhibitory.

To address this situation, the only solution available to MCE has been the practice of manually analyzing test results and predicting the 28 days strength for the purpose of making strategic decision about the quality of batch cement produced daily. In other words, soon after the 3 days compressive strength, chemical and other physical test results are obtained for batch of cements produced daily, quality control experts manually analyze the results and predict as to what the 28 days strength results would be and then authorize the cement for sale even before the 7 days strength results are obtained.

However, this manual analysis of test results and prediction of cement quality is not only difficult for humans due to the size of the cement quality variables and the complex relationships, but also subjective and purely heuristic that demands quality control experts' years of experience.

The most serious problem with this manual analysis is that it is prone to errors. Firstly, batch of cement that is predicted and judged to qualify the minimum quality requirements may turn out to be otherwise when the 7 and/or 28 days strength result are later made available. By the time this knowledge is gained, the low quality cement has already been utilized to construct structures of huge investment and that support lives of many. Secondly, batch of cement may also be predicted to be of low quality while it is not. This results in delaying good quality cement from delivery while there is very high demand for cement. It also results in expensive corrective measures to be taken in the production process unnecessarily. Both cases have serious effects on the country's economy and lives of the citizens.

Therefore, the cement quality assessment and approval decision making process of MCE need to be supported with an automated system that is more accurate, less subjective and requiring less expertise.

The quality control laboratories of MCE perform tests on batches of cements produced daily. The test results then are stored simply being considered as by-product of day-to-day operations of the laboratories; that is, test results are not used for any purpose other than deciding the quality of a particular batch of cement. However, these collected and stored cement quality data seem to have the potential to understand more and provide predictive information on quality of cement for which test results are not yet completed.

This research attempted automating the cement quality assessment task by building classification and regression models using data mining technology on historical cement quality test records. The predictive models provide cement quality assessment and approval decision support assistance thereby facilitating the smooth operations of the Enterprise.

The models enable to minimize the potential risk of delivering inferior quality cement to customers thereby potentially saving loss of expensive investment and lives of people. In addition, the models provide more reliable early warning information by detecting the production of cement that is likely to be of low quality so that measures can be taken in time.

The research give opportunity for the utilization of cement test data that is stored for no purpose, and the models developed also enable less qualified quality control staff to perform cement quality assessment and approval tasks more accurately and confidently and less subjectively.

1.3 JUSTIFICATION OF THE STUDY

A number of researches have been conducted on the possible application of data mining technology for various economic sectors of Ethiopia, such as the airline, bank, telecommunication, insurance, and health sectors.

However, to the knowledge of the researcher, the possible application of data mining technology to the manufacturing sector has not been experimented. It was, therefore, the researcher belief that the study would add to the existing body of knowledge pertaining to the applications of data mining technology.

1.4 OBJECTIVES

1.4.1 GENERAL OBJECTIVE

The general objective of this research is to explore the potential applicability of data mining technology in developing predictive models that can support cement quality assessment and approval decision making process at Mughher Cement Enterprise.

1.4.2 SPECIFIC OBJECTIVES

In order to accomplish the above stated general objective, the research work has carried out the following specific objectives:

- Examine in detail the existing cement quality assessment and approval practices at MCE;
- Study the cement quality standard currently utilized by MCE;
- identify the types of cement quality test results data available at MCE;
- Select the data mining tool and algorithm to be used that support predictive modeling;
- Collect data on which the data mining processes will be carried out;
- Preprocess the data for models building;
- Build and train predictive models;
- Evaluate the models;
- Analyze the outcome of the research and make recommendations based on findings.

1.5 METHODOLOGY ADOPTED

To conduct the research, the following methods were employed:

1.5.1 REVIEW OF RELATED LITERATURE

To gain insight into the subject matters of data mining technology, cement production, cement standards, and cement testing and quality assessment, books, research papers, and materials on the Internet related to the research tasks were reviewed.

1.5.2 DATA COLLECTION

The source of data for the research was obtained from quality control laboratory of Addis Ababa Cement Plant (AACP), the branch factory of the MCE. The data sources contain cement quality test records that were used to verify the quality of cement produced daily and latter stored for no other reason. The data contain the test results of all the chemical and physical parameters tested for two major types of cements (Ordinary Portland Cement [OPC] and Portland Pozzolana Cement [PPC]) that are produced by AACP. More than 1,081 test data records for OPC and 1,271 test records for PPC were collected for this research. All of the data for both OPC and PPC were numerical records.

1.5.3 DATA PREPROCESSING

As a prerequisite to the data mining process, data preprocessing tasks including data cleaning, integration, selection, and transformation were performed on the collected OPC and PPC data. The collected data were analyzed to remove data that were irrelevant for the objective of the research. Finally, the input variables in the OPC and PPC data sets that have significant weight for the research were identified and the OPC and PPC data sets converted in a tab-separated text formats that are appropriate to data mining tool used to build the models.

1.5.4 TRAINING AND BUILDING THE MODELS

The resulting preprocessed OPC and PPC data sets were used to build the models that answer research objectives. However, before the actual model building took place, tree model and ensemble tree model were selected as the modeling techniques, classification and regression models were identified as the types of models to be built, and 10-fold cross validation was determined as validation method for evaluating the models' predictive accuracy. Then, several classification and regression models were built for both OPC and PPC data sets using Rule Discovery System (RDS™) software selected for this research.

1.5.5 EVALUATION OF THE MODELS

The models were evaluated for their predictive performance. Measures of Accuracy, Precision, Recall, Confusion Matrix, Correlation Coefficient, Mean Error, Mean Squared Error and Performance comparison matrix were used to evaluate the predictive performances of the models and to assess the outcomes of the research. Interpretability of the models, in terms of soundness of rules generated, was also additional criteria utilized for evaluating the models developed.

1.6 SCOPE AND LIMITATION

The scope of this research is to investigate the potential applicability of data mining technology to support cement quality assessment and approval decision making process at MCE. Even though the results of the research can be adapted to cement raw materials quality assessment and cement production process control tasks of MCE, this research is limited to explore the possible applicability of data mining in predicting the quality of cement once it has been produced and preliminary test results are made available.

In addition, the outcome of the research is restricted to the OPC and PPC historical test results data that were collected to conduct this research. Due to shortage of time, it was not possible to test the models using new test result data and then compare and contrast the 28 days strength results or cement quality predicted by the models with the actual test results values that are obtained after 28 days.

1.7 ORGANIZATION OF THE THESIS

The thesis is organized into five chapters. The first chapter presents introduction of the research. The second chapter deals with review of related literature divided into four sections. The first section discusses the data mining technology and its relationship with the current research problem. Section two presents overview on cement raw materials and cement production process. The third section describes cement testing and quality assessment. Standard cement specifications are presented in the final section. Chapter three reports the survey of conducted to give insight on the existing cement production and quality assessment procedures at MCE. The fourth chapter illustrates the step-by-step tasks executed while conducting the experiment. The final chapter is dedicated for the conclusion and recommendations based on the findings of the research.

CHAPTER TWO

LITERATURE REVIEW

2.1 DATA MINING TECHNOLOGY

Our capabilities of both generating and collecting data have been increasingly in the last several decades. Contributing factors include the wide spread use of bar codes, the computerization of many business, scientific and governmental transactions, advances in data collection tools ranging from scanned text and image platforms to satellite remote sensing systems. In addition, popular use of the World Wide Web as a global information system has flooded us with a tremendous amount of data. This explosive growth in stored data has generated an urgent need for new techniques and automated tools that can assist us in transforming the vast amounts of data into useful information and knowledge (Han and Kamber, 2001).

Gray (2001) shares the above idea that we are deluged by data- scientific data, medical data, demographic data, marketing data, etc. People have no time to look at this data. Human attention has become a precious resource. We must find ways to automatically analyze, classify, summarize, discover trends in it, and to automatically flag anomalies.

Han and Kamber (2001) also having similar view explain that the abundance of data, coupled with the need for powerful data analysis tools, has been described as a data rich but information poor situation. The fast growing, tremendous amount of data, collected and stored in large and numerous databases, has far exceeded our human capability for comprehension without powerful tools.

In relation to the urgent need for data mining to address these concerns, Witten and Frank (2000) state that "As the world grows in complexity, overwhelming us with data it generates, data mining becomes our only hope for elucidating the patterns that underlie it." Describing the potentials of data mining, Seidman (2001) also states that data mining leverages the computing power of today's servers to transform mountains of raw data into useful information. The ever increasing physical storage capacity of computers, coupled with easy access to powerful processing muscle, make sophisticated data analysis possible in a way that were unimaginable only a few years ago.

Data mining derives its name from the similarities between searching for valuable business information in a large database- for example, finding linked products in gigabytes of store scanner data- and mining a

mountain for a vein of valuable ore. Both processes require either sifting through an immense amount of material, or intelligently probing it to find exactly where the value resides (Thearling, 2003).

The term data mining is defined more or less similarly in several literatures. Some of these include:

- The exploration and analysis of large quantities of data by automatic or semiautomatic means in order to discover meaningful patterns and rules (Berry et al., 1997).
- The efficient discovery of valuable, non-obvious information from a large collection of data (Bigus, 1996).
- An analytic process designed to explore large amounts of data in search of consistent patterns and/or systematic relationships between variables, and then to validate the findings by applying the detected patterns to new subsets of data (StatSoft, Inc., 2003).
- An information extraction activity to find patterns and subtle relationships in data and infer rules that allow the prediction of future results (Two Crows Corporation, 1999).
- The task of discovering interesting patterns from large amount of data where the data can be stored in databases, data warehouses, or other information repositories (Han and Kamber, 2001).
- The process of discovering meaningful patterns and relationships through the automated analysis of large stores of historical data (Seidman, 2001).

From the above definitions, the idea of data mining revolves around finding information that is hidden in large data, which is new and hard to grasp by human being.

Unfolding the steps in the evolution of data mining, Thearling (2003) clarifies that data mining techniques are the result of a long process of development in research areas such as Statistics, Artificial intelligence, and Machine learning. This evolution began when business data was first stored on computers, continued with improvements in data access, and more recently, generated technologies that allow users to navigate through their data in real time. Data mining takes this evolutionary process beyond retrospective data access and navigation to prospective and proactive information delivery. Data mining is ready for application in the business community because it is supported by three technologies that are now sufficiently mature: massive data collection, powerful multiprocessor computers, and data mining algorithms.

2.1.1 DATA MINING AND KNOWLEDGE DISCOVERY IN DATABASES.

The term data mining is used as a synonym for Knowledge Discovery in Databases (KDD) in some data mining literatures. For example, Seidman (2001) uses the term data mining interchangeably with KDD. On the other hand, Han and Kamber (2001) and Two Crows Corporation (1999) view KDD as the overall process of discovering useful knowledge from data, and data mining as a particular but essential step in the process of knowledge discovery in databases. According to them, KDD is an iterative sequence of steps, which are data cleaning, data integration, data selection, data transformation, data mining, pattern evaluation, and Knowledge presentation.

In this research paper, however, the terms data mining and knowledge discovery process both refer to the entire process from data collection through pattern identification and reporting.

2.1.2 DATA MINING PROCESS

Recognizing that a systematic approach is essential to successful data mining, many vendors and consulting organizations, such as SPSS, SAS and Two Crows Corporation, have specified a process model designed to guide the user through a sequence of steps that will lead to good results (Two Crows Corporation, 1999).

The process model for data mining (or Knowledge Discovery process) provides an outline of the life cycle of a data mining project and serve as blueprints for how to organize the process of data mining from the idea to final implementation.

Based on the process models of CRISP-DM (2000), Two Crows Corporation (1999), and works of Han and Kamber (2001) and Seidman (2001), the generally established steps of data mining process is discussed below. Yet, even if the steps appear in a list, the data mining process is not linear; that is, the steps are not always performed in exact order. In fact, data mining is an iterative process - one will inevitably need to loop back to previous steps.

DEFINING THE BUSINESS PROBLEM

According to Two Crows Corporation (1999), understanding the data and business is the first and foremost prerequisite to knowledge discovery. This background will enable to identify the problems that need to be

solved, prepare the data for mining, and correctly interpret the results. Therefore, to make the best use of data mining one must make a clear definition of the business problem and the objective function, as this may direct not only the KDD process but also the data mining modeling involved.

Concerning the issue of analyzing the business problem, Seidman (2001), suggests that the data sources have to be assessed to see whether they meet the criteria for data mining and explains quality and abundance of the data as the primary factors in deciding whether the data is at all suitable. In addition, the expected results of the data mining effort have to be carefully understood to be sure that the existing data contains the right kind of extractable information.

DATA SELECTION AND COLLECTION

At this step of the sources of the data one will be mining is identified and selected. The selection process is important because data sources are heterogeneous, containing a wide Variety of data, not all of which may be appropriate for the analysis at hand. The data relevant to the analysis task are then extracted from its native's sources, such as databases, text files and spreadsheets. Usually, this data is placed in a database or data warehouse; however, depending on the amount of the data, the complexity of the data, and the uses to which it is to be put, a flat file or even a spreadsheet may be adequate (Two Crows Corporation, 1999).

DATA PREPROCESSING

This step constitutes the core of the data mining and take anywhere from 50% to 90% of the time and effort of the entire knowledge discovery process (Two Crows Corporation, 1999). Data preprocessing, according to Han and Kamber (2001), is an important issue for data mining as real-world data tend to be incomplete, noisy, and inconsistent. Data preprocessing techniques improve the quality of the data, thereby helping to improve the accuracy and efficiency of the subsequent mining process.

Data preprocessing include data cleaning, data integration, data transformation and data reduction. Data cleansing routines are used to fill in missing values, smooth noisy data, identify outliers, and correct data inconsistencies. Data integration combines data from multiple sources to form a coherent data store. Data transformation routines convert the data into appropriate forms for mining. Data reduction techniques enable to obtain a reduced representation of the data, while minimizing the loss of information content (Han and Kamber, 2001).

MODEL BUILDING

Model building or data mining is a crucial process where intelligent methods are applied in order to extract patterns (Han and Kamber, 2001). An important function of data mining, also referred to as the model-building step, is the production of a model. A model can be descriptive or predictive. A descriptive model helps in understanding underlying processes or behavior. A predictive model is an equation or set of rules that makes it possible to predict an unseen or unmeasured value (the dependent variable or output) from other, known values (independent variables or input) (Two Crows Corporation, 1999).

The most important thing to remember about model building, according to Two Crows Corporation (1999), is that it is an iterative process. There will be a need to explore alternative models to find the one that is most useful in solving business problem. Searching for a good model may lead to go back and make some changes to the data or even modify problem statement.

VALIDATING THE MODEL

In the validation phase of the KDD process, the model is tested for accuracy. Two Crows Corporation (1999) state that after building a model, one must evaluate its results and interpret their significance.

DEPLOYING THE MODEL

After the model is trained and validated, it is used to analyze new data and make predictions. This use of the model is called deployment (Two Crows Corporation, 1999). 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. That is, the knowledge gained from data will need to be organized and presented in a way that the users can use it in decision making processes.

2.1.3 DATA WAREHOUSE TECHNOLOGY FOR DATA MINING

Since the 1960's, database and information technology has been evolving systematically from primitive file processing systems to sophisticated and powerful database systems. Data can now be stored in many different types of databases. One database architecture that has recently emerged is the data warehouse

(Hans and Kamber, 2001). Data warehouse is defined as "a subject-oriented, integrated, time-variant, and non-volatile collection of data, which supports decision making process in enterprise management." (W.H.Inmon, 1996) as quoted in (Han and Kamber, 2001).

According to Two Crows Corporation (1999), most of the time the data to be mined is first extracted from an enterprise data warehouse into a data mining database or data mart. There is some real benefit if the data mining database is already part of a data warehouse. A data mining endeavor includes the effort to identify, acquire, and cleanse data. The problems of cleansing data for a data warehouse and for data mining are very similar. If the data has already been cleansed for a data warehouse, then it most likely will not need further cleaning in order to be mined. In addition, if the design of the data warehouse includes support for data mining applications, the warehouse facilitates the data mining efforts.

Hans and Kamber (2001) also have similar view to wards significance of data warehousing for data mining. The construction of data warehouses, which involves data cleaning and integration, can be viewed as an important preprocessing step for data mining. Moreover, data warehouses provide On-Line Analytical Processing (OLAP) tools for the interactive analysis of multidimensional data of varied granularities, which facilities effective data mining. Still, many other data mining functions, such as prediction, association, and clustering, can be integrated with OLAP operations to enhance interactive mining of knowledge. Hence, the data warehouse has become an increasingly important platform for data analysis and on-line analytical processing and will provide an effective platform for data mining.

2.1.4 DATA MINING FUNCTIONALITIES

Data mining functionalities are used to specify the kind of patterns to be found in data mining tasks. In general, data mining tasks can be classified into two categories: descriptive and predictive. Descriptive mining tasks characterize the general properties of the data in the database. Predictive mining tasks perform inference on the current data in order to make prediction (Han and Kamber, 2001).

Several data mining functionalities are found in data mining literature. According to Han and Kamber (2001), Two Crows Corporation (1999), and Thearling (2003), the most common functionalities include:

CONCEPT/CLASS DESCRIPTION

Concept description is the most basic form of descriptive data mining. It describes a given set of task relevant data in concise and summarized form, presenting general properties of the data. Concepts (or class) description consists of characterization and comparison (or discrimination). The former summarizes and distinguishes one collection of data, called the target class, whereas the latter summarizes and distinguishes one collection of data, called the target class, from other collection(s) of data, collectively called the contrasting class(es) (Han and Kamber, 2001).

ASSOCIATION ANALYSIS

Association analysis is the discovery of association rules showing attribute-value conditions that occur frequently together in a given set of data. Association analysis is widely used for market basket or transaction data analysis (Han and Kamber, 2001). Association analysis, according to Two Crows Corporation (1999), is a descriptive approach to exploring data that can help identify relationships among values in a database. The two most common approaches are association discovery and sequence discovery. Association discovery finds rules about items that appear together in an event. Market-basket analysis is a well-known example of association discovery. Sequence discovery is very similar in that a sequence is an association related over time.

PREDICTION

Prediction can be viewed as the construction and use of a model to assess the class of unlabeled sample is likely to have. In this view, classification and regression are two major types of prediction problems. Whereas classification predicts categorical labels, prediction models continuous-valued functions (Hans and Kamber, 2001).

The present research is concerned with building classification and regression predictive models using decision trees. Because of this, predictive data mining activity using decision tree, which will be used in the experimental part of this research, is discussed in more detail in section 2.1.6 of this chapter.

CLUSTER ANALYSIS

Clustering is the process of dividing a dataset into mutually exclusive groups such that the members of each

group are as "close" as possible to one another, and different groups are as "far" as possible from one another, where distance is measured with respect to all available variables (Thearling, 2003).

According to Two Crows Corporation (1999), clustering divides a database into different groups or categories. The goal of clustering is to find groups that are very different from each other, and whose members are very similar to each other. Unlike classification, one don't know what the clusters will be in the start, or by which attributes the data will be clustered. Since the categories are unspecified, this is sometimes referred to as unsupervised learning.

EVOLUTION/TIME SERIES ANALYSIS

Data evolution analysis describes and models regularities or trends for objects whose behavior changes over time. Although this may include characterization, discrimination, association, classification, or clustering of time-related data, distinct features of such analysis include time-series data analysis, sequence or periodicity pattern matching and dissimilarity based data analysis (Han and Kamber, 2001).

Time series analysis, according to Two Crows Corporation (1999), predicts unknown future values based on a time-varying series of predictors. Like regression, it uses known results to guide its predictions. Models must take into account the distinctive properties of time; especially the hierarchy of periods (including such varied definitions as the five- or seven-day work week), seasonality, and calendar effects, such as holidays, and special considerations, such as how much of the past is relevant.

2.1.5 DATA MINING APPLICATIONS

Organizations apply data mining technology for two main reasons: to make sense of their past and to make predictions about their future. Ultimately, they use the information taken from this computerized fortuneteller to make decisions about their futures. Data mining is an activity that offers business advantages, as well as solutions to some mounting problems associated with exploring the knowledge embedded within corporate databases. Data mining is particularly valuable for organizations that collect large quantities of historical information. Banks, insurance companies, and even astronomers use this technology to derive critical information from large, unwieldy data samples. Interestingly, data mining is used in many fields, such as medical diagnosis and metrology. In short, the practical applications are endless- any business or academic pursuit that collects and studies large quantities of data is a candidate for data mining (Seidman, 2001).

Data mining offers great promise in helping organizations uncover patterns hidden in their data that can be used to predict the behavior of customers, products and processes. Data mining offers value across a broad spectrum of industries. Telecommunications and credit card companies are two of the leaders in applying data mining to detect fraudulent use of their services. Insurance companies and stock exchanges are also interested in applying this technology to reduce fraud. Medical applications are another fruitful area: data mining can be used to predict the effectiveness of surgical procedures, medical tests or medications. Pharmaceutical firms are mining large databases of chemical compounds and of genetic material to discover substances that might be candidates for development as agents for the treatments of disease (Two Crows Corporation, 1999).

