FACTORS AFFECTING THE USE OF DATA MINING IN MOZAMBIQUE: Towards a framework to facilitate the use of data mining

Constantino Sotomane
DSV Report Series No. 14-012
Factors affecting the use of data mining in Mozambique:

Towards a framework to facilitate the use of data mining

Constantino Sotomane
To my beloved wife Paula, daughters Angélica and Michela and mother Angélica.
Abstract

Advances in technology have enabled organizations to collect a variety of data at high speed and provided the capacity to store them. As a result the amount of data available is increasing daily at a rapid rate. The data stored in organizations hold important information to improve decision making and gain competitive advantage. To extract useful information from these huge amounts of data, special techniques such as data mining are required. Data mining is a technique capable of extracting useful knowledge from vast amounts of data. The successful application of data mining in organizations depends on several factors that may vary in relation to the environment. In Mozambique, these factors have never been studied. The study of the factors affecting the use of data mining is important to determine which aspects require special attention for the success of the application of data mining.

This thesis presents a study of the level of awareness and use of data mining in Mozambique and the factors affecting its use. It is a step towards the development of a framework to facilitate the application of data mining in Mozambique. The study is exploratory and uses multiple case studies in two institutions in Maputo city, the capital of Mozambique, one in the area of agriculture and the other in the field of electricity, and of Maputo city more broadly. The study involved a combination of observations, focus group discussions and enquiries directed at managers and practitioners on aspects of information technology (IT) and data analysis.

The results of the study reveal that the level of awareness and use of data mining in Mozambique is still very weak. Only a limited number of professionals in IT are aware of the concept or its uses. The main factors affecting the use of data mining in Mozambique are: the quality, availability and integration of, access to data, skill in data mining, functional integration, alignment of IT and business, interdisciplinary learning, existence of champions, commitment of top management, existence of change management, privacy, cost and the availability of technology. Three applications were developed in two real settings, which showed that there are problems to be solved with data mining. The two examples in the area of electricity demonstrate how data mining is used to develop models to forecast electricity consumption and how they can enhance the estimation of electricity to be sold to the international market. The application in the area of agriculture extracts associations between the characteristics of small farmers and the yield of maize from a socioeconomic database with
hundreds of attributes. The applications provide practical examples of how data mining can help to discover patterns that can lead to the development of more accurate models and find interesting associations between variables in the dataset. The factors identified in this thesis can be used to determine the feasibility of the implementation of data mining projects and ensure its success.
Acknowledgements

First of all, I wish to thank my supervisors, Associate Professor Lars Asker, Professor Henrik Boström and Associate Professor Venâncio Massingue, for their motivation, guidance and patience, without which I would not have reached this stage. I also thank Associate Professor Panagiotis Papapetrou for useful feedback during the pre-licentiate seminar that contributed to the improvement of the thesis.

My study was financed by the SIDA programme at UEM. I would like to thank UEM for providing the scholarship. I thank the team of this programme for their hard work that made access to the funds for my study possible. Among others, I thank Associate Professor Bhanji (former coordinator), Associate Professor João Paulo, Associate Professor Emilio Mosse, Professor Leif Abrahamsson, Mrs. Zsuzsanna Kristoffi and Mrs. Pravina Gajjar. When SIDA’s funding ended, I was awarded funds from the World Bank through the Ministry of Science and Technology (MCT). I want to thank all those who made this resource available, among others, Dr Jamisse Taimo, Dr Alsácia Atanásio, Mr Mahala, Mrs Verónica and Mrs Rocha.

After starting my studies, many changes happened. I changed jobs and several colleagues also changed jobs. Initially, Dr Américo Muchanga was director of CIUEM and as my boss, he allowed me to join the programme, for which I thank him. Dr Américo was replaced by Mr Francisco Mabila, who continued to support me in my studies. Thank you very much, Mr Mabila. I would like also to thank my colleagues from MICTI for helping me with my tasks when I was unable to fulfil them because of my studies. I extend my gratitude to the Rector and Vice-Rectors of UEM for authorizing my studies and providing the necessary resources.

I thank Professor Venâncio Massingue who, as Minister of Science and Technology, agreed that I could combine my duties at the ministry and my studies. I also thank Professor Louis Pelembe, the current Minister of Science and Technology, for releasing me from several tasks to provide more time for my studies. I also thank the Permanent Secretary of MCT. I would like to thank my colleagues at the Directorate of Infrastructure and Information Systems for their patience and helping me with my work when I was unavailable. Dr Gertrude, Ludmila, Jamo, Benjamim, Ana, Celestina, Rebelo, Zacarias and Ferro, thank you very much. I would also like to thank my colleagues from the office of the implementation of Science and
Technology Park of Maluana: several times, I left the project at a critical stage to travel for my research. Thank you, Zefanias, Tamele, Jamo, Rosita, Tania, Benedeto and Aida. I owe gratitude also to the CEO of the Science and Technology Park Company, Associate Professor Jorge Nhambiu, who accepted my long absence to continue with my studies, even in situations in which there was too much work and too few human resources.

At DSV, I have been helped by several people. I thank all the staff of the DSV for helping me and for being patient: Fatima, Rodolfo, Brigitta, Eija and Sören, to name just a few. Many thanks to Professor Love Ekenberg for introducing me to DSV almost a decade ago and for encouraging me to join. Despite his heavy agenda, Professor Ekenberg has been always available.

Thank you Dr Ravindra Abhyankar and Mr V.P. Sanjei for finding time in your busy schedules to read and comment on the draft of this thesis. I thank Dr Jordi-Gallego for being my co-author in one of the papers and for always being available to discuss my work.

The PhD gave me opportunity to interact with several people. On this journey, I had the privilege to share ideas about studies, life, cultures, traditions, religions and challenges and to collaborate. Thank you, Thashmee for being very kind and helpful. You spent a lot of time discussing my work with me. Thank you, Rasika, Mturi, Karokola, Nfuka, Elli, Florence, Meshari, Javier and Ruben for your friendship and for sharing your cultures and experiences with me. It was very important to have colleagues from Mozambique. Thank you very much, Xavier, Orlando, Avelino, Juvêncio, Eugênio, Condo, Lucílio, Cuvilas and Agnélio for keeping the Mozambican spirit alive all the time.

I thank my family, in particular my mother and my late father, for all their education and love.

Almost last, but certainly not the least, I thank my wife Paula and my daughters Angélica and Michela for allowing my long absence from home and for their unconditional love. Your love kept me standing in all the battles I faced. Please give me a little more time: I still need to cover the last mile.

Thank you, God, for your guidance and protection.
Contents

Chapter 1: Introduction ................................................................. 1
  1.1 Motivation and problem area ............................................. 3
  1.2 Research questions .............................................................. 4
  1.3 Objectives ........................................................................ 5
      General objective .................................................................. 5
      Specific objectives: .............................................................. 5
  1.4 List of included publications .............................................. 5
  1.5 Organization of the thesis ................................................. 7

Chapter 2: Overview of Socioeconomic and ICT Development in Mozambique................................................................. 9
  2.1 Overview of geographic, political and socioeconomic situation .... 9
  2.2 The role and status of ICT ................................................ 10

Chapter 3: The concept of data mining ............................................. 12
  3.1 Data mining methods ........................................................ 15
      Clustering ........................................................................ 15
      Classification ..................................................................... 18
      Association rules ............................................................... 19
      Numerical prediction ......................................................... 21
  3.2 Evaluation of data-mining methods .................................... 23
      Evaluating the results of clustering ................................... 23
      Evaluating the output of classification ............................... 23
      Evaluating association rules .............................................. 24
      Evaluating numerical prediction ...................................... 24
  3.3 Data mining in organizations .............................................. 24

Chapter 4: Research Methodology .................................................. 25
  4.1 Philosophical assumptions, approach, strategy and methodological choices......................................................... 25
4.2 Selection of data-mining cases and samples ........................................... 26
4.3 Data collection and processing and analytic techniques ....................... 27
   Study 1: ICT for automated forecasting of electrical power consumption:
   A case study in Maputo ........................................................................ 27
   Study 2: Extracting patterns from a socioeconomic database to
   characterize small farmers with high and low corn yields in Mozambique:
   A data-mining approach ...................................................................... 27
   Study 3: Factors affecting the use of data mining in Mozambique............ 27
   Study 4: Short-term load forecasting of electricity consumption in
   Maputo .................................................................................................. 28
4.4 Specific methods ................................................................................ 28
4.5 Research development process ............................................................. 28
4.6 Ethical issues ..................................................................................... 29
4.7 Soundness of the research .................................................................. 29