The major reason that data mining has attracted a great deal of attention in the information industry in recent years is due to the wide availability of huge amount of data and the imminent need for turning such data into useful information and knowledge. Data patterns can be mined from many different kinds of databases. Interesting patterns can also be extracted from other kinds of information repositories, such as spatial, time-related, text, multimedia, and the World Wide Web. The information and knowledge gained can be used for applications ranging from business management, production control, and market analysis, to engineering design and science exploration (Han and Kamber, 2001).

The data mining, as described in the above paragraphs, has potential applications in many sectors. Since the research at hand is concerned with exploring the potential application of data mining in the product quality assessment in the manufacturing industry sector, the next sub-section reviews these possible applications.

2.1.5.1 DATA MINING APPLICATIONS IN MANUFACTURING INDUSTRY SECTOR

Several researches have been conducted to explore the potential application of data mining technology in the manufacturing industry sector to support tasks, such as process control, process optimization and product quality control. Two examples of such researches with their objectives are summarized as follows.

■ Data mining application at a steel making factory

The application of Data Mining technology at steel making factory has been explored by Jeon (2000). According to Jeon (2000), hot blast stoves are important to energy consumers. Hot blast sensible heat supplies 10%-15% of the total energy consumed by the blast furnaces. Reduction in energy consumption for hot blast heating has an important impact on hot metal production cost.

With the objective of developing feed-forward guidance simulator based on Neural Network Models and multi-stage combustion pattern control systems, Jeon (2000) applied Data Mining technology to the Pohang Steel Works, Korea, and reported that the introduction of these systems required only minimal operation costs, decreasing the hot blast stove energy costs approximately 15%. In Conclusion, Jeon (2000) states that high productivity stable operation has been achieved at Pohang No. 3 blast furnace through following activities:

- Optimization of Heat Efficiency: Automatic Combustion control system, Calorie Control System.
- Application of New Control Concept: Multi Stage Combustion Control System.
- Optimization of Operation Condition using the Data Mining
- Increase of Model Accuracy using Neural Network Model

■ Data mining application in engineered wood products manufacturing

With the goal of introducing a regression tree method called GUIDE, Kim, Guess, and Young (2002) conducted a case study in the international manufacture of highly reliable engineered wood products. According to them, GUIDE is one of the most powerful regression modeling tools in Data mining. They explain that medium density fiberboard (MDF) is a highly engineered wood composite that is used extensively by the home building and furniture manufacturing industries around the world. They further explain that destructive testing is performed during the MDF manufacturing process. This provides quick feedback to the manufacturer in order to minimize process inputs and maximize product quality within specified limits. Samples from a cross section of the MDF are tested by being pulled apart. The internal bond (strength) of MDF is then measured in pounds per square inch until failure.

The objective of the case study was to use the GUIDE regression tree method to better predict and/or interpret the internal bond (strength) of MDF and assess its reliability. The cost of unacceptable MDF was large as 5% to 10% of total manufacturing costs. Prevention of unacceptable reliability can result in millions of US dollars saved by using better information systems. The regression model, built using 1,477 observations with many detailed variables that describe MDF, correctly detected unusual relationships between the response and the predictors (Kim, Guess, and Young, 2002).

2.1.6 OVERVIEW DATA MINING TECHNIQUE

In this section an attempt has been made to review classification and regression predictive modeling. Emphasis is given to decision tree technique that would be used in the experimental part of the research to build models that would answer the research objective.

2.6.1.1 PREDICTIVE MODELING

The term *Predictive Data Mining* is usually applied to identify data mining projects with the goal to identify a statistical or neural network model or set of models that can be used to predict some response of interest (StatSoft, 2003).

According to Compumine (2004), Predictive modeling is the process of creating predictive models from data. The data is normally stored in a table and consists of example cases, or simply examples, stored as rows, with the properties of the examples as the columns. The purpose of predictive modeling is to create a model that can be used to predict one of the properties of an example provided that the other properties are known. The property one wants to predict is typically something that is not easily available for the example, such as the experimental outcome of a chemical assay. The properties based on which the prediction should be made are normally relatively easily obtained. These properties are commonly referred to as the predictor, independent or background variables- variables that can be used for creating rules of a prediction model. The variable that generated model will try to predict is called prediction, dependent or target variable.

Predictive models are built, or trained, using data for which the value of the response variable is already known. This kind of training is sometimes referred to as supervised learning, because calculated or estimated values are compared with the known results. By contrast, descriptive techniques, such as clustering, are sometimes referred to as unsupervised learning because there is no already-known result to guide the algorithms (Two Crows Corporation, 1999).

Prediction can be viewed as the construction and use of a model to assess the class of unlabeled sample is likely to have. In this view, classification and regression are two major types of prediction problems. Whereas classification predicts categorical labels, regression predicts continuous-valued functions (Hans and Kamber, 2001). Two Crows Corporation (1999), also agree on two types of predictions: classification: predicting into what category or class a case falls, and regression: predicting what number value a variable will have.

Two Crows Corporation (1999) further explain that classification problems aim to identify the characteristics that indicate the group to which each case belongs. This pattern can be used both to understand the existing data and to predict how new instances will behave. Data mining creates classification models by examining already classified data (cases) and inductively finding a predictive pattern. These existing cases may come from an historical database, such as people who have already undergone a particular medical treatment. They may come from an experiment in which a sample of the entire database is tested in the real world and the results used to create a classifier.

Han and Kamber (2001) describe classification similarly as the process of finding a set of models (or functions) that describe and distinguishing data classes or concepts, for the purpose of being able to use the model to predict the class of objects whose class label is unknown. According to them, classification is a two-step process. In the first step, a model is built describing a predefined set of data classes or concepts. In the second step, the model is used for classification.

According to Two Crows Corporation (1999), regression uses existing values to forecast what other values will be. In the simplest case, regression uses standard statistical techniques, such as linear regression. Unfortunately, many real-world problems are not simply linear projections of previous values; they may depend on complex interactions of multiple predictor variables. Therefore, more complex techniques (e.g., logistic regression, decision trees, or neural nets) may be necessary to predict future values.

Hans and Kamber (2001) assert decision trees as one of the most commonly used algorithms used to perform classification. In line with Two Crows Corporation (1999), decision trees can also be used to perform regression. Actually, there are two main types of decision trees: classification trees and regression trees. Decision trees, which are used to predict categorical variables, are called classification trees because they place instances in categories or classes. Decision trees used to predict continuous variables are called *regression trees*.

DECISION TREE FOR CATEGORY PREDICTION

A decision tree is a predictive model that, as its name implies, can be viewed as a tree. Specifically each branch of the tree is a classification question and the leaves of the tree are partitions of the dataset with their classification (Thearling, 2003).

Decision trees are a way of representing a series of rules that lead to a class or value. For example, one may wish to classify cement quality as good or bad. Figure 1 below shows a simple decision tree that solves this problem while illustrating all the basic components of a decision tree: the decision node, branches and leaves.

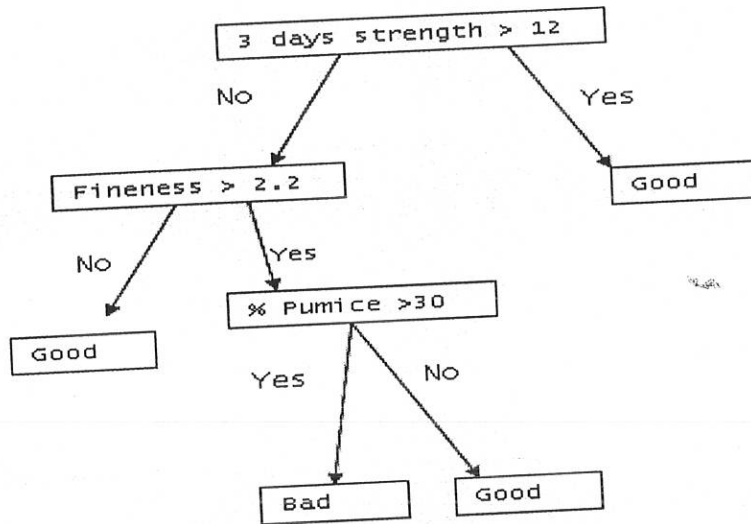


Figure 1: Sample decision tree

The first component is the top decision node, or root node, which specifies a test to be carried out. The root node in this example is "3 days strength > 12". The results of this test cause the tree to split into branches, each representing one of the possible answers. Each branch will lead either to another decision node or to the bottom of the tree, called a leaf node. By navigating the decision tree, one can assign a value or class to a case by deciding which branch to take, starting at the root node and moving to each subsequent node until a leaf node is reached. Each node uses the data from the case to choose the appropriate branch.

Decision trees models are commonly used in data mining to examine the data and induce the tree and its rules that will be used to make predictions. A number of different algorithms may be used for building decision trees including CHAID (Chi-squared Automatic Interaction Detection), CART (Classification And Regression Trees), Quest, and C5.0 (Two Crows Corporation, 1999). According to Seidman (2001), these decision tree algorithms are recommended for predictive tasks that require a classification-oriented model, and as such they are designed for problems best solved by segregating cases into discrete groups.

The training process that creates the decision tree is often called induction. According to Han and Kamber (2001), the basic algorithm for decision tree induction is a greedy algorithm that constructs decision trees in a top-down recursive divide-and-conquer manner as follows:

- The tree starts as a single node representing the training samples
- If the samples are all of the same class, then the node becomes a leaf and is labeled with that class
- Otherwise, the algorithm uses entropy-based measure known as information gain as a heuristic for selecting the attribute that best separates the samples into individual classes. This attribute becomes the 'test' or 'decision' attribute at the node.
- A branch is created for each known value of the attribute, and the samples are portioned accordingly
- The algorithm uses the same process recursively to form a decision tree for the samples at each partition. Once the attribute has reached a node, it need not be considered in any of the node's descendants
- The recursive partitioning stops only when any one of the following condition is true:
 - All samples for a given node belong to the same class
 - There are no remaining attributes on which the samples may further partitioned
 - There are no samples for the branch test-attribute

When a decision tree is built, many of the branches will reflect anomalies in the training data due to noise or outliers. Tree pruning methods address this problem of over-fitting the data. Such methods typically use statistical measures to remove the least reliable branches, generally resulting in faster classification and an improvement in the ability of the tree to correctly classify independent test data (Han and Kamber, 2001).

According to them, there are two common approaches to tree pruning: Prepruning approach and Postpruning approach. Prepruning is when a tree is "pruned" by halting its construction early. Upon halting the nodes become leafs. The leaf may hold the most frequent class among the subset, samples or the probability distribution of those samples. Postpruning removes branches from a "fully grown" tree. However, Prepruning and Postpruning may be interleaved for a combined approach. Postpruning requires more computation than Prepruning, yet generally leads to a more reliable tree.

Han & Kamber (2001) further elaborate that the knowledge represented in decision trees can be extracted and represented in the form of classification 'if-then' rules. One rule is created for each path from the root to a leaf node. Each attribute-value pair along a given path forms a conjunction in the rule antecedent. The leaf node holds the class prediction, forming the rule consequent. The 'if-then' rules may be easier for humans to understand, particularly if the given tree is very large.

DECISION TREES FOR NUMERIC PREDICTION

Whitten and Frank (2000) affirm that the same kind of decision tree and rule representation can be used for numeric prediction, but the leaf nodes of the tree, or the right-hand side of the rules, would contain a numeric value which is the average of all the training set values that the leaf, or rule, applies to. According to them, because statisticians use the term regression for the purpose of computing an expression that predicts a numeric quantity, decision trees with averaged numeric values at the leaves are called *regression trees*. A linear regression model that predicts the class value of instances that reach the leaf is called a *model tree*. Model tree is combination of regression equation and regression tree; that is, it is a tree whose leaves contain linear expressions (regression equation) rather than single predicted value. The model tree approximates continuous functions by linear 'patches', a more sophisticated representation than either linear regression or regression tree.

Regression and model trees are constructed by first using a decision tree induction algorithm to build an initial tree. However, whereas most decision tree algorithms choose the splitting attribute to maximize the information gain, it is appropriate for numeric predictions to instead minimize the intra-subset variation in the class values down each branch. The only difference between regression and model tree induction is that for the latter, each node is replaced by a regression plane instead of a constant value. When a model tree is used to predict the value for test instance, the tree is followed down the leaf in the normal way, using the instance's attribute values to make routing decisions at each node. The leaf will contain a linear model based on some of the attribute values, and this is evaluated for the test instance to yield a raw predicted value (Whitten and Frank, 2000).

TECHNIQUES FOR ESTIMATING ACCURACY OF PREDICTIVE MODEL

According to Han and Kamber (2001), estimating the accuracy is important in that it allows one to evaluate how accurately a given model will predict future data and also helps in the comparison of different models. Using training data to derive a model and then to estimate the accuracy of the model can results in misleading

over optimistic estimates due to over simplification (over-fitting) of the learning algorithm (or model) to the data.

Techniques for estimating model accuracy, according to Han and Kamber (2001), Whitten and Frank (2000), and Seidman (2001) include hold-out, K-fold cross validation, bootstrapping, and leave-one-out. Techniques for increasing model accuracy include bagging and boosting.

MEASURING PERFORMANCE OF MODEL

For classification model, accuracy is the number of correctly predicted test examples divided by the total number of test examples. Han and Kamber (2001) and Whitten and Frank (2000) argue that accuracy is not enough to judge performance of a classification model and suggest that sensitivity, specificity and precision as useful measures in addition to accuracy, particularly when the main class of interest is in the minority.

However, these measures are not appropriate for regression models that perform numeric prediction. According to Whitten and Frank (2000) and Compumine (2004), Mean Error, Mean Squared Error, and Correction Coefficient can be used to evaluate the performance of regression models.

2.2 CEMENT PRODUCTION

In this section, an attempt has been made to review cement production process, technologies involved, the raw materials used and their chemical compositions. The review made here is based on the work of Hewlett (1998), the technical report "Cement production and quality control" of workshop held in Jordan in 2001, and publication of the seminar on "production and quality control in cement industry" held Qatar in December 2003.

2.2.1 OVERVIEW OF CEMENT

Cement may be defined as any material that hardens and becomes strongly adhesive after application in plastic form. The term *cement* is often used interchangeably with glue and adhesive. However, in engineering and building construction the term usually refers to a finely powdered substance that hardens and adheres after being mixed with water (Tomasetti, 2004).

According to British Standard BS:12 (1996), cement is defined as a hydraulic binder; it is finely ground inorganic material which, when mixed with water, forms a paste which set and harden by means of hydration reactions and after hardening, retains its strength and volume stability even under water.

The term *Portland cement* was first used in 1824 by Joseph Aspdin, a British cement maker, because of the resemblance between concrete made from his cement and Portland stone, which was much used in building in England. Portland cement is manufactured from lime-bearing materials, usually limestone, together with clays or blast-furnace slag containing Alumina (Al_2O_3) and Silica (Si_2O_3), in the approximate proportions of 60% lime (CaO), 19% Silica, and 8% Alumina, 5% Iron(Fe_2O_3), 5% Magnesia (MgO), and 3% Sulfur trioxide (SO_3) (Hewlett, 1998).

2.2.2 CEMENT PRODUCTION PROCESS

In the manufacture of cement the raw materials are ground together, the mixture is heated until it fuses into a clinker, and the clinker is ground into a fine powder. The heating is usually accomplished in rotary kilns more than 150 m long and 3.7 m or more in diameter. The kilns are slightly tilted from the horizontal, and the raw material is introduced at the upper end, either in the form of a dry rock powder or as a wet paste composed of ground-up rock and water. As the charge progresses down through the kiln, it is dried and heated by the hot gases from a flame at the lower end. As it comes nearer the flame, Carbon dioxide is driven off, and in

the area of the flame itself the charge is fused at temperatures between 1350 and 1450°C. The material takes approximately 6 hours to pass from one end of the kiln to the other. After it leaves the kiln, the clinker is cooled quickly and ground, and then conveyed by a blower to storage silos (Tomasetti, 2004).

Cement kilns are massive cylindrical structures, lined with refractory brick, into which a lime-bearing material and fuel such as coal or gas is fed. There is a range of different kiln designs but all rely on the same basic processes-raw feed passing through the kiln is heated to very high temperatures by the burning of fuel and is transformed chemically and physically into a grey material called clinker. Clinker is ground to produce cement.

According to Hewlett (1998), the three broad processes of cement manufacturing are:

- Raw milling (the preparation of the kiln feed);
- Calcining and burning or clinkering (the conversion process that takes place within the cement kiln and associated equipment); and
- Finish milling (the grinding of clinker to produce cement).

In addition to the processes outlined, other processes included in cement production are mining (quarrying), primary and secondary crushing and transport of raw materials, pre-blending, storage, loading and discharge of clinker and cement powder, and bagging of cement.

2.2.2.1 RAW MATERIALS

Minerals of natural origin as well as industrial products can be used for production of cement. The main components available for the manufacture of cement are limestone and clay.

Limestone is a sedimentary rock composed of mainly the mineral Calcium Carbonate (CaCO_3) and serves as a source of Lime (CaO). Limestone also contains Magnesium, Aluminum and Iron combined as Carbonates and Silicates. Clay is soil formed by weathering of sedimentary rocks. It is the source of SiO_2 , Al_2O_3 and Fe_2O_3 . If the amount of SiO_2 is less in clay, sandstone may be used as a corrective additive to supply SiO_2 . Another important raw material is Gypsum ($\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$): non-metallic natural mineral used in cement manufacturing to delay the rapid setting of cement. About 5% gypsum is added when grinding clinker to produce cement.

The four oxides CaO , SiO_2 , Al_2O_3 , and Fe_2O_3 are the principal oxides in cement manufacturing. The sum of these four oxides is just around 96%. The remaining four percent is composed of raw material contaminants or impurities; that is, some minerals or elements are present in cement due to their original presence in raw materials as impurities. These elements, described below, have adverse effect on cement manufacturing process and product quality.

- **Alkalies Na_2O and K_2O (Sodium oxide and Potassium oxide)**

Alkalies Na_2O and K_2O exist in cement raw materials, especially in clay. Excess Alkalies affect kiln operation (system clogging) and product quality (alkali aggregate reactivity). Certain concrete aggregates contain alkali sensitive components, which under unfavorable conditions can react in an expansion phenomena- alkali expansion. The Alkalies in cement also have the potential to react with and corrode the steel used to reinforce the concrete. The amount of Alkalies content in cement is specified by standard specification.

- **Chloride (Cl)**

Chloride is one of the undesired elements, which exist in the raw materials of cement and have a negative effect in cement manufacturing. Chloride reacts with alkalies in kiln forming alkali chlorides that cause potential system clogging. So, the maximum permissible value of Alkalies content in cement is specified by standard specification.

- **Magnesium Oxide (MgO)**

Magnesium Oxide is one of the auxiliary components of cement raw materials whose quantity in cement is limited by standard specification. This component is combined up to 2% by weight with clinker phases; beyond that amount it appears in the clinker as free Magnesia (free MgO or Periclase). During cement hydration, Periclase or free MgO reacts with water to give $\text{Mg}(\text{OH})_2$; that is, $\text{MgO} + \text{H}_2\text{O} \rightarrow \text{Mg}(\text{OH})_2$.

But this reaction proceeds slowly while the other hardening reactions are already concluded. Since the $\text{Mg}(\text{OH})_2$ occupies a larger volume than MgO , it can split apart the binding of the hardened cement paste resulting in expansion cracks in concrete or mortar (Magnesia expansion). The proportion of Magnesium Oxide in cement is limited to 5% in most standard specifications.

2.2.2.2 PROPORTIONING, MIXING AND RAW MILLING

Proportioning involves determining the correct proportions of raw materials to be combined to form the raw mix. The proportions are governed by the chemical properties of the materials to be used and the desired

properties of the clinker that is produced. After proportioning, the raw materials are mixed to produce 'raw-meal'. Mixing is an important step in the process to ensure an even distribution of the properly proportioned components of the raw mix so that the clinker will be of a uniform quality.

The limestone and clay components have to be mixed each other in proportions depending on their own and the required final composition. For the production of cement it is necessary to have or make raw material mixture whose chemicals composition is within certain limits. The continuous production of high quality cement is possible only if the raw mix possesses optimum composition and further more if variation in this composition remains within the narrowest possible range (Rao, 2003).

For practical purposes, the raw material composition is usually characterized by certain ratios, often called "moduli". These ratios are in fact proportioning formulas into which the percentage of the various oxides as determined by chemical test should be. According to Rao (2003), the principal ratios used as chemical parameters in cement production are:

- **Silica modulus (SM)**

$$SM = [\%SiO_2] \div [\%Al_2O_3] + [\%Fe_2O_3]$$

The amount of melt phase in the kiln is a function of SM. When the SM is high, the amount of melt is low and vice versa. When the SM is too high, chemical reaction may be too slow and amount of dust increases that impede the kiln operation. Typical value for SM is 2.3-3.5.

- **Iron modulus (IM)**

$$IM = [\%Al_2O_3 \div \%Fe_2O_3]$$

IM determines the viscosity of the melt phase. The higher the IM the higher the viscosity and vice versa. The temperature by which the melt phase forms in kiln depends on the IM. The lowest temperature is obtained when $IM \approx 1.6$. This is optimum regarding the formation of clinker minerals. Typical value of IM is 1.5-2.5.