Chapter 5: Contributions ........................................................................... 31
5.1 Overview of the contributions .............................................................. 31
5.2 Study 1: ICT for the automated forecasting of electrical power
   consumption: A case study of Maputo ................................................ 31
   Lessons from Study 1 ......................................................................... 36
5.3 Study 2: Extracting patterns from a socioeconomic database to
   characterize small farmers with high and low corn yields in Mozambique:
   a data mining approach ..................................................................... 38
   Lessons from Study 2 ......................................................................... 39
5.4 Study 3: Factors affecting the use of data mining in Mozambique ...... 42
5.5 Study 4: Short-term load-forecasting model for Maputo ....................... 42
   Lessons from Study 3 ......................................................................... 45

Chapter 6: Conclusions and Future Directions ............................................ 47
6.1 Conclusions ........................................................................................ 47
6.2 Future research directions ................................................................. 48

References ............................................................................................... 50

Appendix A: Paper I
Appendix B: Paper II
Appendix C: Paper III
Appendix D: Paper IV
## List of Figures

<table>
<thead>
<tr>
<th>Figure</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>Figure 1</td>
<td>Structure of the thesis</td>
<td>8</td>
</tr>
<tr>
<td>Figure 2</td>
<td>Map of Mozambique, bordering countries and the Indian Ocean</td>
<td>9</td>
</tr>
<tr>
<td>Figure 3</td>
<td>Overview of steps in the KDD process following Fayyad et al</td>
<td>12</td>
</tr>
<tr>
<td>Figure 4</td>
<td>Overview of steps in the KDD process following CRISP-DM methodology</td>
<td>13</td>
</tr>
<tr>
<td>Figure 5</td>
<td>Example of clustering of the Iris dataset related to the type of plants</td>
<td>17</td>
</tr>
<tr>
<td>Figure 6</td>
<td>Example of a decision tree classifying plants</td>
<td>18</td>
</tr>
<tr>
<td>Figure 7</td>
<td>Result of the hotspot algorithm for the analysis of labour negotiation data</td>
<td>21</td>
</tr>
<tr>
<td>Figure 8</td>
<td>Example of the model for the prediction of electricity consumption</td>
<td>23</td>
</tr>
<tr>
<td>Figure 9</td>
<td>Common approaches, philosophical assumptions, strategies and choices</td>
<td>25</td>
</tr>
<tr>
<td>Figure 10</td>
<td>Research development process</td>
<td>29</td>
</tr>
<tr>
<td>Figure 11</td>
<td>Three-dimensional view of electricity consumption in Maputo</td>
<td>32</td>
</tr>
<tr>
<td>Figure 12</td>
<td>Long-term view of electricity consumption in Maputo</td>
<td>32</td>
</tr>
<tr>
<td>Figure 13</td>
<td>Comparison of temperature and electricity consumption in Maputo</td>
<td>33</td>
</tr>
<tr>
<td>Figure 14</td>
<td>Comparison of trends in GDP per capita and electricity consumption</td>
<td>33</td>
</tr>
<tr>
<td>Figure 15</td>
<td>Clustering of electricity consumption data for Maputo city</td>
<td>34</td>
</tr>
<tr>
<td>Figure 16</td>
<td>Comparison of forecasts and real electricity consumption over five summer working days</td>
<td>35</td>
</tr>
<tr>
<td>Figure 17</td>
<td>Comparison of forecasts and real electricity consumption over six winter working days</td>
<td>35</td>
</tr>
<tr>
<td>Figure 18</td>
<td>Comparison of forecasts and real electricity consumption over six Saturdays</td>
<td>36</td>
</tr>
<tr>
<td>Figure 19</td>
<td>Comparison of forecasts and real electricity consumption of five Sundays</td>
<td>36</td>
</tr>
<tr>
<td>Figure 20</td>
<td>Association rules tree for small farmers’ households with high yield of maize</td>
<td>39</td>
</tr>
<tr>
<td>Figure 21</td>
<td>Association rules tree for small farmers’ households with low yields of maize</td>
<td>41</td>
</tr>
<tr>
<td>Figure 22</td>
<td>Comparison of the results of the new model developed (model</td>
<td>43</td>
</tr>
</tbody>
</table>
Figure 23  Comparison of underestimated electricity using the EDM model and the newly-developed model

Figure 24  Comparison of overestimated electricity using the EDM model and the newly-developed model
List of Tables

Table 1  Mapping of Fayyad and CRISP-DM KDD process models  15
Table 2  Lessons learned during the Study 1 regarding: Awareness, use and factors affecting the use of data mining  37
Table 3  JRip decision rules characterizing small farmers household with high yield of maize  38
Table 4  Lessons learned during the Study 2 regarding: Awareness, use and factors affecting the use of data mining  40
Table 5  Factors affecting the use of data mining  42
Table 6  Comparison of mean absolute error (MAE) of Model EDM and Model EDM  43
Table 7  Comparison of financial implications between Model RR  45
Table 8  Lessons learned during the Study 5 regarding: Awareness, use and factors affecting the use of data mining  46
Abbreviations

<table>
<thead>
<tr>
<th>TERM</th>
<th>DESCRIPTION</th>
</tr>
</thead>
<tbody>
<tr>
<td>AGNES</td>
<td>AGlomerative NESting</td>
</tr>
<tr>
<td>CEO</td>
<td>Chief Executive Officer</td>
</tr>
<tr>
<td>CIUEM</td>
<td>Informatics Centre of Eduardo Mondlane University</td>
</tr>
<tr>
<td>CRISP-DM</td>
<td>CRoss-Industry Standard Process for Data Mining</td>
</tr>
<tr>
<td>CRM</td>
<td>Customer Relationship Management</td>
</tr>
<tr>
<td>CSF</td>
<td>Critical Success Factor</td>
</tr>
<tr>
<td>DAM</td>
<td>Day Ahead Market</td>
</tr>
<tr>
<td>DIANA</td>
<td>Divisive ANAlysis</td>
</tr>
<tr>
<td>DSV</td>
<td>Department of Computer and Systems Sciences</td>
</tr>
<tr>
<td>EDM</td>
<td>Electricidade de Moçambique</td>
</tr>
<tr>
<td>e-Government</td>
<td>Electronic Government</td>
</tr>
<tr>
<td>EM</td>
<td>Expectation-Maximization</td>
</tr>
<tr>
<td>e-Mail</td>
<td>Electronic Mail</td>
</tr>
<tr>
<td>GDP</td>
<td>Gross Domestic Product</td>
</tr>
<tr>
<td>GPS</td>
<td>Global Positioning Systems</td>
</tr>
<tr>
<td>HDI</td>
<td>Human Development Index</td>
</tr>
<tr>
<td>ICT</td>
<td>Information and Communication Technology</td>
</tr>
<tr>
<td>INAM</td>
<td>National Institute of Metrology</td>
</tr>
<tr>
<td>IRT</td>
<td>Iron Triangle</td>
</tr>
<tr>
<td>IT</td>
<td>Information Technology</td>
</tr>
<tr>
<td>KDD</td>
<td>Knowledge Discovery in Databases</td>
</tr>
<tr>
<td>MAE</td>
<td>Mean Abslolute error</td>
</tr>
<tr>
<td>MAPE</td>
<td>Mean Absolute Percent error</td>
</tr>
<tr>
<td>MCT</td>
<td>Ministry of Science and Technology</td>
</tr>
<tr>
<td>RFID</td>
<td>Radio Frequency Identification</td>
</tr>
<tr>
<td>RR</td>
<td>Robust Regression</td>
</tr>
<tr>
<td>SAPP</td>
<td>Sothern African Power Pool</td>
</tr>
<tr>
<td>SIDA</td>
<td>Swedish International development Cooperation Agency</td>
</tr>
<tr>
<td>SISTAFE</td>
<td>State financial management system</td>
</tr>
<tr>
<td>STL_F</td>
<td>Short-Term Load Forecasting</td>
</tr>
<tr>
<td>SVM</td>
<td>Support Vector Machine</td>
</tr>
<tr>
<td>TAM3</td>
<td>Technology Adoption Model 3</td>
</tr>
<tr>
<td>TIA</td>
<td>Trabalho de Inquérito Agrário (Agricultural Survey)</td>
</tr>
<tr>
<td>TSR</td>
<td>The Square Route Framework</td>
</tr>
<tr>
<td>UEM</td>
<td>Eduardo Mondlame University</td>
</tr>
</tbody>
</table>
Information technology (IT) is an instrument that can provide competitive advantage and transform organizations and people [2]. Today, the role of IT is not only to enable the automation of a large number of batch transactions as it used to be; it also serves to provide accurate and timely information for the support of tactical and strategic decision making in organizations. This development has led to the birth of decision support systems and techniques, such as data mining, which are becoming vital to the success of many organizations [3]. With the advances in IT, organizations have acquired the capacity to collect a variety of data at high speed and to store large volumes of data. In terms of variety, in addition to traditional structured data, organizations today collect and maintain raw, semi-structured and unstructured data [4].