- **Lime saturation factor (LSF)**

$$LSF = 100 \times (\%CaO) \div [(2.8 \times \%SiO_2) + (1.2 \times \%Al_2O_3) + (0.65 \times \%Fe_2O_3)]$$

The LSF establishes the required proportions between the four oxides to make clinker. The word saturation factor refers to the maximum amount of lime (CaO) that the raw mix tolerates based on its silica, Alumina and Iron oxide content. In controlling the burning process, the amount of un-reacted CaO (free lime) is monitored. The lower the free lime, the more complete the reaction. LSF is used to measure the free CaO. Typical value for LSF is 92-96%.

As mentioned previously, the four oxides CaO , SiO_2 , Al_2O_3 , and Fe_2O_3 are the principal oxides in cement manufacturing and combined form clinker according to the following equation:



From the above equation, it is clear to see that lime (CaO) is by far the most important clinker components. Lime deficiency leads to poor cement strength, while excess lime (CaO) leads to the presence of free lime which causes unsoundness or expansion problem. Similarly, a deficiency in silica (SiO_2) leads to low early strength, while excessive silica leads to unburnability problems and kiln upset. Iron (Fe_2O_3) has the desirable property of acting as a fluxing agent in the kiln, facilitating the formation of other components of the cement at somewhat lower temperature. Therefore, raw mix proportioning is a serious task (Rao, 2003).

Optimum chemical composition is an important factor in determining the cement quality. However, Hewlett (1998) claims that overall chemical composition is not the only deciding factor to produce good quality cement and suggests a number of other significant factors that must be considered to make good quality clinker and hence cement, such as raw mix grinding, raw mix homogenization, burning temperature, burning time (duration of burning), and cooling rate of clinker.

2.2.2.3 BURNING OF THE RAW MIX IN A KILN TO FORM CLINKER

According to Tomasetti (2004), there are four stages to this process- evaporation and preheating, calcining, clinkering, and cooling. Evaporation and preheating remove moisture and raise the temperature of the raw mix preparatory to calcining. Calcining takes place at $800-900^\circ\text{C}$ and breaks the calcium carbonate (CaCO_3) down into Calcium Oxide (CaO) and Carbon dioxide (CO_2), which is evolved in the process. Clinkering completes the calcination stage and fuses the calcined raw mix into hard small grey pebbles called clinker. Kiln temperatures in the clinkering zone range from $1350-1450^\circ\text{C}$.

Clinkering is critical to the quality of cement and requires accurate control of the energy input (Asim, 2003). Insufficient heat will cause the clinker to be under-burnt, containing unconverted lime, and reducing the hydration (setting and hardening) properties of the resulting cement. Excess heat will shorten the life of the refractory bricks lining the kiln, may damage the kiln shell and diminish product reactivity.

According to Asim (2003), the quantitative characterization of the kinetics of clinker formation, however, is extremely complicated due to:

- Cement clinker having a range of chemical compositions and being produced from a wide spectrum of different minerals
- The transformation takes place via a series of sequential and simultaneous reaction steps of various types
- The non-availability of a simple precise method for determining the extent of reaction.

2.2.2.4 CEMENT CLINKER COMPOSITION

Cement clinker consists of substantially of the four crystalline clinker phases listed below

- **TRICALCIUM SILICALTE (C_3S):**

$$C_3S = [(4.071 \times \%CaO) - (7.6 \times \%SiO_2) - (6.718 \times \%Al_2O_3) - (1.430 \times \%Fe_2O_3) - (2.852 \times \%SO_3)]$$

 C_3S is the most important constituent and main contributor to both early and late strength of cement. Average proportion of C_3S in cement clinker is 60 % (Max 80 %, Min 40%).

- **DICALCIUM SILICATE (C_2S):**

$$C_2S = (2.867 \times \%SiO_2) - (0.7544 \times C_3S)$$

 C_2S is the second most important constituent that mainly contributes to late strength of cement. Average proportion of C_2S in cement clinker is 15% (Max 30%).

- **TRICALCIUM ALUMINATE (C_3A):**

$$C_3A = (2.650 \times \%Al_2O_3) - (1.692 \times \%Fe_2O_3)$$

 C_3A is the third important constituent that contributes to early strength. The reaction of C_3A with water is violent and leads to the very rapid setting and immediate stiffening of the cement paste. As a result, C_3A is responsible for the workability of the mortar or concrete. To regulate the rapid setting, gypsum ($CaSO_4 \cdot 2H_2O$) is added to cement. Average proportion of C_3A in cement clinker is 11% (Max 15%, Min 7%).

- **TETRACALCIUM ALUMINO FERRITE (C_4AF):**

$$C_4AF = (3.043 \times \%Fe_2O_3)$$

 C_4AF is forth constituent that governs the color of cement. Typical cement color is gray, but may also be dark gray-green depending on the content of MgO . Average proportion of C_4AF in cement clinker is 8% (Max 15%, Min 4%).

2.2.2.5 FINISH MILLING OF CLINKER

Finish milling is the grinding of clinker to produce a fine grey powder. Gypsum ($\text{CaSO}_4 \cdot 2\text{H}_2\text{O}$) is blended with the ground clinker, along with other materials, to produce finished cement. The gypsum controls the rate of hydration of the cement in the cement-setting process. The amount of gypsum added is connected to the C_3A content and fineness of cement. Fineness of cement increases the gypsum needed in cement. However, too much gypsum (sulfate) in cement is liable to make expansion. So, the maximum permissible value of sulfate expressed as SO_3 content in cement is specified by standard specification (Hewlett, 1998).

2.2.2.6 SPECIAL CEMENTS

By varying the percentage of its normal components or adding others, Portland cement can be given various desirable characteristics, such as rapid-hardening and resistance to sulfates. Rapid-hardening cements, sometimes called high-early-strength cements, are made by increasing the proportion of Tricalcium silicate (C_3S) or by finer grinding of clinker. For concrete work exposed to sulfate condition that which attacks it, Sulfate Resistant Cements with low Aluminum (Al_2O_3) or low Tricalcium Aluminate (C_3A) content are generally employed (Tomasetti, 2004).

2.2.2.7 SUPPLEMENTARY CEMENTITIOUS MATERIALS

Supplementary cementitious materials (or cement extenders) are materials that exhibit inherent or latent cementitious properties. They are added to cement through inter-grinding with cement clinker. They extend the cementing or binding capacity of a given quantity of cement because the material has latent cementitious properties (Khan, 2003).

Latent cementitious properties are activated by the presence of Portland cement in a moist environment. They become part of the binding component of the cement paste which in combination with other materials hardens to form concrete. The addition of the cement extender may allow a reduction in the quantity of cement. According to Khan (2003), the main types of supplementary cementitious materials are Pozzolanic compounds such as *fly ash*, *blast-furnace slag*, *silica fume* and *natural Pozzolana*. Pumice is a typical example of natural Pozzolana.

2.3 CEMENT TESTING AND QUALITY ASSESSMENT

In this section an attempt has been made to review typical tests conducted and methods used to assess quality of cement. The review made here, similar to the previous section 2.2, is based on the works of Hewlett (1998), the technical report "*Cement production and quality control*" of workshop held in Jordan in 2001, and publication of the seminar on "*production and quality control in cement industry*" held Qatar in December 2003. In addition, cement testing specifications of the British Standard (BS:12), European Pre-standard (ENV-197), and American Standard for Testing and Materials (ASTM) have been utilized.

2.3.1 OVERVIEW OF QUALITY

Quality is a term that is widely used by customers searching for good products and by companies trying to convince the customers with their products. According to ISO 9000 (2000), quality is defined as "Totality of characteristics of an entity that bear on its ability to satisfy stated or implied needs."

Quality of manufactured products is always subject to a certain amount of variation. No production process is good enough to produce all items of product exactly alike. Some stable variation is inherent in any particular scheme of production and inspection of variation within this stable pattern is inevitable (Nadeem, 2003).

To provide customers with good quality products, there is a need for a suitable quality assurance system. ISO 9000 (2000) define quality assurance as "All the planned and systematic activities implemented within the quality system, and demonstrated as needed, to provide adequate confidence that an entity will fulfill requirements for quality" and remarked that for specific product or service, quality assurance involves verifications, audits, and the evaluation of the quality factors that affect the specification, production, inspection, and use of product.

In cement industry, the cement consumer expects a product with the most favorable characteristics for application - a high level of strength and other desirable cement properties. The chemical and physical properties of cement are of prime importance to users and the production of cement complying with national and international quality standards is the main objective of the producers (Dr. Khatib, 2003).

The overall quality of cement is governed by chemical and physical variables. These variables are tested and monitored in order to assess the quality of cement produced daily. The typical chemical and physical tests that are employed for such purpose are discussed below.

2.3.2 CHEMICAL TESTS

The chemical tests are performed on chemical variables of cement quality. Typical chemical tests specified in most popular cement standards and utilized to assess the quality of manufactured cement, by determining the percentage of chemical composition of cement, are depicted in table 1 below.

Chemical variable	Description
SiO ₂	Silicon dioxide
Al ₂ O ₃	Aluminum Oxide
Fe ₂ O ₃	Ferric Oxide
CaO	Calcium Oxide
MgO	Magnesium Oxide
SO ₃	Sulfur trioxide
Cl ⁻	Chloride
L_O_I	loss on ignition
I_R	Insoluble residue
Na ₂ O	Sodium oxide
K ₂ O	Potassium oxide
% Pozzolana	Percentage of Pozzolana in cement
Free CaO	Uncombined Calcium Oxide

Table 1: Chemical test variables

However, depending on the cement type, the particular chemical test may include all or some of the variables. The chemical tests of SO₃, Al₂O₃, Na₂O, SiO₂, CaO, K₂O, Fe₂O₃, MgO, and Cl⁻ assess the amount (percentage) of the respective minerals in cement. The L_O_I test assesses the amount of volatile organic matter in cement at about 1000°C. The Free CaO test assesses the amount of free or uncombined lime. The I_R test assesses the amount of insoluble material in cement. The total sum of the results of these tests is less than or equal to 100 and the value 100 minus the total sum is considered as elements not determined by the chemical analysis.

All the above chemical variables, except %Pozzolana, are usually tested for OPC. The chemical variables that are usually tested for PPC include SO₃, MgO, L_O_I, %Pozzolana and Cl⁻. The %Pozzolana test is a special test used for PPC and assesses the amount of Pozzolana material (for example, Pumice) added in cement.

The chemical analysis of the components of cement can be performed by reacting cement with certain chemicals and following standard procedure. Two categories of such tests include *Gravimetric* and *Complexometric* methods. The analysis of elements in cement can also be carried out using modern technologies, such as *X-ray* and *Atomic Absorption Spectrometry* (Natsheh, 2001).

2.3.3 PHYSICAL TESTS

The physical tests are performed on physical variables of cement quality. Typical physical tests specified in most popular cement standards and utilized to assess the quality of manufactured cement can be grouped into four categories, are shown in table 2 below.

Category	Physical variable	Description	Typical Unit
Compressive strength	3d	3 days compressive strength	N/mm ²
	7d	7 days compressive strength	N/mm ²
	28d	28 days compressive strength	N/mm ²
Fineness	S_R	Sieve Residue	percentage
	S_S	Specific Surface	cm ² /gm
Expansion	S	Soundness	Millimeter (mm)
Setting time	I_S	Initial set	Minutes
	F_S	Final Set	Minutes

Table 2: Physical test variables

The purpose of assessing physical variables of cement listed in the above table as well as the methods and technologies available for conducting the tests are discussed below.

COMPRESSIVE STRENGTH

The most important property of cement is the strength development that is usually tested in the first 3, 7 and 28 days (Hewlett, 1998). According to British Standard (BS:12), European Pre-standard (ENV-197), and American Standard for Testing and Materials (ASTM) cement standards, this strength development property of cement is dependent on a number of factors, including % C₃S (the higher the C₃S, the higher the early and later strength), % C₂S (the higher the C₂S, the higher the late strength), % C₃A and % C₄AF (these contribute to strength but less than C₃S and C₂S), mixing conditions and environment, water to cement ratio, fineness of cement, particle size distribution, % SO₃ (the amount of gypsum added), curing conditions, additives (the amount, type and quality of additives added to cement), and overall chemical composition of cement.

Cement strength test is carried out using a compressive strength testing machine that may be mechanical or fully automatic with computer interface. However, in both cases, the testing process involves applying a compressive force on mortar cubes of cement and recording the point at which the mortar cube fail to resist the force as its compressive strength (usually in N/mm^2).

To carry out the test, first a mortar specimen of one part cement and three part sand is prepared. The prepared mortar is placed in cubic mortar mould that is made of metal. The mortar in the mould is placed in curing cabinet for 24 hours. For every batch of cement produced daily, nine such mortar cubes are prepared. After 24 hours, the metal moulds are removed and the harden mortar cubes are placed in water.

After 3 days, compressive strength test is done on the first three mortar cubes and the average result is recorded as a 3 days compressive strength. Similar testing and recording processes are carried out on each of the three remaining mortar cubes after 7 days and 28 days. All these 3, 7 and 28 days strength tests will be the respective 3, 7 and 28 days compressive strength results of cement that was produced 28 days ago.

FINENESS

Fineness of cement is a measure that assesses how fine the cement has been ground to. Fineness is a decisive factor for the strength of cement. When the fineness of cement increases, then the amount of cement particles that potentially react with water increases, thereby increasing the strength of cement.

However, particle size distribution of cement is more important than fineness for providing strength of cement. If the particle size distributions is the same then any increase in the fineness will result in strength increase. But if the particle size distribution is not the same, fineness alone can not be suitable as a parameter for strength estimation.

There are two popular techniques to determine the fineness of cement. Both these methods can be carried out using devices operated manually or automatically. The first technique involves the use of sieve with predefined diameter between the openings of the sieve, usually 25, 45 or 90 micrometer (μm). With this method fixed amount of cement (usually 1gm) is vibrated so that cement particles with diameter less than the diameter of sieve opening pass through the openings and the percentage of cement that fail to pass is calculated as percentage of sieve residue. Testing sieve residue of cement using sieves is convenient to monitor the production process. However, it does not measure or determine the particle size distribution.

The specific surface measurement is second technique used to test and determine the fineness as well as the particle distribution of cement. This test is carried out using a device called Blain Air Permeability Apparatus. Using this apparatus, the fineness of cement is measured as specific surface by observing the time taken for a fixed quantity of air to flow through a compacted cement bed of specified dimension and porosity.

Under standardized conditions, the specific surface of cement is proportional to $\sqrt{t}$, where t is the time for given quantity of air to flow through the compacted bed of cement. The number and size range of individual pores in the specified bed are determined by the cement particle size distribution which also determines the time for the specified air flow. The specific surface is reported in unit of cm^2/gm or m^2/kg .

EXPANSION

Expansion or Soundness of cement means the chemical requirements to avoid the excessive expansion of cement paste during hydration. In cement certain amount of free CaO and uncombined MgO exist. When these hydrate to Ca(OH)_2 and Mg(OH)_2 respectively, the cement paste should not exceed certain limits of expansion; otherwise cracks might occur in the concrete. Soundness, therefore, is a test performed to identify such possible expansion. The two popular methods to conduct soundness test are the Le-Chatelier test and Autoclave test. Both these test methods involve applying heat to cement paste and measuring the distance (usually in millimeter) between the indicator points as the amount of expansion.

SETTING TIME

Setting time is the duration of time elapsing before the cement paste change from plastic mass to rigid mass. Setting time governs the workability of concrete mix or mortar. Factors affecting setting time include water to cement ratio, temperature, C_3A content, gypsum content (SO_3 content), fineness of cement and mixing condition. The device mostly used to test the setting time of cement is called *VICAT needle*, which depends on the needle penetration in the cement paste. The penetration operation of the needle can be controlled manually or automatically.

The initial setting time, measured in minutes or hours, is the time elapsed until the needle can not penetrate the paste totally. The final set, measured similarly, is the time when the needle can not penetrate any more in the solid paste. To measure the initial and final setting time, two types of needles that differ in diameter are used.

2.4 CEMENT STANDARDS

In this section an attempt has been made to review the standard physical and chemical requirement specifications of three most popular and widely adopted standards in cement manufacturing: the British Standard (BS:12), European Pre-standard (ENV-197), and American standard for Testing and Materials (ASTM C-150, C-595).

2.4.1 OVERVIEW

The requirements on composition and performance of cement are outlined in the cement standards. The cement standards govern, to a significant degree, the quality of cement and thus have a direct effect on market position. Further more, the standards are recognized terms of reference in relations with customers. The standards are published by a society or by the government (Samir, 2001).

In general, the standards comprise test methods, definitions, recommended practices, classifications and specifications. A test method covers sampling and describes the subsequent testing procedures used to determine the composition or performance of the material to be specified. A specification is a precise statement of a set of requirements to be satisfied by a material, product system, or service (Hani, 2001).

According to Samir (2001), the most popular and widely adopted standards in cement manufacturing are the British standard (BS), the American Society for Testing and Materials (ASTM) standard and the European Pre standard (ENV-197). These standards are the backbone of most other national standards. Most Latin American countries use the ASTM or modified ASTM.

The major chemical and physical requirements specifications of these standards are discussed below.

2.4.2 THE BRITISH STANDARD (BS:12)

According to the British Standard (BS:12), cement, when mixed with aggregate and water to produce mortar or concrete, should retain workability for specified time and shall after definite period attain specified strength level and also possess long term of volume stability.

The chemical requirements specification of the BS:12 standard for Ordinary Portland Cement (OPC) is shown in table 3 below:

Property	Requirement
Loss on ignition (L.O.I)	$\leq 3.0\%$
Insoluble residue (I.R)	$\leq 1.5\%$
Sulfate expressed as SO_3	$\leq 3.5\%$
Chloride (Cl^-)	$\leq 0.01\%$
CaO and SiO_2	$\geq 50\%$
C_3S and C_2S	$\geq 66\%$
CaO/ SiO_2 ratio	$\geq 2.0\%$
MgO	$\leq 5.0\%$
Additives	$\leq 1.0\%$

Table 3: Chemical requirement of BS:12 Standard for OPC

The BS specification BS:12 for Ordinary Portland Cement (OPC) identify three classes of cement according to the compressive strength at 28 days as 32.5N, 42.5N and 52.5N. Additionally, two sub classes for early strength are defined as 32.5R and 42.5R. The **N** and **R** classes are for ordinary early strength and high early strength classes. Based on the strength classes, the strength requirements of the BS:12 standard for OPC is shown in table 4 below:

Strength class	Early strength		Standard strength	
	2 days (N/mm^2)	7 days (N/mm^2)	28 days (N/mm^2)	
32.5N	-	≥ 16	≥ 32.5	≤ 52.5
32.5R	≥ 10	-	≥ 32.5	≤ 52.5
42.5N	≥ 10	-	≥ 42.5	≤ 62.5
42.5R	≥ 20	-	≥ 42.5	≤ 62.5
52.5N	≥ 20	-	≥ 52.5	-
62.5N	≥ 20	-	≥ 62.5	-

Table 4: Strength requirements of BS:12 Standard for OPC

The setting time and expansion/soundness physical requirements of BS:12 standard for OPC are shown in table 5 below:

Property	Requirement
Initial setting time	- Minimum of 60 minutes - Minimum of 45 minutes for class 52.5N and 62.5N
Soundness	Maximum of 10 mm

Table 5: Setting time and expansion specification of BS: 12 for OPC

2.4.3 THE EUROPEAN PRE STANDARD (ENV-197)

The purpose of this prestandard is to specify the composition, requirements and conformity criteria for common cements. Types based on composition and classifications based on strength have been introduced in order to take into account the different cements included.

The European prestandard specifies properties of the constituents of common cements and the proportions in which they are to be combined to produce a range of types and classes of cement. It then specifies the physical and chemical requirements for these types and classes as characteristic values and states the rules for assessing their conformity to these requirements.

Common cements conforming to the European Pre-Standard ENV197-1 are subdivided into five main types as shown in table 6 below:

Cement Type	Name
CEM I	Portland cement
CEM II	Portland-composite cement
CEM III	Blast-furnace cement
CEM IV	Pozzolan cement
CEM V	Composite cement

Table 6: Common cement types of the European Pre-Standard ENV 197-1

The notation and designation of main constituents of cements, other than Portland cement clinker, identified by the European Pre-Standard ENV 197-1 standard is given in table 7 below:

Notation	Designation
S	Granulated blast furnace slag
D	Silica fume
P	Pozzolana, natural
Q	Pozzolana, industrial
V	Fly ash, siliceous
W	Fly ash, calcareous
T	Burnt shale
L	Limestone
M	Composite, two of the above

Table 7: The ENV 197-1 notation and designation of constituents of cement

Common cement types and their composition (proportion by mass) defined in the ENV 197-1 is given in table 8 below:

Cement	Designation	Notation	Clinker	S	D	P	Q	V	W	T	L
CEM I	Portland cement	I	95-100	-	-	-	-	-	-	-	-
	Portland slag cement	II/A-S	80-94	6-20	-	-	-	-	-	-	-
		II/B-S	65-79	21-35	-	-	-	-	-	-	-
	Portland silica fume cement	II/A-D	90-94	-	6-10	-	-	-	-	-	-
	Portland Pozzolanic cement	II/A-P	80-94	-	-	6-20	-	-	-	-	-
		II/B-P	65-79	-	-	21-35	-	-	-	-	-
		II/A-Q	80-94	-	-	-	6-20	-	-	-	-
		II/B-Q	65-79	-	-	-	21-35	-	-	-	-
	Portland fly ash cement	II/A-V	80-94	-	-	-	-	6-20	-	-	-
		II/A-V	65-79	-	-	-	-	21-35	-	-	-
		II/A-W	80-94	-	-	-	-	-	6-20	-	-
		II/B-W	65-79	-	-	-	-	-	21-35	-	-
CEM II	Portland burnt shale cement	II/A-T	80-94	-	-	-	-	-	-	6-20	-
		II/B-T	65-79	-	-	-	-	-	-	21-35	-
	Portland lime stone cement	II/A-L	80-94	-	-	-	-	-	-	-	6-20
		II/B-L	65-79	-	-	-	-	-	-	-	21-35
	Portland composite cement	II/A-M	80-94	6-20							
		II/B-M	65-79	21-35							
	Blast furnace cement	III/A	35-64	36-65	-	-	-	-	-	-	-
		III/B	20-34	66-80	-	-	-	-	-	-	-
		III/C	5-19	81-95	-	-	-	-	-	-	-
CEM IV	Pozzolanic cement	IV/A	65-89	-	11-35				-	-	-
		IV/B	45-64	-	36-55				-	-	-
CEM V	Composite cement	V/A	40-64	18-30	-	18-30			-	-	-
		V/B	20-39	31-50	-	31-50			-	-	-

Table 8: Common cement types and composition defined in the ENV 197-1 standard

As seen in table above there are five main cement types. Each of these is classified into three strength classes 32.5, 42.5 and 52.5. In the European pre-standard, the standard strength of cement is the compressive strength determined at 28 days. Two classes of strength are further defined for each of standard strength, a class with ordinary strength and a class with high early strength indicated by *R*.

The physical requirements of cements defined by the ENV 197-1 are given in table 9 below:

Compressive strength (N/mm2)					Initial setting time (Min)	Expansion (mm)
Class	Early strength		Standard strength			
	2 days	7 days	28 days			
32.5	-	≥ 16	≥ 32.5	≤ 52.5	≥ 60	≤ 10
32.5 R	≥ 10	-				
42.5	≥ 10	-	≥ 42.5	≤ 62.5		
42.5 R	≥ 20	-				
52.5	≥ 20	-	≥ 52.5	-	≥ 45	
52.5 R	≥ 20	-				

Table 9: Physical requirements of the ENV 197-1 Standard

The chemical requirements of cements, according to the ENV 197-1, are given in table 10 below:

Property	Cement type	Strength class	Requirement
Loss on ignition	CEM I CEM III	All classes	≤ 5.0%
Insoluble residue	CEM I CEM III	All classes	≤ 5.0%
Sulfate (SO ₃)	CEM I CEM II CEM IV CEM V	32.5 32.5R 42.5	≤ 3.5%
		42.5 R 52.5 52.5 R	≤ 40%
		All classes	
	CEM III	All classes	≤ 0.10%
Chloride	All types	All classes	

Table 10: Chemical requirements of the ENV 197-1 Standard

2.4.5 THE AMERICAN STANDARD FOR TESTING AND MATERIALS (ASTM)

The ASTM standard has two standard specifications C-150 and C-595 for Portland cement and Blended hydraulic cements respectively. According to the ASTM C-150 standard specification, Portland cement (or Ordinary Portland Cement) is defined as hydraulic cement produced by pulverizing clinker consisting of essentially of hydraulic calcium silicates, usually containing one or more forms of calcium sulfate as an inter-ground addition. The ASTM C-595 standard is for blended hydraulic cements, for both general and special applications, using slag and Pozzolana, or both, with Portland cement or Portland cement clinker.

For each of the eight types Portland cement in the table 11, ASTM C-150 states the respective standard physical requirements prescribed in table 13 below:

Property	Cement type							
	I	IA	II	IIA	III	IIIA	IV	V
Fineness, specific surface, air permeability test, m ² /kg	280	280	280	280	-	-	160	160
Autoclave expansion, max%	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
1 day Compressive strength, MPa	-	-	-	-	12.0	10.0	-	-
3 days Compressive strength, MPa	12.4	10.0	10.0	8.0	24.0	19.0	-	8.0
7 days Compressive strength, MPa	19.3	16.0	17.0	14.0	-	-	7.0	15.0
28 days Compressive strength, MPa	24.1	22.0	28.0	22.0	-	-	17.0	21.0
Time of setting, Vicat test, minutes								
• Initial set	≥45	≥45	≥45	≥45	≥45	≥45	≥45	≥45
• Final set	≤375	≤375	≤375	≤375	≤375	≤375	≤375	≤375

Table 13: Physical specifications of ASTM C-150 for Ordinary Portland cements

The ASTM C-595 specification pertains to five classes of blended hydraulic cements for both general and special applications, using slag and Pozzolana, or both, with Portland cement or Portland cement clinker.

The types of blended cement covered by this specification are listed below:

- ◆ **Portland Blast-furnace slag cement** - One type with three optional provisions as follows
 - **Type IS** – Portland blast furnace slag cement for use in general concrete construction
 - Moderate sulfate resistance, air entrainment, or moderate heat of hydration or any combination may be specified using the suffixes (MS), (A), or (MH)
- ◆ **Portland-Pozzolana cement** – Two types, each with three optional provisions as follows
 - **Type IP** – Portland-Pozzolana cement for use in general concrete construction
 - Moderate sulfate resistance, air entrainment, or moderate heat of hydration or any combination may be specified using the suffixes (MS), (A), or (MH)
 - **Type P** – Portland- Pozzolana cement for use in concrete construction where high strength at early ages are not required

- Moderate sulfate resistance, air entrainment, or low heat of hydration or any combination may be specified using the suffixes (MS), (A), or (LH)
- ◆ **Slag cement**- One type as follows
 - **Type S** – Slag cement for use in combination with Portland cement in making concrete
 - Air entraining can be specified by adding the suffix (A)
- ◆ **Pozzolana-Modified Portland cement** – One type is as follows
 - **Type I(PM)** – Pozzolana-modified Portland cement for use in general concrete construction
 - Moderate sulfate resistance, air entrainment, or moderate heat of hydration or any combination may be specified using the suffixes (MS), (A), or (MH)
- ◆ **Slag-modified Portland cement** – One type as follows
 - **Type I(SM)** – Slag-modified Portland cement for use in general concrete construction
 - Moderate sulfate resistance, air entrainment, or moderate heat of hydration or any combination may be specified using the suffixes (MS), (A), or (MH)

For blended cement types above, ASTM C-595 states the respective standard chemical requirements prescribed in table 14 below:

Property	Cement types		
	I(SM), I(SM)-A, IS, IS-A	S, SA	I(PM), I(PM)-A, P, PA, IP, IP-A
MgO, Max %	-	-	5.0
SO ₃ , Max %	3.0	4.0	4.0
Insoluble residue, Max %	1.0	1.0	-
Loss on ignition, Max %	3.0	4.0	5.0
Alkali, Max %	-	0.03	-

Table 14: Chemical specifications of ASTM C-595 for Blended cements

For blended cement types, ASTM C-595 states the respective standard physical requirements prescribed in table 15 below:

Property	Cement type							
	I(SM), IP I(PM), IS	I(SM)-A, IS-A, I(PM)-A, IP-A,	IS(MS), IP(MS)	IS-A(MS), IP-A(MS)	S	SA	P	PA
Autoclave expansion, max%	0.80	0.80	0.80	0.80	0.80	0.80	0.80	0.80
1 day Compressive strength, MPa	-	-	-	-	12.	10.0	-	-
3 days Compressive strength, MPa	12.4	10.0	11.0	9.0	-	-	-	-
7 days Compressive strength, MPa	19.3	16.0	18.0	14.0	5.0	4.0	11.0	9.0
28 days Compressive strength, MPa	24.1	20.0	25.0	20.0	11.0	9.0	12.0	18.0
Time of setting, Vicat test, min								
• Initial set	≥ 45	≥ 45	≥ 45	≥ 45	≥ 45	≥ 45	≥ 45	≥ 45
• Final set	≤ 420	≤ 420	≤ 420	≤ 420	≤ 420	≤ 420	≤ 420	≤ 420

Table 15: Physical specifications of ASTM C-595 for Blended cements

CHAPTER THREE

A SURVEY OF CEMENT PRODUCTION AND QUALITY ASSESSMENT AT MCE

3.1 OVERVIEW

Understanding the data and business itself is the first and foremost prerequisite to data mining. Without this understanding, no algorithm, regardless of sophistication, is going to provide with a result in which one should have confidence. Without this background, one will not be able to identify the problems, prepare the data for mining, or correctly interpret the results (CRISP-DM, 2000).

In line with this, this survey intends to assess the existing cement production and quality assessment process at Muger Cement Enterprise (MCE). The purpose is to conduct an analysis of the current business process, to identify the important functions and activities involved, and to identify and assess the availability of data sources that can support to develop models that reliably predict cement quality with acceptable accuracy.

Accordingly, this chapter is devoted to description of the cement production and quality assessment process at MCE. It starts by giving a brief overview of MCE, followed by a detailed description of cement production, quality assessment processes, and the procedures involved in cement quality approval process. The findings of the survey are included at the end of the chapter.

In the process of conducting this survey, a visit to the Enterprise to observe cement production process and a series of interviews were made with quality control experts. Additionally, secondary sources of information, such as documents, monthly and annual reports, and publications of the Enterprise have also been utilized.

3.2 MUGHER CEMENT ENTERPRISE

Mugher cement enterprise (MCE) is a state owned Enterprise established with a purpose of producing and supplying cement. The Enterprise, established with the authorized capital of 334,716,000 Birr, was formed in 1999 through amalgamation of two formerly independent factories: Muger Cement Factory and Addis Ababa Cement Factory. The Council of Ministers declared the unification of the two factories to form the MCE with a Proclamation Act number 53/1999 published in "Negart Gazetta" 5th year, number 67. The MCE currently is run by board of directors appointed by State Owned Enterprises Agency (Public Relations Department, MCE, 2002).

In addition to the two cement factories- Muger Cement Factory and Addis Ababa Cement Plant- MCE also owns its own paper sack making plant. Over view of the two cement plants of the MCE and their cement production processes are discussed below.

3.2.1 MUGHER CEMENT FACTORY (MCF)

Muger Cement Factory (MCF), the mother plant of the Enterprise, has two production lines (Kilns). The first line started operation in 1984 and the second line in 1990. The factory was established with long term loan secured from the Development Bank of Ethiopia and the former East Germany Government plus a financial outlay from the treasury of the Ethiopian government- totally 436,909,975 Birr capital.

The MCF is located 105Kms west of Addis Ababa. The design annual production capacity of the factory is 630,000 tons of cement. The type of technology used by MCF is a dry process five-stage pre-heater kiln with rotary cooler. Initially the factory was intended to produce Ordinary Portland Cement (OPC). However, in the past years Portland Pozzolana Cement (PPC) has been the main type of cement produced by the factory. The raw materials used by the factory are limestone, clay, sand stone, gypsum and pumice. Limestone and gypsum are quarried and transported to the factory by conveyer belt of 12Kms long. Clay and sandstone are supplied from near by quarries by dump trucks direct to the storage halls within the factory. Pumice is also transported by trucks from Koka quarry site located 206Kms from the factory.

Based on the annual reports of the Enterprise, the production history of the MCF is shown in table 16 below

Year	Annual Cement Production (tons)
2000	700,230
2001	745,000
2002	800,025
2003	849,980
2004	901,229

Table 16: Annual cement production history of MCF

3.2.1.1 CEMENT PRODUCTION PROCESS AT MCF

The cement production at MCF undergoes a number of processes that are more or less similar to processes followed in typical cement factory; the significant difference being the raw materials utilized and the technology adopted.

The cement production process can be divided into two categories. The first process involves clinker production and the second process involves grinding the clinker to produce cement. The clinker production process can schematically be represented in the following format.

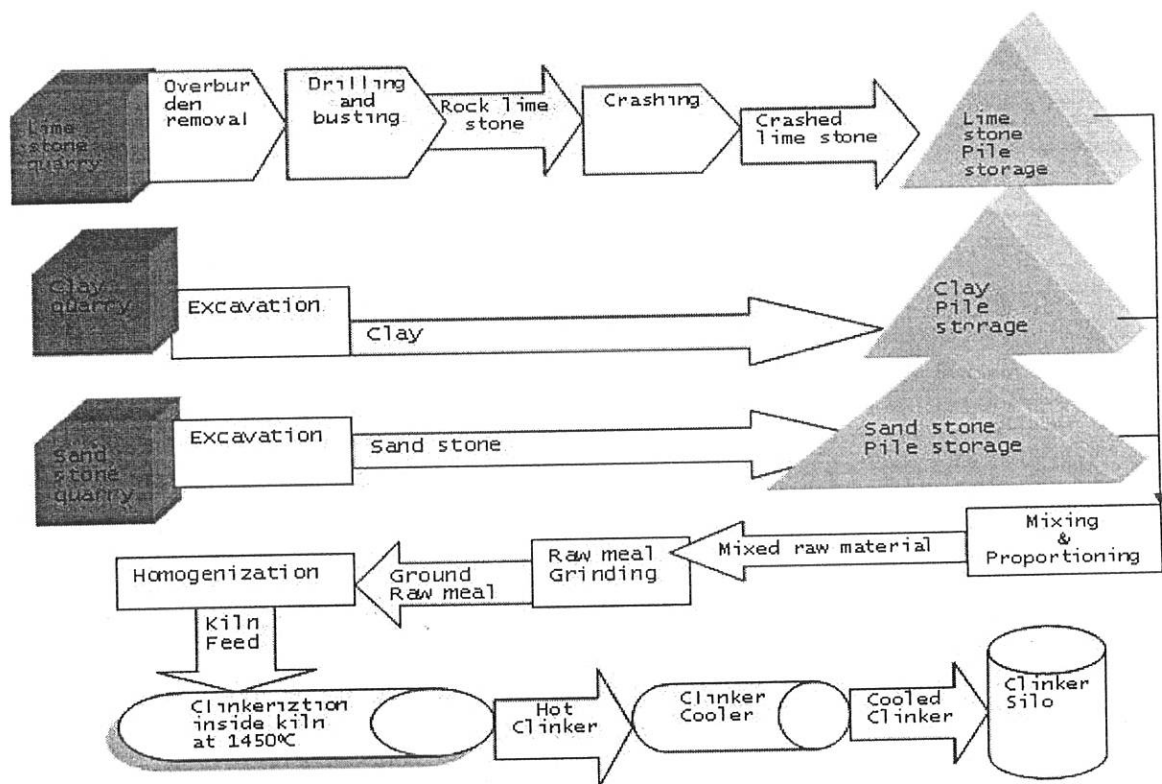


Figure 2: Clinker production process at MCF

The production of the two types of cements by the factory involves two processes that are more or less the same except the use of pumice raw material in the production of PPC. The OPC production process can schematically be represented in the following format.

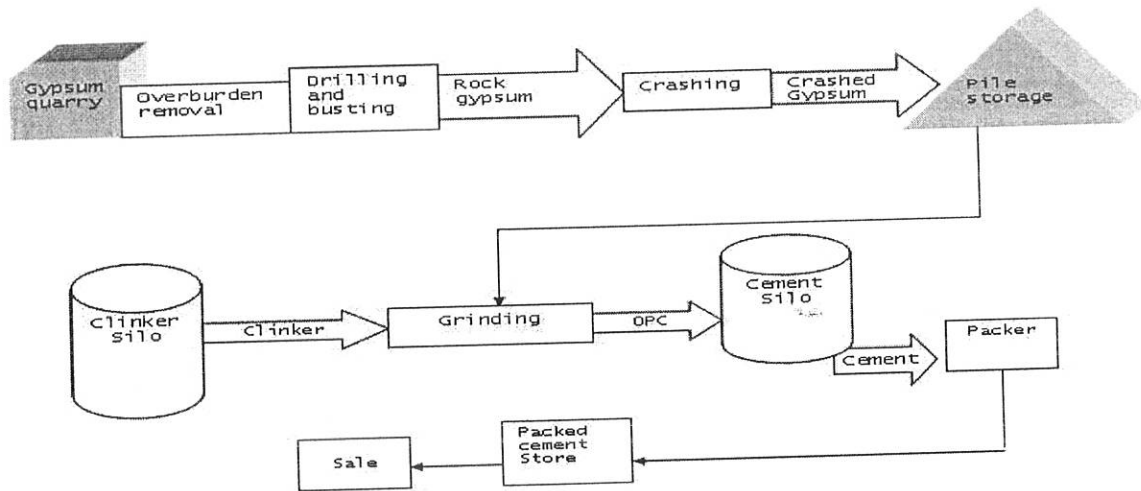


Figure 3: OPC production process at MCF

The PPC production process can schematically be represented in the following format.

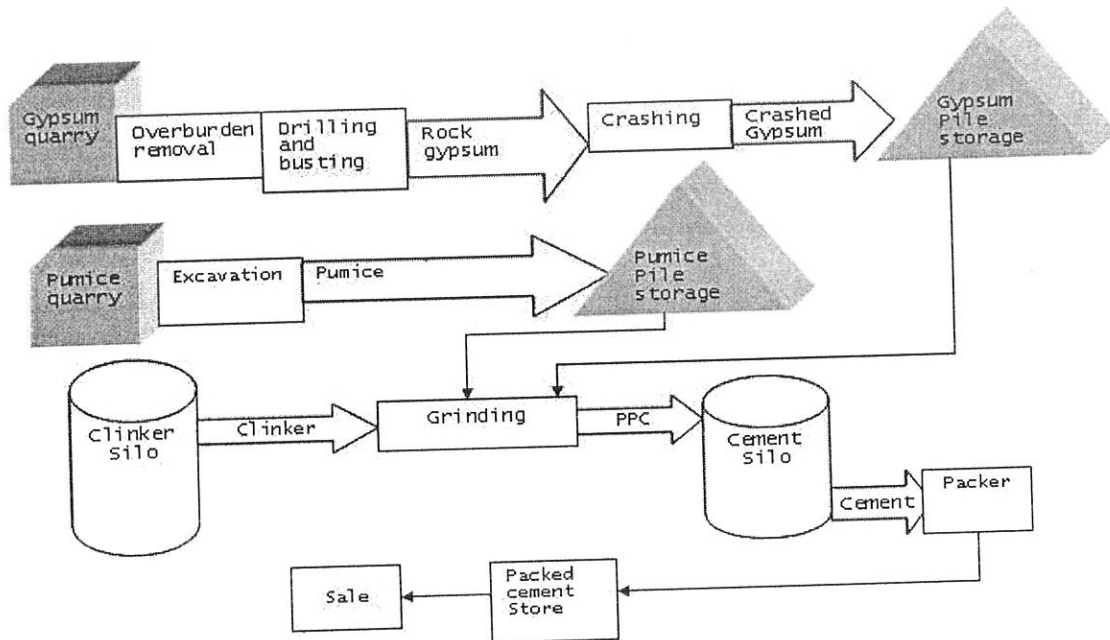


Figure 4: PPC production processes at MCF

The following steps are carried out in the production cement at Mughar Cement Factory:

- The overburden over the limestone is removed using dozers. Then, a hole is drilled using a very high air pressure. In the hole drilled, explosive is inserted and blasted turning the limestone into large pieces of rocks. The rock is transported to a near by hammer crusher that crushes the large limestone rocks into smaller ones. The hammer crusher has a capacity crush 350 to 380 tons of limestone per hour. The crushed lime stone is transported to the deposited hall located in the factory using a 12Kms long conveyer belt. At the limestone deposit hall, the transported limestone is mixed using a computer controlled machine called *stacker and reclaimer*. The annual limestone consumption of the factory is about 875,000 tons.
- Similar steps are carried to quarry and transport gypsum to the factory. However, the transported gypsum is stored in a separate deposit hall. The annual gypsum consumption of the factory is about 50,000 tons
- Clay raw material is excavated from a clay quarry located near the factory and transported to its own deposit hall inside the factory.
- Sandstone raw material is excavated from a sand stone quarry located near the factory and transported to its own deposit hall inside the factory. The factory uses sandstone as a corrective additive since the available clay that the factory uses is deficient in silica (SiO_2).
- Pumice raw material is excavated from a pumice quarry, located at Koka 206Kms from the factory, and transported to and stored its own deposit hall inside the factory. The annual pumice consumption of the factory is about 200,000 tons.
- The limestone, clay and sandstone raw materials are proportioned and fed in to Raw Mill. The factory has two Raw Mills that use steel balls as grinding mechanism and each with a capacity of grinding 90 tons per hour. Output of Raw Mills, called kiln-feed, is an input to the kiln operation.
- The kiln-feed is transported into homogenization silo, where it is further mixed with air to achieve a uniform distribution of the chemical components.
- The kiln-feed prepared this way is fed into the kiln. Inside the kiln, it is burnt at temperature of 1450°C using furnace oil to form what is known as clinker. The factory has two kilns, each with a capacity of producing 300,000 tons of clinker per year. The annual consumption of furnace oil by the factory is 63 million lts. The kilns and their operations are the heart and cost center of the factory.
- Each kiln is equipped with a rotary cooler that takes the clinker outputs. Inside the rotary cooler, the hot clinker is cooled using air fans.

- The cooled clinker output of the rotary coolers is transported to clinker silo using bucket elevators. The factory has four clinker silos, each capable of storing 1,000 tons of clinker.
- The factory produces two types of cements: Ordinary Portland Cement (OPC) and Portland Pozzolana Cement (PPC). When producing OPC, about 95% clinker and 5% gypsum are mixed and transported to Cement Mills using a weigh belt feeder. Inside the Cement Mills, the mix is ground into fine powder to produce OPC. Similar operations are used to produce PPC except that when producing PPC, about 5% gypsum, 20 to 30% pumice and 65 to 75% clinker are mixed and fed into Cement Mills as input. The factory has two Cement Mills that use steel balls as grinding mechanism and each with a capacity of grinding 90 tons per hour.
- The cement output of the Cement Mills is transported to Cement Silos using pneumatic (air) conveyers. The factory has four Cement Silos, each capable of storing 1,000 tons of cement.
- Using screw conveyers, the powder cement is transported from Cement Silos to cement packer, where it is packed into a 50Kg paper sacks labeled either PPC or OPC. The factory has two cement packers each capable of packing 50 tons of cement per hour.
- The packed cement is transported to Addis Ababa, where it is distributed to customers. For Customers who provide their own transport trucks, cement is also distributed at the factory premises by loading packed cement directly onto their trucks.

3.2.2 ADDIS ABABA CEMENT PLANT

The Addis Ababa Cement Plant (AACP), formerly known as Addis Ababa Cement Factory, was established in 1964. The factory is located in Addis Ababa and initially had a production capacity of 70, 000 tons of cement. However, following the failure of the factory's Electrostatic Precipitator, which prevents dust emission, the plant stopped its production of clinker in 1996. The plant, however, is reconditioned to produce PPC and OPC using semi-processed clinker transported from MCF.

In addition to transporting around 100, 000 tons of clinker from MCF, the plant has gypsum and limestone quarry located North West of Addis Ababa, at a distance of 70Kms. From this quarry site, mainly gypsum is transported to the plant using trucks. The plant also transports pumice from Koka pumice quarry site that is located east of the plant at distance of 60Kms.

Based on the annual reports of the enterprise, the production history of the AACP is shown in table 17 below:

Year	Annual Cement Production(tons)
2000	87,050
2001	95,440
2002	120,234
2003	125,030
2004	125,500

Table 17: Annual cement production history of AACP

3.2.2.1 CEMENT PRODUCTION PROCESS AT AACP

The cement production processes carried out at AACP are the subset of the processes performed at MCF. The AACP produces three types of cements: Ordinary Portland Cement (OPC), Portland Pozzolana Cement (PPC) and Sulfate Resisting Portland Pozzolana Cement (SRPPC). The OPC production process can schematically be represented in the following format.

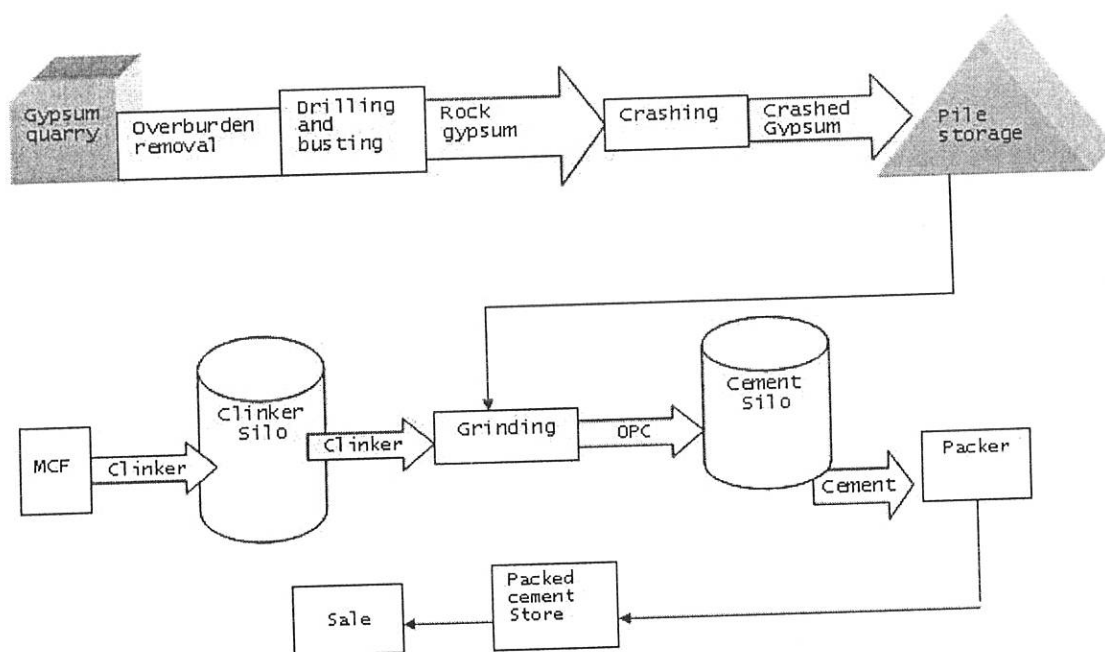


Figure 5: OPC production processes at AACP

The processes involved in the production of PPC and SRPPS are identical except for the percentage of pumice utilized. Thus, processes involved in the production these two types of cements are schematically represented the following format

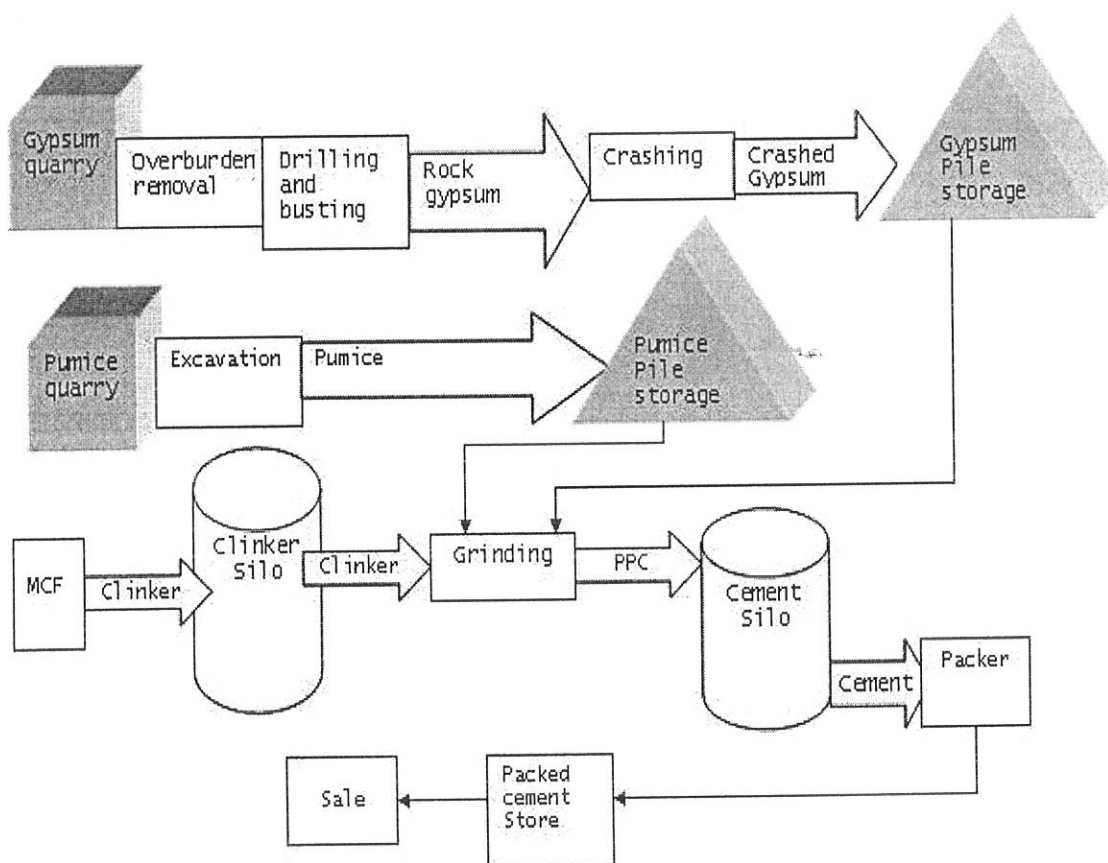


Figure 6: PPC and SRPPC production processes at AACP

The following steps are carried out in the production cement at Addis Ababa Cement Plant:

- Clinker is transported from Mughher cement factory to a deposit hall inside the factory. About 100,000 tons of clinker is transported per year.
- The factory's gypsum quarry is located north-west of Addis Ababa at a distance of 70Kms. Quarrying activities that are similar to MCF are carried out to gypsum from this quarry site, but this time the gypsum is transported to the plant using trucks. The annual gypsum consumption of the plant is about 5,000 tons.

- Pumice raw material is excavated from a pumice quarry located at Koka 60Kms west from the factory and then transported by trucks to a deposit hall inside the plant. The annual pumice consumption of the plant is about 30,000 tons.
- The AACP produces three types of cements: Ordinary Portland Cement (OPC), Portland Pozzolana Cement (PPC) and Sulfate Resisting Portland Pozzolana Cement (SRPPC). When producing OPC, about 95% clinker and 5% gypsum are mixed and transported to Cement Mills using a weigh belt feeder. Inside the Cement Mills, the mix is ground into fine powder to produce OPC. When producing PPC, about 5% gypsum, 20 to 30% pumice and 65 to 75% clinker are mixed and transported to Cement Mills using a weigh belt feeder. Inside the Cement Mills, the mix is ground into fine powder to produce PPC. Similar operations are used to produce SRPPC except that when producing SRPPC, about 5% gypsum, 40 to 45% pumice and 50 to 55% clinker are mixed and fed into Cement Mills as input. The factory has two Cement Mills that use steel balls as grinding mechanism and each with a capacity of grinding 12.5 tons per hour.
- The cement output of the Cement Mills is transported to cement silos using pneumatic (air) conveyers. The factory has six cement silos each capable of storing 100 tons of cement.
- Using screw conveyers, the powder cements are transported to cement packer, where it is packed into a 50Kg paper sacks labeled OPC, PPC or SRPPC. The factory has one cement packer capable of packing 50 tons of cement per hour.
- The packed cement is transported to deposit hall, where it is distributed to customers.

3.3 QUALITY ASSESSMENT PROCESS AT MCE

The Muger Cement Enterprise has two quality control laboratories for each of its cement plants. The quality control laboratory at the Muger Cement Factory is responsible for controlling quality of raw materials used by the factory and proportioning raw materials mix. It is also responsible for assessing and approving the quality of clinker and cement produced daily. Based on cement quality test results, the laboratory authorizes the sale of cement stored in silos.

The quality control laboratory at Addis Ababa Cement Plant also carries out similar activities, except monitoring clinker production process. The quality control laboratory is responsible for assessing and approving the quality of cement produced at AACP. Based on test results, it then authorizes the sale of cement in silos.

Even if the production capacities of AACP and MCF differ, the two quality control laboratories of the Enterprise have similar organization, functionality and follow same procedures for quality assessment tasks. This is true regardless of clinker production monitoring activity being performed only by the quality control laboratory of MCF. In fact, the quality control laboratory of AACP used to perform tests on clinker before its production is stopped for reason stated earlier.

The quality control laboratories of AACP and MCF have *process control*, *physical test* and *chemical analysis* sections that carry out similar tasks to assess and approve cement quality. The tasks carried out in these laboratory sections quality are discussed below.

PROCESS CONTROL SECTION

About 500 gram of sample of cement is collected from cement mills every 2 hours. From this sample, one gram is taken to carry out sieve residue test; one gram of cement is used to test percentage of pumice content, if PPC is the type of cement currently produced. The remaining sample is used for tests conducted in the physical and chemical analysis sections

The sieve residue and pumice content test results are recorded in log sheets. The average of sieve residue and pumice content tests conducted for 24 hours are recorded on *physical test of milled cement form* (attached as annex 2) as the sieve residue and pumice content results of cement produced daily.

PHYSICAL TEST SECTION

The cement samples that have been collected every two hours for 24 hours are mixed properly to represent the factory's daily production sample. This 24 hours long cement sample is used by the physical test and chemical analysis sections.

Using the sample that has been collected every two hours for 24 hours, the physical test section prepared nine mortar cubes for 3, 7, and 28 days compressive strength tests. When the test dates are due, the respective 3, 7, and 28 days compressive strength tests are performed on the three mortar cubes, using a computer controlled compressive strength testing machine. The average test result of the tests on conducted on the three mortar cubes is recorded and saved in a computer file.

Using the same sample, the physical test section also carry out setting time and expansion tests. The setting time test is conducted using Vicat apparatus, while the expansion test is done using Le-Chatelier mould. The initial, final set and expansion tests results are recorded on *physical test results of milled cement form*.

CHEMICAL ANALYSIS SECTION

Using the sample collected every two hours and accumulated for 24 for hours, the chemical analysis section performs several tests depending on the type of cement. For both PPC and OPC, this section performs specific test using Blain Apparatus and record the result as the specific surface value of the cement produced daily on *physical test results of milled cement form*. This test is considered as a physical test even if the chemical Analysis Section is the one performing the test.

If the 24 hours sample is OPC, the chemical analysis section daily performs the chemical tests of SO_3 , SiO_2 , Fe_2O_3 , $\text{L}_\text{O}_\text{I}$, I_R , Cl^- , Al_2O_3 , CaO , and MgO and record the results as the chemical analysis result of cement produced daily on *Chemical analysis form* (attached as annex 3). Using chemical test results obtained, the Chemical Analysis section also compute percentages of complex chemical compounds in OPC and record on the same form. The values recorded include SM, IM, LSF, C_3S , C_2S , C_3A , and C_4AF .

If the 24 hours sample is PPC, the chemical analysis section does not perform chemical tests on the 24 hours sample. Instead, the 24 hours samples are further accumulated for number of days ranging from four to eight days. Then average sample is taken and chemical tests of MgO , SO_3 , Cl^- and, $\text{L}_\text{O}_\text{I}$ are performed and then recorded on *Chemical analysis form*.

3.4 CEMENT QUALITY APPROVAL PROCESS AT MCE

The two quality control laboratories of MCE use the ASTM C-150 and C-595 standard chemical and physical requirements specifications to assess and approve cement quality. The OPC and PPC produced by the two factories of the Enterprise and tested in both quality control laboratories are designated as **Type I** of the ASTM C-150 standard and **Type IP** of the ASTM C-595 standard respectively.

Accordingly, the above-mentioned physical and chemical tests conducted by the laboratories are carried out in accordance with the test methods specified in the ASTM standard. The test results obtained for OPC and PPC are then compared with the requirements specifications outlined in the ASTM C-150 for OPC **Type I** and

C-595 for PPC *Type IP* standards respectively. Cements with test results complying with the ASTM standards are considered of good quality. For example, the 3, 7, and 28 days compressive strength test results obtained for OPC are compared with the ASTM C-150 standard strength requirements for *Type I* (12.4N/mm², 19.3N/mm², and 24.14N/mm² respectively) to judge OPC quality in terms of strength. Similar comparison with the ASTM standard is performed for PPC.

Even if the ASTM standard give a very strong emphasis on and define the 28 days compressive strength test result as a standard to judge quality of cement, the MCE can not afford to wait even the 7 days compressive strength test to complete in order to verify quality and authorize sale of cement to customer.

To authorize the sale of cement soon after produced, quality control experts follow the following procedure in the process of assessing and approving quality of cement produced daily:

- The chemical test for both OPC and PPC are completed with in a day and results are checked to see if they are with in the limits of the ASTM C-150 and C-595 standard chemical requirements respectively.
- The 3 days compressive strength test results are obtained within 72 hours, while the setting time, sieve residue and specific surface test results are obtained with in a day.
- The 3 days compressive strength results are checked first to see if they satisfy the ASTM C-150 and C-595 standard physical requirements for OPC and PPC respectively. If the 3 days compressive test result complies with the ASTM standard, the experts, taking into account other chemical and physical test results, predict as to what the 28 days compressive strength would be using their expertise. They then authorize the sale of cement in the silo represented by the test if predicted 28 days strength comply with the ASTM standard.
- If the 3 days compressive strength test result and/or the predicted 28 days strength do not comply with the ASTM standard, the cement in the cement silo represented by the test is considered as 'low quality' cement. The experts take a measure of ordering the production of 'higher quality' cement that can be mixed with the 'low quality' cement.
- For OPC this order usually involves grinding the cement finer than the previous days. For PPC the order may be either finer grinding of cement or decreasing the content of pumice or both.
- Several samples taken from both the 'higher quality' and the 'low quality' cements with varying proportion. On each sample, compressive strength test is conducted, and a 3 days compressive

strength test result of sample that comply with the ASTM standard is used as mixing proportion when packing the 'higher quality' and 'low quality' cements for sale.

- Up on request, Test Certificate is given to customers for the batch of cement supplied to them using OPC and PPC test certificate forms (attached as annex 4).

3.5 FINDINGS OF THE SURVEY

The existing cement quality assessment and approval process at MCE involves carrying out chemical and physical tests on representative sample of cement produced daily. A manual process of analyzing test results and predicting the 28 days strength follows this. The obtained test results and predicted values are compared with the ASTM C-150 and C-595 standard specifications, and cement quality is then approved.

The cement quality approval and hence authorization of cement for sale is effected even before the 7 days compressive strength test results are completed. That is, soon after the 3 days compressive results are obtained, while the 7 and 28 days compressive strength tests are still in progress.

In this cement quality approval process, the 3 days compressive strength test results are used as a main deciding factor. As a result, cement is already in the hands of customers and being utilized in construction works by the time the 'true' 7 and 28 days compressive strength test results are made available.

Even if the greatest precautions and experts' years of experience is used, this cement quality approval process, which involves the manual analysis of test results to predict the 28 days strength results, is prone to two types errors.

The first type of error is wrongly predicting the 28 days strength result to be below the minimum standard requirement while it is not. Such error cause the MCE to incur huge expenses that arise from the delay of delivery and further processing required by the production process to correct the cement in silo that is deemed to be of low quality while it not.

The second type of error, which is the most risky one, is wrongly predicting the 28 days strength result to be above the minimum standard and delivering low quality cement to customers. Such error prevents corrective measures from being taken before the cement is supplied and utilized in construction works. Consequently, this

type error may cause constructed building and structures to collapse resulting in loss of huge capital investment and lives of many.

Meanwhile, as seen in table 16 and table 17, MCE has been increasing its cement production capacity in the past years. At the same time, there is a high demand for cement in the country that every batch cements produced daily needs to be delivered as soon as possible to satisfy the demand. This situation has exerted too much pressure on quality control experts for they have the greatest responsibility to predict cement quality with the high degree of accuracy.

This leads to a question that how accuracy of predicting the 28 days compressive strength results can be quantified, measured and how this predictive accuracy may used in cement quality approval and sales authorization decision making process by quality control experts. To the researcher's belief, the answer lies in the application of data mining technology.

The chemical and physical test results of the quality control laboratories are used only for the purpose of determining the quality of a particular batch of cement for which the tests were conducted for. Then after, the test results are simply kept in record and used for no other purposes. The researcher has found out that the *chemical analysis* and *physical test on milled cement* forms and the compressive strength test results saved on computer file holds a wealth of cement test records data, which could help to discover and get a better understanding knowledge required to generate rules for predicting cement quality. The researcher believes that exploring these cement test records data may reveal new information that may provide cement quality predictive assistance that will used in cement quality approval decision-making process.

It is the researcher's belief that the availability of many cement test records data and business requirements necessitate the use of data mining techniques, for extracting hidden information that could help to make a more accurate prediction of cement quality. Employing data mining techniques would minimize the risks and costs that may arise due to possible errors that may occur in the manual cement quality assessment and approval decision-making process.

Based on the survey results, the next task is the collection and exploration of cement quality data from the available sources, and building predictive models by using data mining techniques in order to determine whether the results potentially be used in the cement quality assessment and approval decision-making process.

CHAPTER FOUR

EXPERIMENTATION

4.1 OVERVIEW

This chapter, as the fundamental part of this research work, first describes the tool used for conducting the experiment and then illustrates the major steps carried out when the experiments of the research were conducted, and finally presents and discusses the evaluation of experiments results.

The present research primarily follows and is structured based on the Cross-Industry Standard Process for Data Mining (CRISP-DM) process cycle that characterizes typical data mining steps (CRISP-DM, 2000). Thus, the major steps undertaken were data mining goal setting, data understanding, data preprocessing, modeling, and evaluation.

4.2 DATA MINING GOAL

As noted in chapter three, the existing practice of cement quality assessment and approval process at quality control laboratories of MCE is based on information gathered from chemical and physical tests performed on samples of cement taken from the production process. Therefore, the first data mining goal was to identify the possible data sources.

After the data sources have been identified, the next task performed is identification of the variables that are used to determine the quality of cement and using these variables to build models that can be used to predict quality of cement. This is believed by the researcher to facilitate the cement quality assessment and approval process of the MCE by supporting the tasks performed by quality control experts and providing predictive assistance in the decision making process.

Even if neural networks and other techniques could also be used, decision tree was chosen as appropriate data mining technique for building predictive models because the models can be built interactively, viewed and explained easily.

For this research work, the success factor is building highly accurate and interpretable models that would be used to predict the 28 days strength result and the quality of cement at the time the cement is produced and

the chemical test and the 3 days compressive strength results are available along with other physical tests. Provided that acceptable cement quality predictive models are built, the MCE could devise a means to minimize the problems and risks associated with the existing manual cement quality assessment and approval decision making process.

4.2.1 Data Mining Tool Selection

In conjunction with understanding the data mining goal, the key task is the selection of right data mining tool for conducting experiments of the research work. The data mining tool should be selected based on the tool's ability to support different methods for different steps of the experiments to be carried out.

According to Two Crows Corporation (1999) and Han and Kamber (2001), the check list for selecting data mining tool include:

- The data mining tasks that the tool is intended for;
- Algorithms supported that match with the characteristics of the problem and its sensitivity to missing or incorrect values;
- The system architecture on which the tool runs (stand-alone or client-server);
- The operating system on which the tool runs (Unix, Windows or other);
- Data preparation is by far the most time-consuming aspect of data mining. Everything a tool can do to ease this process will greatly expedite model development. Some of the functions that a product may provide include:
 - Data cleanup, such as handling missing data or identifying integrity violations;
 - Data description, such as row and value counts or distribution of values;
 - Data transformations, such as adding new columns, performing calculations;
 - Data sampling for model building or for the creation of training and validation data sets;
 - Selecting predictors from the space of variables;
- The ability to consolidate data from multiple sources;
- Scalability-the maximum number of records the tool can handle;
- Model evaluation and interpretation- facilities provided to understand the results (for example, providing measures of accuracy, significance, etc., in useful formats, such as confusion matrices).

Even though there are several data mining products that may fulfill all or some the above criteria, the data mining tool selection process in this research was mainly based on the recommendation given by the research advisor, Dr. Lars Asker, to use the Rule Discovery System software (RDS™) Modeling Edition for Windows, Version 2.0 from the Compumine AB (www.compumine.com). In addition to recommending the tool, Dr. Lars Asker also facilitated the issuance of free temporary Academic license by the licensor (Compumine AB) to the Department of Information Science; W/ro Ethiopia Tadesse, the current Department Chair Person, signed the academic license agreement on behalf of the department and downloaded the software from the Internet (license information is attached as Annex1).

This facilitated situation to acquire the tool for free of charge, in addition to the suitability of the tool for the research problem, contribute significantly for selecting this particular data mining tool. Therefore, the Rule Discovery System software (RDS™) Modeling Edition for Windows, Version 2.0 was the sole software used as the tool for the experiments conducted since it fulfilled most of the criteria set above and performed well during experimentation.

4.3 DATA UNDERSTANDING

After a good understanding of the data mining goal was gained, the next step to be performed was establishing an understanding of the data to be used. A good source of data for data mining purpose is the corporate data warehouse (Hans and Kamber (2001); Two Crows Corporation (1999)).

However, the available sources of information that held cement quality test records were stored in both manual and electronic format: the 3, 7, and 28 days compressive strength test records were in electronic format while the other physical test and all chemical records were in manual format.

The 3, 7 and 28 days compressive strength tests are carried out using computer controlled testing machine and the results are directly saved on the computer. These saved data are easily exportable to Microsoft Excel. The two documents that are used record the remaining physical and all chemical test records were *Physical Test Result of Milled Cement* form (attached as Annex 2) and *Chemical Analysis* form (attached as Annex 3) respectively.

The *Physical Test Result of Milled Cement* form contains test date, test number, fineness (specific surface and sieve residue), expansion (soundness) and setting time (initial and final) physical test result records and also

pumice content test result records. The *Chemical Analysis* form contains test date and all the chemical tests record. According to the survey conducted and based on experts' opinion, these manual forms together with strength test results saved on computer provided the required information for the experimentation.

Even though MCE has two cement factories one located in Addis Ababa (Addis Ababa Cement Plant) and other factory located 105 Km west of Addis Ababa (Mugher Cement Factory), the cement test results data from the quality control laboratory of Addis Ababa Cement Plant was only considered for two main reasons. Firstly, the two laboratories perform similar chemical and physical tests on cement using similar testing methods. Secondly, due to time constraint to access the data from quality control laboratory of Mugher Cement Factory that is geographically located far from Addis Ababa. Accordingly, the quality control laboratory of Addis Ababa Cement Plant was the branch laboratory whose data was collected as a sample for the present research.

The decision on the period to be covered by the research was one of the most important issues that needed careful consideration and experts' opinion before the actual data collection. Even though the quality control laboratory of Addis Ababa Cement Plant has been operational over a number of decades, the laboratory used to follow the British Cement Standard for assessing product quality before adopting the American Standard in January 2000. This makes the test records before January 2000 to be incompatible or non-comparable to those after January 2000. Consequently, test records starting from January 2000, which are compatible with the current practices of the lab, were considered and test records earlier than January 2000 were not considered and hence not collected.

April 2005 was the time that the data was collected from the quality control laboratory of Addis Ababa Cement Plant. Even though there were chemical and physical test result records for April 2005 by the time data was collected, these test records were not included since some of 7 days and most 28 days strength results were not yet completed and recorded. Therefore, March 2005 was the end of the period for the data collected. Consequently, the time period covered by the research was a period of about six years: from January 2000 to March 2005.

The final issue that needed consideration before data collection was determining the types of cement for which data would be collected. The Addis Ababa cement Plant produces three types of cements (OPC, PPC and SRPPC) and the quality control laboratory records test results for these three types of cements. However, a small number test results data were available for the Sulfate Resisting Portland Pozzolana (SRPPC). Asked

about the reason, quality control experts explained that SRPPC is produced occasionally with special order from customers and not many tests are performed for this type of cement in the laboratory. Accordingly, OPC and PPC were the only types of cements that test results data were collected for the purpose of building predictive models.

4.3.1 INITIAL DATA COLLECTION

As mentioned earlier, the data identified for the present research were available in both manual and electronic format. Therefore, the data in manual format had to be converted into an electronic format, since it is compulsory step to exercise data mining activity on data. Three activities were carried out in the data collection process for OPC and PPC data.

The first task was entering the OPC and PPC data that were in a manual format one-by-one into a separate Microsoft Excel files. The OPC data were entered into a single table **OPC_phy_chem** file and the PPC data were entered into a single table **PPC_phy_chem** file. This activity of entering the data was very difficult and time consuming because the researcher had to go through each *Chemical Analysis* and *Physical Test Result of Milled Cement* forms for both type of cements.

The second task was exporting the compressive strength test records of OPC and PPC that were saved on computer files to a separate Microsoft Excel files. The compressive strength test records of OPC were exported to a single table **OPC_comp_stre** file and the compressive strength test records of PPC were exported to a single table **PPC_comp_stre** file.

The data were entered in Microsoft Excel format to take advantage of easier data calculation and graphical facilities. Having the data in Microsoft Excel table had an advantage because the software chosen for analysis (i.e. Rule Discovery System software) has the facility to import data from text files and Microsoft Excel can easily save the file in tab-separated format that suitable is for the RDS tool once data manipulation is completed.

The third task was merging the two Microsoft Excel files for each type of cement into single table files; that is, merging **OPC_phy_chem** and **OPC_comp_stre** into a single table Excel file **OPC_dataset**, and merging **PPC_phy_chem** and **PPC_comp_stre** into a single table Excel file **PPC_dataset**. In order to match the **OPC_phy_chem** table with the **OPC_comp_stre** table, records in both tables needed to have the same test

date. This was also the case when merging *PPC_phy_chem* and *PPC_comp_stre*. This activity of merging the respective two tables for OPC and PPC data was very crucial as the cement test result records were contained in separate Excel files and the *OPC_dataset* and *PPC_dataset* files must contain only chemical and physical testes whose test dates match exactly.

Unfortunately, for both OPC and PPC data, many records existed in both respective tables without corresponding matching test dates; that is, there were many chemical test results whose matching or corresponding physical tests were not found and vises versa for both OPC and PPC data. In the *OPC_dataset* and *PPC_dataset* files, there were also records with different units of measure; for OPC data records, one field had missing values. A procedure was used correct the units and to fill the missing values in the 'data preprocessing' phase.

The resulting *OPC_dataset* and *PPC_dataset* Excel files contained 1,126 and 1,307 records of cement test results and were saved in a tab-separated text format, which is suitable for the RDS tool used to conduct experiments.

4.3.2 DESCRIPTION OF DATA COLLECTED

For the OPC cement type, the fields that are identified and collected from *Chemical Analysis* form, *Physical Test Result of Milled Cement* forms, and the compressive strength test records computer file are listed as follows:

1. Fields collected from the *Chemical Analysis* form:

- Test date
- Sulfur trioxide (SO₃)
- Silicon dioxide (SiO₂)
- Aluminum Oxide (Al₂O₃)
- Ferric Oxide (Fe₂O₃)
- Chloride (Cl⁻)
- Insoluble residue (I_R)
- loss on ignition (L_O_I)
- Calcium Oxide (CaO)
- Magnesium Oxide (MgO)

2. Fields collected from the *Physical Test Result of Milled Cement* form:

- Test date
- Test No.
- Initial set (I_S)
- % Sieve Residue (S_R)
- Specific Surface (S_S)
- Final Set (F_S)
- Expansion/Soundness (S)

3. Fields collected from the compressive strength test records computer file:

- Test date
- Test No.
- 3d (3 days compressive strength)
- 7d (7 days compressive strength)
- 28d (28 days compressive strength)

For the PPC cement type, the fields that are identified and collected from *Chemical Analysis* form, *Physical Test Result of Milled Cement* form and the *compressive strength test records* computer file are listed as follows:

1. Fields collected from the *Chemical Analysis* form:

- Test date
- Sulfur trioxide (SO₃)
- Magnesium Oxide (MgO)
- loss on ignition (L_O_I)

2. Fields collected from the *Physical Test Result of Milled Cement* form:

- Test date
- Test No.
- Initial set (I_S)
- Final Set (F_S)
- % Sieve Residue (S_R)
- Expansion/Soundness (S)
- Specific Surface (S_S)
- % pumice

3. Fields collected from the *compressive strength test records* computer file:

- Test date
- Test No.
- 3d (3 days compressive strength)
- 7d (7 days compressive strength)
- 28d (28 days compressive strength)

4.3.3 DATA QUALITY VERIFICATION

The modeling data for this research, as indicated earlier, is collected from both computer files and documents in manual format. The data from the computer files were exported to and the manual data were converted into Microsoft Excel files. Since wrong data have negative effect on the models to be built, utmost care was taken while performing these tasks.

The resulting data files were crosschecked with their corresponding original source documents and verified that, for both OPC and PPC data, each pair of physical and chemical tests records contain identical test dates. In consequence, decision was made to use the *OPC_dataset* and *PPC_dataset* files for the research at hand after the subsequent data preprocessing tasks are completed.

4.4 DATA PREPROCESSING

The main purpose of this phase was to generate the data sets that would be used for building the predictive models. Data preprocessing is an important activity in data mining process. Han and Kamber (2001) explain that it would enhance the excellence of the data thereby helping to improve the precision and effectiveness of

the subsequent data mining process. In accordance with CRISP-DM (2000), the main tasks during data preprocessing include data cleaning, data selection, data integration, and data transformation.

The OPC and PPC data that were collected for this research underwent essential steps of preprocessing. The tasks carried out under each case are discussed below. However of some tasks under data preprocessing were already done during the data collection process. For example, the OPC and PPC test result records were obtained from two manual documents and computer files. These data had to be integrated from different sources in one record for each type of cement during data collection. Therefore, task of data integration was done during data collection. However, each of the remaining tasks under data preprocessing are described as follows.

4.4.1 DATA CLEANING

This activity is about raising the quality of data to the level required by the research. According to CRISP-DM (2000), this may involve filling missing values, smoothing noisy data, and correcting data inconsistencies. The specific task that is carried out in this phase was handling missing values and is described below:

MISSING VALUES

Both the OPC and PPC data collected had some physical test records with missing values for '7 days' and/or '28 days' compressive strength attributes. That is, from the initially collected 1,126 OPC and 1,307 PPC test records, 45 OPC records and 36 PPC records had no value for 7 days and/or 28 days compressive strength attributes. The missing values of these records, as remarked by domain experts, were very important for determining the quality of cement. Therefore, decision was made to remove these records with missing values from the list of OPC and PPC data sets.

The collected OPC data also had many records with missing values for the 'Chloride' attribute. Actually, only 62 OPC chemical test records had recorded value for this attribute. After consulting the domain experts about cause for missing values, the Chloride (Cl⁻) chemical test was found to be performed occasionally.

For numeric attribute, Two Crows Corporation (1999) and Han and Kamber (2001) suggest that missing values to be replaced with the mean value for that field. Accordingly, this technique was adopted to handle missing values of the numeric 'Chloride' attribute. The mean value of the existing 62 values for the 'Chloride' attribute was 0.003644 and this mean value was used to replace the missing values.

4.4.2 DATA SELECTION

During this phase, selection decisions were made on the data sets to be used for developing models. As described in the data understanding phase, selection of data to be incorporated in the research work took into consideration the compatibility of cement quality test results and ease of access to data. Accordingly 1,126 OPC and 1,307 PPC test result records that span time period from January 2000 to March 2005 were initially collected from the quality control laboratory of Addis Ababa Cement Plant as modeling data for the present research.

Upon the opinions of domain experts, 45 (3.99%) OPC records and 36 (2.75%) PPC records with missing values of 7 days and/or 28 days strength results were disqualified. As a result, from the 1,126 OPC and 1,307 PPC records that were collected initially, the numbers of records after data selection were 1,081 OPC and 1,271 PPC data records with complete chemical and physical test results.

4.4.3 DATA TRANSFORMATION

This task, according to CRISP-DM (2000), includes data preparation operations, such as the production of derived attributes or transforming values for existing attributes. It is often necessary to construct new predictors variables derived from the raw data. Certain variables that have little effect alone may need to be combined with others using various arithmetic or algebraic operations (Two Crows Corporation, 1999).

The initially collected OPC and PPC data comprise attributes with different scale of measurements. For example, the Initial and final setting time attributes' values were found in both units of *Hours* and *Minutes*. These attributes' values were all transformed into unit of *Minutes*. The 3, 7, and 28 days compressive strength attributes' values were also found in both units of *Kilogram Force (KgF)* and *Mega Pascal (MPa)* and were all transformed in to unit of N/mm^2 .

According to quality control experts, the cement quality assessment process for OPC also takes into consideration supplementary derived attributes, whose values are obtained by combining values of attributes that are actually tested. Such derived attributes are 'Total', 'Undetermined', 'SM', 'IM', 'LSF', 'C₃S', 'C₂S', 'C₃A', and 'C₄AF'.

The 'TOTAL' attribute indicate the total sum of chemical test results except 'Insoluble Reside'. The 'Undetermined' attribute, which is 100 minus 'TOTAL', indicates the amount of other elements in cement but not analyzed. The 'SM', 'IM', 'LSF', 'C₃S', 'C₂S', 'C₃A', and 'C₄AF' attributes represent rations and amount of complex chemical compounds in OPC.

As suggested by domain experts, these attributes of OPC were derived from the values of existing fields and were incorporated to OPC data set as new fields.

Explanation on the how the supplementary derived attributes of OPC were computed is shown below:

- **Total** = (%SO₃ + %SiO₂ + % Fe₂O₃ + %L_O_I + %Cl + %Al₂O₃ + %CaO + %MgO)
- **Undetermined** = (100- Total)
- **SM** = [%SiO₂] ÷ [%Al₂O₃] + [% Fe₂O₃]
- **IM** = [%Al₂O₃] ÷ [% Fe₂O₃]
- **LSF** = (%CaO-) ÷ [(2.8x %SiO₂) + (1.2x %Al₂O₃) + (0.65x% Fe₂O₃)]
- **C₃S** = [(4.071 x %CaO) - (7.6x %SiO₂) - (6.718x %Al₂O₃) - (1.430 x % Fe₂O₃) - (2.852x%SO₃)]
- **C₂S** = (2.867x %SiO₂) - (0.7544x C₃S)
- **C₃A** = (2.650 %Al₂O₃) - (1.692 x % Fe₂O₃)
- **C₄AF** = (3.043 x % Fe₂O₃)

After the data preprocessing tasks were completed, a total of 26 fields were considered for the OPC data to be used for model building. The names and description of the fields are listed in the table 18 below:

No.	Field name	Description	unit	Data type	Source
1	SO ₃	Sulfur trioxide	Percent	Numeric	Chemical Analysis form
2	SiO ₂	Silicon dioxide	Percent	Numeric	Chemical Analysis form
3	Fe ₂ O ₃	Ferric oxide	Percent	Numeric	Chemical Analysis form
4	L_O_I	Loss on ignition	Percent	Numeric	Chemical Analysis form
5	I_R	Insoluble residue	Percent	Numeric	Chemical Analysis form
6	Cl ⁻	Chloride	Percent	Numeric	Chemical Analysis form
7	Al ₂ O ₃	Aluminum oxide	Percent	Numeric	Chemical Analysis form
8	CaO	Calcium oxide	Percent	Numeric	Chemical Analysis form
9	MgO	Magnesium oxide	Percent	Numeric	Chemical Analysis form
10	Total	Sum of above nine	Percent	Numeric	computed
11	Undetermined	100 minus Total	Percent	Numeric	computed
12	SM	Silica Modulus	Ratio	Numeric	computed
13	IM	Iron modulus	Ratio	Numeric	computed
14	100 LSF	Lime saturation factor	Percent	Numeric	computed
15	C ₃ S	Tircalcium silicate	Percent	Numeric	computed
16	C ₂ S	Dicalcium silicate	Percent	Numeric	computed
17	C ₃ A	Tri calcium Aluminate	Percent	Numeric	computed
18	C ₄ AF	Tetra calcium Alminoferrite	Percent	Numeric	computed
19	I_S	Initial set	Minutes	Numeric	Physical Test Result of Milled Cement form
20	F_S	Final Set	Minutes	Numeric	Physical Test Result of Milled Cement form
21	S_R	% Sieve Residue	Percent	Numeric	Physical Test Result of Milled Cement form
22	S	Soundness	mm	Numeric	Physical Test Result of Milled Cement form
23	S_S	Specific Surface	cm ² /gm	Numeric	Physical Test Result of Milled Cement form
24	3d	3 days compressive strength	N/mm ²	Numeric	Physical Test Result of Milled Cement form
25	7d	7 days compressive strength	N/mm ²	Numeric	Physical Test Result of Milled Cement form
26	28d	28 days compressive strength	N/mm ²	Numeric	Physical Test Result of Milled Cement form

Table 18: OPC data variables used for model building

After the data preprocessing tasks were completed, a total of 12 fields were considered for the PPC data to be used for model building. The names and description of the fields are listed in the table 19 below:

No.	Field name	Description	unit	Data type	Source
1	SO ₃	Sulfur trioxide	Percent	Numeric	Chemical Analysis form
2	L_O_I	Loss on ignition	Percent	Numeric	Chemical Analysis form
3	% pumice	Pumice content	Percent	Numeric	Chemical Analysis form
4	MgO	Magnesium oxide	Percent	Numeric	Chemical Analysis form
5	I_S	Initial set	Minutes	Numeric	Physical Test Result of Milled Cement form
6	F_S	Final Set	Minutes	Numeric	Physical Test Result of Milled Cement form
7	S_R	% Sieve Residue	Percent	Numeric	Physical Test Result of Milled Cement form
8	S	Soundness or expansion	mm	Numeric	Physical Test Result of Milled Cement form
9	S_S	Specific Surface	cm ² /gm	Numeric	Physical Test Result of Milled Cement form
10	3d	3 days compressive strength	N/mm ²	Numeric	Physical Test Result of Milled Cement form
11	7d	7 days compressive strength	N/mm ²	Numeric	Physical Test Result of Milled Cement form
12	28d	28 days compressive strength	N/mm ²	Numeric	Physical Test Result of Milled Cement form

Table 19: PPC data variables used for model building

4.5 MODELING

The tasks performed during the modeling phase were the selection of the modeling techniques, laying out a test design, building and evaluating models.

4.5.1 SELECTION OF MODELING TECHNIQUE

The major objective of this research is building classification and regression predictive models, using decision tree algorithms, for predicting cement quality. According to Compumine (2004), modeling techniques, methods or algorithms are instructions to RDS tool about how the predictive models should be constructed.

There are two types of decision tree models that can be constructed with Rule Discovery System for both classification and regression models:

■ **Tree model**

Tree model is a tree like representation of a prediction model consisting of hierarchically organized rules generated by recursive-partitioning. In tree model, the rules discovered are hierarchically organized, and the entire model can be visualized as a tree. The formal name for this model is recursive-partitioning and the resulting model tree is called a decision- tree.

■ **Ensemble model**

Ensemble model is a prediction model created by an ensemble method (an algorithm, such as bagging, for generating several prediction models and combining their individual predictions) and thus consists of multiple decision trees or rule sets. Ensembles are collections of predictions models that form consensus or composite models.

Bagging (bootstrap aggregation) generates an ensemble of models by repeatedly sampling the training data with replacement and building models based on each of the sub samples. The predictions of all the models in the set are combined to yield a single, ensemble prediction. Ensemble models are particularly useful when high prediction accuracy is wanted. Since the Ensemble models typically consists of a very large number of rules, they are not suitable for visualization

For selecting a particular modeling technique, Compumine (2004) remark that there is a trade off between *Creating an Interpretable model* verses creating a *Powerful Predictive model* as outlined below:

- Sometimes the interpretability may be even more important than the performance, if the goal is to let the human user gain an understanding of the problem domain. As a general rule, it is easier to understand and interpret a model which was not constructed using an ensemble scheme. Since an ensemble scheme uses a set of several models whose predictions are combined to yield the final prediction, the user would have to inspect several rule sets and understand how they are used in the classification (or regression) process. This could get very confusing.
- When the goal is to maximize a model's predictive power, it is often worthwhile to use an ensemble scheme, even if this leads to a less easily interpretable model. There is, however, no predictive algorithm

that always performs better than any other algorithm; the effectiveness of a given type of algorithm depends on properties of the data set. In practice, randomized bagging with 50 iterations using model tree as the base strategy usually produces good results.

The present research was concerned with building models for predicting quality of batches cements as good or bad, and also predicting 28 days strength values batches of cements. Therefore, accuracy of models created was important in modeling technique selection. Yet, the interpretability of the models created was also equally important. To address the need of creating models that are powerful and at the same time Interpretable, a decision was made to use both tree model and ensemble model techniques for building both classification and regression models.

The need to build both classification and regression predictive models for both OPC and PPC data sets were identified for the following reasons:

- a. If classification model is built for the cement data sets, then the model will classify a particular batch of cement to be of good or bad quality based on the new test results. Such information provided by the classification model is important for cement quality approval decision making process. However, the model does not tell the value of the 28 days strength. It just tells the class of cement as good or bad.
- b. If a regression model is built for the cement data set, then the model will tell as to what the value of the 28 days strength would be for the new example. The regression model has the advantage of predicting numeric values for the predicted variable, but does not tell about the quality status- the user has to decide by interpreting the generated result. Hence, the usability of the model is based on further involvement of person with skills to interpret the output.
- c. Classification has the advantage of easily determining Accuracy, Precision, Recall, and Confusion Matrix, while regression has Correlation Coefficient, Mean Error and Mean Squared Error.
- d. Classification and regression models are not comparable to each other since different tests are used for comparing the models. However, it would be very interesting to translate the results from regression models into classes and then calculate the Accuracy, Precision and Recall afterwards and see how this relates to the results obtained from the classification models.

With this understanding of the pros and cons of classification and regression models and the need to evaluate the performance of one model in terms of the other, decision was reached to build both types models for both OPC and PPC data sets.

4.5.2 TEST DESIGN

After the modeling techniques were selected and the types of models to be constructed have been identified, the next task was determining how to assess the models' predictive accuracy in a suitable way. In this step what was performed was selection of method that separate the data sets into training and testing set so as to build the models on the training set and estimate its performance on the testing set.

According to Compumine (2004), the method by which data is treated with the purpose of estimating the performance of models is called validation method. The validation method estimates the performance of model that can be created from the entire data set. The available validation method that can be selected and configured in RDS include *Split sample*, *External test set*, *N-fold cross validation*, *leave-X-out cross validation* and *group-wise cross validation*.

According to Compumine (2004), there are issues that need to be considered before selecting a particular validation method. When data is scarce and one wants to make the most of it, cross-validation can be a good option. If the data set is larger, it may be more prudent to use a separate test set, which has not been used at all during model construction.

As mentioned earlier, after data preprocessing, the number of test result records for OPC and PPC are 1,081 and 1,271 respectively. The researcher considered the sizes of these data to be small and opted for the 10-fold cross validation method to evaluate models' validity.

After selecting the validation method to be used, the next steps were running the modeling tool on the preprocessed OPC and PPC data sets to create the best classification and regression models. However, before starting the actual models building process, there were two more final issues that needed consideration and opinion of domain experts. The first issue was whether to use or not the 7 days strength variables in both OPC and PPC data set when building models. The second issue was related to class labels to be used when building classification model for both OPC and PPC data sets.

According to the findings of the survey outlined in chapter 3, all chemical and 3 days strength test result, together the rest physical test results are used in the cement quality assessment and approval process. Consequently, every batch of cement is already sold even before the 7 days compressive strength results are

obtained. This situation, hence, gave rise to make decision on using or not using the 7 days strength variable in building classification and regression models for both OPC and PPC data sets. To address this issue, the following three options were proposed:

- a. Exclude the 7 days strength variable and build the models using 28 days strength as dependent variable.
- b. Include the 7 days strength variable and build the models using 28 days strength as dependent variable.
- c. Build separate and independent models: one model using the using 7 days strength as dependent variable and second model using the 28 days strength as dependent variable. The two models would be used separately for assessing 7 and 28 days strength variables respectively.

These options were presented to and discussed with domain experts. It was found out that the first option is most appropriate to the MCE. The second option was not accepted since the Enterprise can not afford to wait four more days for the 7 days strength tests to be completed. The third option was also not interesting for the domain experts. Consequently, decision was made to build both classification and regression models excluding the 7 days strength variable from both OPC and PPC data sets.

Concerning the second issue, the option of using 'Good', 'Doubtful' and 'Bad' class labels as values of the 'quality' dependent variable was presented to domain experts. However, it was found out that setting up the threshold values between the three different classes is very problematic and subjective.

The domain experts then suggested that this issue should be addressed based on the minimum 28 days strength requirement of the ASTM C-150 and C-595 cement quality standards adopted by the MCE. Accordingly, it was decided to classify both types of cements into two classes of quality using class labels of 'Good' and 'Bad' for cement test records with 28 days strength value above and below the minimum strength requirements respectively.

Now that the all issues that need to be considered are resolved, the next step was running the RDS tool on the preprocessed OPC and PPC datasets to create the classification and regression models and evaluate accordingly.

4.5.3 MODEL BUILDING

EXPERIMENT ONE: *Classification model building for OPC data set*

Using the Rule Discover System, the steps required to build predictive models are: define a new modeling experiment, import a modeling data, verify or change variable types, select and configure validation method, add modeling methods to the experiment, and start the experiment. These steps carried out when building models are discussed below.

Step 1: Define a new modeling experiment

In RDS, a modeling experiment is what is done when one wants to build and validate predictions models from a modeling data set. To build a predictive model from the modeling data set, a modeling experiment need to be defined. Therefore, at this step, OPC classification modeling experiment was defined.

Step 2: Import a modeling data

The next step is importing a modeling data. The OPC data set Excel file saved in tab-separated text format was selected to be imported by the RDS software.

Step 3: Verify or change variable types

According to Compumine (2004), having the correct representation of the data is a prerequisite for obtaining good results. The representation is the set of variables, including their data types that are used to construct the models. The modeling data set must have exactly one prediction variable defined. The prediction variable is the variable that generated models will try to predict. Several variable types are included in RDS. However, the important variable types for the present research were:

- **Numeric**

The *numeric* type specifies that the values of the column are used by RDS to find numeric intervals based on the possible values of the variable in the training examples. The values of a numeric column must be numbers.

- **Ignore:**

The ignore type is used specify that a column should not be used in model construction.

- **Class**

The *class* type specifies that the column having this type will be used as the prediction variable for classification modeling. The *class* type specifies that the value of this column will be used as class labels. The *class* variable type is used to indicate that the resulting prediction model is to be used for classification tasks.

- **Regression**

The *regression* type is used to specify that the numeric values of the column will be used as prediction targets for regression modeling. The *regression* variable type is used to indicate that the resulting prediction model is to be used for regression tasks.

In the OPC and PPC data sets, all variables are numeric. This was verified by displaying the data clicking "modeling data set" hyperlink in RDS program window. In addition, since experiment one was concerned with building classification model, this was configured so by setting the dependent variable 'quality' to be of 'Class' type. The 7 days strength variable was set to 'Ignore' type.

Step 4: Select and configure validation method

In order to verify models generated by RDS are valid and assess their accuracy, they need to be validated on data that was not used to generate them. The available validation method that can be selected and configured in RDS include *Split sample*, *External test set*, *N-fold cross validation*, *leave-X-out cross validation* and *group-wise cross validation*. However, for reasons explained in section 4.5.2, the 10-fold cross validation was method selected for purpose of estimating the performance of the classification models.

Step 5: Add modeling methods to the experiment

The available modeling method or algorithms that can be selected and configured in RDS include *Tree model*, *Rule set model*, and *Ensemble model* that can also be further customized. However, for reasons explained in section 4.5.1, the *tree model* and *ensemble tree model* were the modeling methods selected for purpose of instructing the RDS about how the classification models should be constructed. The tree model method was configured with no additional parameters. For Ensemble model method *Bagging*, *Random Variable Selection*, and the number of models in ensemble to be 50 were configured.

Modeling results of Experiment one

The model statistics (Accuracy, Precision, Recall, and Confusion matrix) for each fold of **tree model** is shown in table 20 below:

Method performance							
Method	Number of rules	Accuracy	Correct class			Predicted class	
				Precision	Recall	bad	good
Tree fold 1	2	96.330	bad	0.667	1.000	8	0
			good	1.000	0.960	4	97
Tree fold 2	2	100.000	bad	1.000	1.000	8	0
			good	1.000	1.000	0	100
Tree fold 3	2	99.074	bad	0.889	1.000	8	0
			good	1.000	0.990	1	99
Tree fold 4	3	96.296	bad	0.667	1.000	8	0
			good	1.000	0.960	4	96
Tree fold 5	3	98.148	bad	0.800	1.000	8	0
			good	1.000	0.980	2	98
Tree fold 6	6	99.074	bad	0.889	1.000	8	0
			good	1.000	0.990	1	99
Tree fold 7	6	96.296	bad	0.700	0.875	7	1
			good	0.990	0.970	3	97
Tree fold 8	2	99.074	bad	0.889	1.000	8	0
			good	1.000	0.990	1	99
Tree fold 9	2	95.370	bad	0.615	1.000	8	0
			good	1.000	0.950	5	95
Tree fold 10	3	98.148	bad	0.778	1.000	7	0
			good	1.000	0.980	2	99

Table 20: Model statistics for tree model of experiment one

The model statistics (Accuracy, Precision, Recall, and Confusion matrix) for each fold of **Ensemble tree model** is shown in table 21 below:

Method performance							
Method	Number of rules	Accuracy	Correct class			Predicted class	
				Precision	Recall	bad	good
Ensemble fold 1	1198	96.330	bad	0.667	1.000	8	0
			good	1.000	0.960	4	97
Ensemble fold 2	1303	100.000	bad	1.000	1.000	8	0
			good	1.000	1.000	0	100
Ensemble fold 3	1234	98.148	bad	1.000	0.750	6	2
			good	0.980	1.000	0	100
Ensemble fold 4	1284	98.148	bad	0.800	1.000	8	0
			good	1.000	0.980	2	98
Ensemble fold 5	1205	98.148	bad	0.800	1.000	8	0
			good	1.000	0.980	2	98
Ensemble fold 6	1282	98.148	bad	0.800	1.000	8	0
			good	1.000	0.980	2	98
Ensemble fold 7	1291	96.296	bad	0.833	0.625	5	3
			good	0.971	0.990	1	99
Ensemble fold 8	1237	99.074	bad	0.889	1.000	8	0
			good	1.000	0.990	1	99
Ensemble fold 9	1086	93.519	bad	0.545	0.750	6	2
			good	0.979	0.950	5	95
Ensemble fold 10	1241	98.148	bad	0.857	0.857	6	1
			good	0.990	0.990	1	100

Table 21: Model statistics for ensemble tree model of experiment one

Over all result of experiment one is shown in table 22 and the performance comparison is depicted in table 23 below:

Method performance							
Method	Number of rules	Accuracy	Correct class		Predicted class		
				Precision	Recall	bad	good
Tree	3.1	97.780	bad	0.772	0.987	78	1
			good	0.999	0.977	23	979
Ensemble	1236.1	97.595	bad	0.798	0.899	71	8
			good	0.992	0.982	18	984

Table 22: Over all result of experiment one

Performance comparison matrix		
P-values calculated by McNemar's test		
	Tree	Ensemble
Tree	-	0.79
Ensemble	0.79	-

Table 23: Performance comparison of models of experiment one

EXPERIMENT TWO: Regression model building for OPC data set

This experiment to build regression model for the OPC data set was conducted using steps and parameter settings similar to the previous experiment (experiment one) conducted for building classification model, except in this case the dependent variable is the 28 days strength and it is of 'regression' type.

Modeling results of Experiment two

The model statistics (Correlation Coefficient, Mean Error and Mean Squared Error) for each fold of **tree model** is shown in table 24 below:

Method performance				
Method	Number of rules	Mean error	Mean squared error	Correlation coefficient
Tree fold 1	51	2.741	12.384	0.814
Tree fold 2	39	3.505	18.803	0.783
Tree fold 3	51	3.038	14.891	0.822
Tree fold 4	34	3.216	16.926	0.806
Tree fold 5	42	3.154	16.525	0.785
Tree fold 6	43	3.241	19.404	0.732
Tree fold 7	39	2.943	14.784	0.641
Tree fold 8	42	3.064	14.483	0.845
Tree fold 9	37	3.057	17.156	0.794
Tree fold 10	50	2.976	16.673	0.766

Table 24: Model statistics for tree model of experiment two

The model statistics (Correlation Coefficient, Mean Error and Mean Squared Error) for each fold of **Ensemble tree model** is shown in table 25 below:

Method performance				
Method	Number of rules	Mean error	Mean squared error	Correlation coefficient
Ensemble fold 1	16349	2.457	10.774	0.866
Ensemble fold 2	16382	3.259	16.125	0.806
Ensemble fold 3	16500	2.953	13.814	0.847
Ensemble fold 4	16445	3.264	16.554	0.818
Ensemble fold 5	16589	3.023	15.198	0.808
Ensemble fold 6	16455	3.021	15.189	0.790
Ensemble fold 7	16535	2.981	15.474	0.606
Ensemble fold 8	16553	3.157	14.715	0.855
Ensemble fold 9	16466	2.994	14.815	0.821
Ensemble fold 10	16324	2.921	14.560	0.811

Table 25: Model statistics for ensemble tree model of experiment two

Over all result of experiment two is shown in table 26 and the performance comparison is depicted in table 27 below:

Method performance				
Method	Number of rules	Mean error	Mean squared error	Correlation coefficient
Tree	42.8	3.093	16.199	0.784
Ensemble	16459.8	3.002	14.718	0.809

Table 26: Over all result of experiment two

Performance comparison matrix		
P-values calculated by paired t-test		
	Tree	Ensemble
Tree	-	9.96E-5
Ensemble	9.96E-5	-

 Significant difference

Table 27: Performance comparison of models of experiment two

EXPERIMENT THREE: Classification model building for PPC data set

This experiment to build classification model for the PPC data set was conducted using steps and parameter settings similar to experiment one.

Modeling results of Experiment Three

The model statistics (Accuracy, Precision, Recall, and Confusion matrix) for each fold of **tree model** is shown in table 28 below:

Method performance							
Method	Number of rules	Accuracy	Correct class		Predicted class		
				Precision	Recall	bad	good
Tree fold 1	6	96.875	bad	1.000	0.886	31	4
			good	0.959	1.000	0	93
Tree fold 2	6	98.425	bad	1.000	0.943	33	2
			good	0.979	1.000	0	92
Tree fold 3	7	93.701	bad	1.000	0.771	27	8
			good	0.920	1.000	0	92
Tree fold 4	8	98.425	bad	0.971	0.971	34	1
			good	0.989	0.989	1	91
Tree fold 5	7	99.213	bad	1.000	0.971	34	1
			good	0.989	1.000	0	92
Tree fold 6	7	95.276	bad	0.892	0.943	33	2
			good	0.978	0.957	4	88
Tree fold 7	8	99.213	bad	1.000	0.971	34	1
			good	0.989	1.000	0	92
Tree fold 8	9	97.638	bad	0.971	0.943	33	2
			good	0.978	0.989	1	91
Tree fold 9	6	98.425	bad	0.971	0.971	33	1
			good	0.989	0.989	1	92
Tree fold 10	12	96.850	bad	0.969	0.912	31	3
			good	0.968	0.989	1	92

Table 28: Model statistics for tree model of experiment three

The model statistics for each fold of **Ensemble tree model** is shown in table 29 below:

Method performance							
Method	Number of rules	Accuracy	Correct class			Predicted class	
				Precision	Recall	bad	good
Ensemble fold 1	1484	96.875	bad	0.970	0.914	32	3
			good	0.968	0.989	1	92
Ensemble fold 2	1596	99.213	bad	1.000	0.971	34	1
			good	0.989	1.000	0	92
Ensemble fold 3	1563	96.850	bad	0.970	0.914	32	3
			good	0.968	0.989	1	91
Ensemble fold 4	1474	99.213	bad	1.000	0.971	34	1
			good	0.989	1.000	0	92
Ensemble fold 5	1588	98.425	bad	0.971	0.971	34	1
			good	0.989	0.989	1	91
Ensemble fold 6	1450	98.425	bad	1.000	0.943	33	2
			good	0.979	1.000	0	92
Ensemble fold 7	1713	99.213	bad	1.000	0.971	34	1
			good	0.989	1.000	0	92
Ensemble fold 8	1546	99.213	bad	0.972	1.000	35	0
			good	1.000	0.989	1	91
Ensemble fold 9	1459	98.425	bad	1.000	0.941	32	2
			good	0.979	1.000	0	93
Ensemble fold 10	1575	96.850	bad	0.941	0.941	32	2
			good	0.978	0.978	2	91

Table 29: Model statistics for ensemble tree model of experiment three

Over all result of experiment three is shown in table 30 and performance comparison is depicted in table 31 below:

Method performance							
Method	Number of rules	Accuracy	Correct class			Predicted class	
				Precision	Recall	bad	good
Tree	7.6	97.404	bad	0.976	0.928	323	25
			good	0.973	0.991	8	915
Ensemble	1544.8	98.269	bad	0.982	0.954	332	16
			good	0.983	0.993	6	917

Table 30: Over all result of experiment three

Performance comparison matrix		
P-values calculated by McNemar's test		
	Tree	Ensemble
Tree	-	0.043
Ensemble	0.043	-

 Significant difference

Table 31: Performance comparison of models of experiment three

EXPERIMENT FOUR: Regression model building for PPC data set

This experiment to build regression model for the PPC data set was conducted using steps and parameter settings similar to experiments two.

Modeling results of Experiment four

The model statistics (Correlation Coefficient, Mean Error and Mean Squared Error) for each fold of **tree model** is shown in table 32 below:

Method performance				
Method	Number of rules	Mean error	Mean squared error	Correlation coefficient
Tree fold 1	55	2.152	7.982	0.829
Tree fold 2	58	2.023	6.846	0.800
Tree fold 3	50	2.065	6.945	0.838
Tree fold 4	60	2.201	8.594	0.781
Tree fold 5	52	2.003	6.345	0.836
Tree fold 6	47	2.233	8.703	0.811
Tree fold 7	57	2.199	8.235	0.832
Tree fold 8	35	1.964	6.844	0.833
Tree fold 9	57	1.985	7.699	0.826
Tree fold 10	50	2.143	8.132	0.829

Table 32: Model statistics for tree model of experiment four

The model statistics (Correlation Coefficient, Mean Error and Mean Squared Error) for each fold of Ensemble model tree is shown in table 33 below:

Method performance				
Method	Number of rules	Mean error	Mean squared error	Correlation coefficient
Ensemble fold 1	19891	1.883	5.954	0.873
Ensemble fold 2	20017	1.656	4.482	0.869
Ensemble fold 3	19908	1.740	4.806	0.896
Ensemble fold 4	19939	1.807	6.150	0.848
Ensemble fold 5	19959	1.776	4.939	0.878
Ensemble fold 6	19985	1.979	6.707	0.865
Ensemble fold 7	19930	1.800	5.613	0.891
Ensemble fold 8	19979	1.692	4.744	0.890
Ensemble fold 9	20001	1.815	5.304	0.893
Ensemble fold 10	19923	1.923	6.119	0.872

Table 33: Model statistics for ensemble model tree of experiment four

Over all result of experiment four is shown in table 34 and the performance comparison is depicted in table 35 below:

Method performance				
Method	Number of rules	Mean error	Mean squared error	Correlation coefficient
Tree	52.1	2.097	7.633	0.822
Ensemble	19953.2	1.807	5.482	0.877

Table 34: Over all result of experiment four

Performance comparison matrix		
P-values calculated by paired t-test		
	Tree	Ensemble
Tree	-	0.65
Ensemble	0.65	-

Table 35: Performance comparison of models of experiment four

4.5.4 EVALUATION OF MODELS

According to Hans and Kamber (2001) and Compumine (2004), classification models can be evaluated and compared according to the following criteria: Accuracy, Precision, Recall, Confusion Matrix, and interpretability. For comparing and evaluating performance of regression models, Whitten and Frank (2000) and Compumine (2004) suggest using Mean error, Mean squared Error, Correction coefficient, and interpretability. Accordingly, the models developed were evaluated and compared as follows.

Evaluation and comparison of Models from Experiment one: *Classification models for OPC data set*

In this experiment, classification models for OPC were built using tree model and ensemble tree methods. As seen in table 22, the classification model built using tree method has an accuracy of 97.780%, while the model built using the ensemble method has an accuracy of 97.595%. Even if higher predictive accuracy was expected from the model built with ensemble method, this was not the case- both these models have similar high predictive accuracy and the performance comparison of the two models, as shown in table 23, is not statistically significant since the *P-values* calculated by McNemar's test is 0.79.

The two models also have somehow similar precision but different recall and confusion matrix. The model built with tree model method failed to correctly classify one example as 'bad' quality, while the model built with the ensemble method failed to correctly classify 8 examples as 'bad' quality. In predicting 'good' quality examples, the models fail to correctly predict 23 and 18 examples respectively. Error in predicting 'bad' quality cement is more risky than error in predicting 'good' quality cement. Domain experts also share the same idea.

In terms of interpretability, the model built with ensemble method is inherently less interpretable-since ensemble models typically consist of a very large number of rules, they are not suitable for visualization. As seen in table 22, this was proved to be true for the ensemble model built as it has an average of 1,236.1 rules, while the model built with tree model method has an average of 3.1 rules. The soundness of the rules generated by the model built with the tree model method was also considered as additional evaluation criteria. According to domain experts, the cement quality variables used at the root and subsequent nodes of the decision tree model built with the tree model method were superior variables and the rules generated were sounder to the problem domain.

As a result, with all things considered, the predictive classification model built using the tree model method was preferred for its performance and interpretability.

Evaluation and comparison of Models from Experiment two: Regression models for OPC data set

In this experiment, predictive regression models for OPC were built using tree model and ensemble tree methods. As seen in table 26, the mean error, mean squared error and correlation coefficient values of regression model built using tree model method are 3.093, 16.199, and 0.784 respectively. These values for the model built using ensemble method are 3.002, 14.718, and 0.809 respectively. Hence, the model built using ensemble method performed better. In addition, as shown in table 27, the performance comparison of the two models is statistically significant since the *P-values* calculated by *t-test* is 9.96E-5.

In addition to the above comparison the models based on mean error, mean squared error and correlation coefficient values, an attempt was made to translate the prediction results from the two regression models into classes and then calculate the Accuracy, Precision and Recall afterwards in order to compare these values with the Accuracy, Precision and Recall values achieved through classification models of experiment one.

For this purpose, the prediction results of the two regression models for the dependent variable (28 days strength) were exported from RDS tool to Microsoft Excel and then translated to class values of 'Good quality' and 'Bad quality'. Using Excel, Accuracy, Precision and Recall values were calculated afterwards. These values obtained for the two regression models are shown in table 36 below together with the corresponding Accuracy, Precision and Recall results of the classification models of experiment one. This activity allowed the comparison of regression and classification models which other wise were not comparable to each other.

Experiment	Model	Method	No of rules	Accuracy	Correct class			Predicted class	
						precision	recall	bad	good
one	Classification	Tree	3.1	97.780					
					Bad	0.772	0.987	78	1
					Good	0.999	0.977	23	979
two	Regression	Tree	42.8	95.745		precision	recall	bad	good
					Bad	0.739	0.646	51	28
					Good	0.972	0.82	18	984
one	Classification	Ensemble	1236.1	97.595		precision	recall	bad	good
					Bad	0.798	0.899	71	8
					Good	0.992	0.982	18	984
two	Regression	Ensemble	16,459.8	93.154		precision	recall	bad	good
					Bad	0.667	0.127	10	69
					Good	0.935	0.995	5	997

Table 36: Performance comparison of models of experiment one and two

As seen in table 36 above, the two regression models have different predictive accuracy: the regression model built using the tree model method has 95.745%, while the regression model built using the ensemble method has 93.154%. The precision and recall values and the confusion matrices of the two regression models also differ significantly. The regression model built using the tree model method wrongly predicted 28 examples of 'Bad quality' as 'Good quality' and 18 examples of 'Good quality' as 'Bad quality', while the regression model built using the ensemble method wrongly predicted 69 examples of 'Bad quality' as 'Good quality' and 5 examples of 'Good quality' as 'Bad quality'.

As noted earlier, error in predicting 'Bad quality' cement is more risky than error in predicting 'Good quality' cement. Accordingly, the regression model built using the tree model method performed better than the regression model built using the ensemble method in terms of accuracy, precision and recall.

Further evaluation of the two regression models was performed by comparing the accuracy, precision and recall of the regression models built with tree model method and ensemble method with the accuracy, precision and recall of the corresponding classification models of experiment one.

Tables 36 above shows that the accuracy of the classification and regression models built with tree model method are 97.780% and 95.745% respectively. Even if these accuracy values are somehow similar, the precision and recall values and the confusion matrices of these models differ significantly. Table 36 also shows that the accuracy of the classification and regression models built with ensemble method are 97.595% and 93.145% respectively. This time, not only the precision and recall values and the confusion matrices, but also the accuracy values of the classification and regression models built with ensemble method differ significantly.

In terms of interpretability, the regression model built using ensemble method is less interpretable since, as shown in table 26, it has an average of 16,459.8 rules, while the regression model built with tree model method has an average of 42.8 rules.

The final criteria used to evaluate the regression model built using the tree model method was examination of the soundness of the rules generated. This was done together with domain experts and rules generated were found to be sounder to the problem domain.

As a result, with all things considered, the regression model built using the tree model method was preferred for its performance and interpretability.

Evaluation and comparison of Models from Experiment three: Classification models for PPC data set

In this experiment, predictive classification models for PPC were built using tree model and ensemble tree methods. Evaluation and comparison of the classification models of this experiments was done using similar approaches used evaluate and compare classification models of experiment one.

As seen in table 30, the classification model built using tree model method has an accuracy of 97.404%, while the model built using the ensemble method has an accuracy of 98.269%. Both theses models have high predictive accuracy and the model built with ensemble method has a higher predictive accuracy as expected from such model. Besides, as shown in table 31, the performance comparison of the two models is statistically significant since the *P-values* calculated by McNemar's test is 0.043.

The precision and recall values and the confusion matrices of the two classification models also differ. As seen in table 30, the model built with tree method wrongly classified 25 examples as 'Bad quality'; the model built with the ensemble method wrongly classified 16 examples as 'Bad quality'. In predicting 'Good quality' examples, the models wrongly predicted 8 and 6 examples respectively. As a result, the model built with the ensemble method has a higher predictive performance.

In terms of interpretability, the classification model built using ensemble method is less interpretable since, as shown in table 30, it has an average of 1,544.8 rules, while the classification model built with tree model method has an average of 7.6 rules.

As final evaluation criteria, the soundness of the rules generated by the classification model built with the tree model method were assessed by the domain experts, and were found to be sounder to the problem domain. However, even if the models generated by the ensemble model could not be visualized, domain experts still considered the classification model built using ensemble method to be the best mainly due to the higher prediction error of the model built with tree model method to correctly predict 'Bad quality' cement, which is the type of error that is more risky.

As a result, with all things considered, the classification model built using the ensemble method was preferred mainly for its performance.

Evaluation and comparison of Models from Experiment four: Regression models for PPC data set

In this experiment, predictive regression models for PPC were built using tree model and ensemble tree methods. Evaluation and comparison of the regression models of this experiments was done using similar approaches used evaluate and compare regression models of experiment two.

As seen in table 34, the mean error, mean squared error and correlation coefficient values of regression model built using tree model method are 2.097, 7.633, and 0.822 respectively. These values for the model built using ensemble method are 1.807, 5.482, and 0.877 respectively. Hence, the model built using ensemble method performed better. However, as shown in table 35, the performance comparison of the two models is not statistically significant since the *P-values* calculated by *t-test* is 0.65.

In addition to the above comparison the models based on mean error, mean squared error and correlation coefficient values, an attempt was made to translate the prediction results from the two regression models into classes and then calculate the Accuracy, Precision and Recall afterwards in order to compare these values with the Accuracy, Precision and Recall values achieved through classification models of experiment three.

For this purpose, the prediction results of the two regression models for the dependent variable (28 days strength) were exported from RDS tool to Microsoft Excel and then translated to class values of 'Good quality' and 'Bad quality'. Using Excel, Accuracy, Precision and Recall values were calculated afterwards. These values obtained for the two regression models are shown in table 37 below together with the corresponding Accuracy, Precision and Recall results of the classification models of experiment three.

Experiment	Model	Method	No of rules	Accuracy	Correct class			Predicted class	
						precision	recall	bad	good
three	Classification	Tree	7.6	97.404					
					Bad	0.976	0.928	323	25
					Good	0.973	0.991	8	915
four	Regression	Tree	52.1	97.325		precision	recall	bad	good
					Bad	0.978	0.922	321	27
					Good	0.971	0.992	7	916
three	Classification	Ensemble	1544.8	98.269		precision	recall	bad	good
					Bad	0.982	0.954	332	16
					Good	0.983	0.993	6	917
four	Regression	Ensemble	19953.2	97.639		precision	recall	bad	good
					Bad	0.990	0.922	321	27
					Good	0.971	0.997	3	920

Table 37: Performance comparison of models of experiment three and four

CHAPTER FIVE

CONCLUSION AND RECOMMENDATION

5.1 CONCLUSION

The objective for undertaking this research was to explore the possible application of data mining technology at the Mugher Cement Enterprise for developing classification and regression models that could provide predictive assistances in cement quality assessment and approval decision making processes.

The CRISP-DM process model was the methodology adopted for undertaking this research that included steps of data collection, data preprocessing, model building and evaluation. However, these steps were not followed strictly in linear order for the fact that data mining is an iterative process that necessitated the need to go back and forth between the steps.

In order to build the classification and regression models, 1081 and 1271 cement quality test records were collected for Ordinary Portland Cement (OPC) and Portland Pozzolana Cement (PPC) respectively. Several models were built, using both tree and ensemble tree modeling techniques and employing 10-fold cross-validation. The models built were then evaluated for their predictive performance and interpretability in order to identify best models.

The classification model identified to be the best for OPC was the one built using the tree model method and has predictive accuracy of 97.780%. The classification model identified to be the best for PPC was the one built using ensemble method and has predictive accuracy of 98.269%.

The regression model identified to be the best for OPC was the one built using the tree model method and has mean error, mean squared error and correlation coefficient of 3.093, 16.199, and 0.784 respectively. The regression model identified to be the best for PPC was the one built using the tree model method and has mean error, mean squared error and correlation coefficient of 2.097, 7.633, and 0.822 respectively.

The prediction results from the regression models of OPC and PPC were translated to class labels of 'Good quality' and 'Bad quality' and accuracy, precision, and recall were calculated afterwards. The regression

model identified to be the best for OPC has predictive accuracy of 95.745% and the regression model identified to be the best for PPC has predictive accuracy of 97.325%.

The role of domain experts during the entire course of the research was not only supportive but also crucial. Resolution of some issues, especially evaluation of models, was only possible with suggestions and recommendations provided by domain experts.

The results found from this research are very encouraging and have revealed the potential applicability of data mining technology in support of cement quality assessment and approval decision making process. The results have demonstrated that cement quality can be reliably predicted using models developed using data mining technology. However, other than providing predictive assistance in cement quality decision making process, the models by no means would replace the domain experts.

The results found in this research are so inviting for exploring possible applications of data mining technology in product quality assessment tasks in other manufacturing industry sectors.

5.2 RECOMMENDATIONS

Basically, this research was conducted for an academic purpose. However, it has revealed the potential applicability of data mining technology to support the cement quality assessment activity at Mugher Cement Enterprise based on historical test results data accumulated on batches of cements. Moreover, it is the researcher's belief that the contribution of this research work could be a good experience for further future studies in the application of data mining technology in quality assessment activities in other manufacturing industry sectors.

Thus, based on the findings of this research work, the researcher would like to make the following recommendations particularly in relation to the possible application of data mining technology in supporting cement quality assessment and to other activities in Mugher Cement Enterprise:

- The availability of data in electronic format is prerequisite for data mining tasks. Therefore, Mugher Cement Enterprise (MCE) should take measures to store all its cement quality test result records in an electronic format.

- Data mining technology could provide predictive assistance in cement quality assessment and approval decision making process. Thus, it could be more important for MCE to use data mining technique as a tool in the cement quality assessment and decision making process.
- RDS tool used for this research comes in two flavors: the Modeling Edition and the Deployment Edition. The Modeling Edition can be used for developing models, while the Deployment Edition allows using existing models (developed by the Modeling Edition) for making predations. Hence, the MCE should buy license and use the Deployment Edition RDS software with the stored models as the interface.
- Although in this study encouraging results were obtained, a small sample data was used for developing models. However, it is known that a data mining task benefits from and provides better results through the use of larger data set. Hence, it is appropriate to conduct the experiments with larger datasets to generate predictive models with higher performance.
- Further experiments should also be made on the same OPC and PPC data sets using neural network or other techniques in order to make comparison between the corresponding models and determine which techniques will result in a better predictive performance.
- Further experiments on the possible application of data mining to support cement production process control and cements raw materials quality assessment activities at the Mugher Cement Enterprise (MCE) need to be explored.

REFERENCE

1. Ali Nadeem (2003): Cement Quality: Production and quality control in cement industry seminar technical report, Qatar
2. ASTM C150-Standard Specification for Portland Cement (1998): America Society for Testing Materials, Pennsylvania
3. ASTM C595-Standard specification for Blended Cement (1998): America Society for Testing Materials, Pennsylvania
4. Atef Natsheh (2001): Analytical methods: Cement production and quality control workshop technical report, Jordan
5. Berry, M. J. A. et al. (2000): Mastering Data Mining. John Wiley & Sons Inc, New York
6. Berry, M. J.A. et al. (1997): Data Mining Techniques, For Marketing Sales and Customer Support: John Wiley and Sons Inc, New York.
7. Bigus, J.P. (1996). Data Mining With Neural Networks in solving Business Problems- From Application to Development to Decision Support: McGraw-Hill, New York
8. British Standard BS:12 (1996):British standard institute (BSI), London
9. Claude Seidman (200): Data mining with Microsoft SQL server 2000. Microsoft press, Washington.
10. Compumine Company AB (2004): Rule Discover System™ 2.0 Modeling Edition User Guide: <http://www.compumine.com>
11. Compumine Company AB (2004): Rule Discover System™ 2.0 Modeling Edition Command Line Manual and Reference Guide: <http://www.compumine.com>
12. CRISP-DM. CRISP-DM 1.0 (2000): Step-by-step data mining guide: <http://www.crisp-dm.org>.
13. Dr. Kamel Al Khatib (2003): Cement industry: Production and quality control in cement industry seminar technical report, Qatar
14. ENV197-1 European Pre-standard (1998): Standard Cement Specification. European Committee for Standardization, Brussels
15. H. Whitten, I. & Frank, E. (2000): Data Mining : Practical Machine Learning Tools and Techniques with Java Implementations, Academic Press, USA
16. Han, J., & Kamber M. (2001):Data Mining, Concepts and Techniques, Academic Press, USA
17. Hani Salah (2001): General concepts of quality and quality control: Cement production and quality control workshop technical report, Jordan
18. Henock Woubishet (2002): Application of Data mining technique to support customer relationship

- management at Ethiopian Airlines. Addis Ababa University, Addis Ababa.
19. Hyunjoong Kim, Frank M. Guess, and Timothy M. Young (2002): Using data mining tools of decision trees in reliability applications. University of Tennessee, Knoxville, TN: www.utk.edu
 20. ISO 9000 (2000): Quality Management and Quality Assurance Guidelines, International Organization for Standardization (ISO)
 21. Jim Gray (2001): Microsoft Research. Microsoft Corporation, Washington
 22. Jong-Hag Jeon (2000): Data Mining Applications for a Six-Sigma Project, POSCO, Pohang, South Korea: www.posco.co.kr
 23. Mohd Asim (2003): Process Kinetics: Production and quality control in cement industry seminar technical report, Qatar
 24. Mugher Cement Enterprise (2002): Overview of MCE, Addis Ababa
 25. Mugher Cement Enterprise (2004): Production Report 2000-2004, Addis Ababa
 26. Peter C. Hewlett (1998): Lea's Chemistry of Cement and Concrete. Arnold Publishers, London.
 27. Public Relations Department (2002): Mugher Cement Enterprise. Addis Ababa
 28. Ravi M. Rao (2003): Clinker and cement process: Production and quality control in cement industry seminar technical report, Qatar
 29. Richard L. Tomasetti (2004): Portland cement, Microsoft Encarta Encyclopedia, Microsoft Corporation, Washington
 30. Samir Akkawi (2001): Cement standards: Cement production and quality control workshop technical report, Jordan
 31. StatSoft (2003): Data mining techniques: <http://www.statsoftinc.com/textbook>
 32. Thearling K. (2003): An Introduction to Data Mining:
<http://www3.primuhost.com/~kht/text/dmwhite/dmwhite.html>
 33. Thearling K. (2003): An Overview of Data Mining Techniques: <http://www.thearling.com/dmwhite/index.htm>
 34. Two Crows Corporation (1999): Data Mining Glossary: <http://www.twocrows.com/glossary.htm>
 35. Two Crows Corporation (1999): Introduction To Data Mining And Knowledge Discovery
<http://www.twocrows.com>
 36. Zulfiqar A. Khan (2003): Cement raw materials: Production and quality control in cement industry seminar technical report, Qatar

ANNEX 1: license Agreement

ANNEX 2: Physical test result of milled cement form

Physical Test Result of Milled Cement

98

ANNEX 3: Chemical analysis form

Chemical Analysis

Month_____

Year_____

[illegible]

ANNEX 4: OPC and PPC test certificate forms

Mugher Cement Enterprise
Addis Ababa Cement Plant
Cement Quality Test Certificate

Type of cement: Portland Pozzolana Cement (PPC), Type IP

Cement Standard: ASTM C595

Date : _____

Production period: _____

I. CHEMICAL REQUIREMENTS

Property	Standard specification	Test results
Magnesium oxide (MgO), %	Max 5.0	
Sulfur trioxide (SO ₃), %	Max 4.0	
Loss on ignition (L.O.I), %	Max 5.0	

II. PHYSICAL REQUIREMENTS

Property	Standard specification	Test results
Time of setting, Vicat test		
a Initial set, Minutes	Min 45	
a Final set, Minutes	Max 120	
Compressive strength, N/mm ²		
a 3 days	Min 14.4	
a 7 days	Min 19.3	
a 28 days	Min 24.1	
Fineness, air permeability, cm ² /gH	Min 280	
Soundness, mm	Max 10	

Comment: _____

Signature: Quality control head Manager

Quality control head

Manager

Mugher Cement Enterprise
Addis Ababa Cement Plant
Cement Quality Test Certificate

Type of cement: Ordinary Portland Cement (OPC), Type I

Cement Standard: ASTM C150

Date: _____

Production period:

I. CHEMICAL REQUIREMENTS

Property	Standard specification	Test results
Magnesium oxide (MgO), %	Max 6.0	
Sulfur trioxide (SO ₃), %	Max 3.0	
Loss on ignition (LOI), %	Max 3.0	
Insoluble residue (I.R.), %	Max 0.75	

II. PHYSICAL REQUIREMENTS

Property	Standard specification	Test results
Time of setting, Vicat test		
a Initial set, Minutes	Min 45	
a Final set, Minutes	Max 375	
Compressive strength, N/mm ²		
a 3 days	Min 12.4	
a 7 days	Min 19.3	
a 28 days	Min 24.1	
Fineness, air permeability, m ² /g ^h	Min 280	
Soundness, mm	Max 10	

Comment: _____

Signature: Quality control head Manager

Quality control head

Manager

DECLARATION

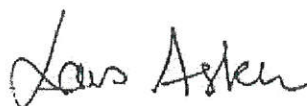
This thesis is my original work, has not been presented for a degree in any other university and all sources of material used for the thesis have been duly acknowledged.



Abenet Tale Workenh

July 2005

THE THESIS HAS BEEN SUBMITTED FOR EXAMINATION WITH MY APPROVAL AS UNIVERSITY ADVISOR



Dr. Lars Asker

Stockholm University

July 2005