The availability of the large volume and variety of data is an opportunity for organizations to gain insights into improving decision-making and competitiveness if they employ appropriate techniques, such as data mining [5, 6]. Data mining is the process of extracting knowledge from large volumes of data [7]. The results of data mining can be used immediately in the decision-making process or can serve to develop hypotheses for further validation [8]. Due to its ability to extract patterns hidden in large amounts of data, data mining is applied in almost all areas of activity; some examples include marketing, banking, finance, manufacturing, health care and agriculture [9].

The application of data mining in organizations is not a straightforward process. It is influenced by several factors, i.e. a limited number of conditions that have a direct and serious impact on the success of projects [10]. A data mining project is successful when it has been implemented and satisfied the business objectives for which it was intended. Thus, it is important to study success factors to determine when and under what conditions data mining can be implemented in organizations.

There are several existing studies related to factors affecting the use of data mining. Hermiz [11] identified four critical success factors in data mining projects, namely: a clearly articulated business problem, sufficient and good quality data, the recognition of data mining as a process with different components and dependencies, and their appropriate management. Hart [9, 12] conducted three studies over a five-year period in South Africa and identified several factors that he grouped into five categories: business issues, organizational issues, resource issues, operational issues and technological issues. Eun-Jeong Cho [13] studied factors affecting data mining in Korea and identified data quality, the understanding of data, data mining outsourcing strategies, process documentation and the request for data mining by the user. Hilbert [14] developed a framework for data mining implementation and tested it using data from German organizations; he found that the following factors affect the use of data mining: the commitment of top management, the existence of change management, the existence of a
budget for the project, the integration of the data mining process with the IT landscape and quality of data. Barko and Nemati [15] used the square route (TSR) framework to study factors affecting the use of data mining. To test the model, they surveyed experts using the internet. The result of this study was the identification of the following factors: quality of data, data integration, technological integration and expertise, outsourcing strategy, level of end user expertise, timeliness, resources and functionalities. Yun Chen [16] identified three factors negatively affecting data mining in China: scarcity of expertise in data mining, lack of data warehousing and limited knowledge of data mining. Sim [17], in his PhD dissertation, found that factors affecting the successful implementation of data mining are related to data issues, i.e. quality of data, integration of data, sufficiency of data, scalability and maintenance of data. Bole et al. [18], based on a literature review and interviews with experts, determined the following factors for data mining success: stakeholder support, quality of data, interpersonal skills, interdisciplinary learning, a focus on solving problems, the presence of a business champion and change management. Nie et al. [19] undertook a study of factors affecting data mining based on decision analysis and identified the following factors: quality of data, human factors, financial factors and the support of the executives. Finally, Dahlan [20] identified the clarity of business, user skills and experience, a data-driven culture and the quality of data as factors determining readiness to adopt data mining technology in the telecommunications industry in Malaysia.

Some of the aforementioned studies present general factors [11, 15, 17-19] and others present factors related to specific countries [9, 12-14, 20, 21]. Among these studies, no equivalent lists of factors can be discerned, although there are similarities between some of them, such as data issues and skills. Moreover, the list of factors derived from the literature is long and not very adequate for the purposes of this study given that the interest is to identify the limited number of factors that are most influential and require special attention for the success of data mining. Thus, it is proposed in this study to identify factors affecting the use of data mining in Mozambique as there is no previous study related to this country.

The application of data mining in Mozambique is in its very early stages [22], which increases the probability of failure during and after the implementation if adequate measures are not taken to ensure its success. Therefore, determining the factors that affect success is important to develop mechanisms to enhance the probability of success of data mining projects.

The four studies that compose the thesis were conducted using case studies in two institutions in Maputo city and the use of data mining in the city more broadly. The two institutions were observed and the case studies of Maputo city were undertaken using a combination of focus group discussions and questionnaires.
1.1 Motivation and problem area

The main motivation for applying data mining is the availability of huge amounts of data and the need to transform this data into useful knowledge [7]. To obtain advantages from the data, several organizations are adopting data mining. In Mozambique, as in other parts of the world, the amount of data collected by organizations is increasing rapidly [3, 6]. There are two main reasons for this rapid increase:

- **Technological development**: with the advance of technology, the capacity to collect and store large amounts of data inexpensively in organizations and at the individual level has increased.

- **The awareness of the importance of data**: with the increase in the level of awareness of the importance of data as an element of competitive advantage, organizations are adopting mechanisms for data collection, such as periodical censuses and surveys, distributing forms for data collection, computerizing the process, employing the internet and intranet, using satellite and other remote sensing equipment and mobile devices for data collection.

There are several examples indicating the growth in activities and capacities of data gathering in organizations, including at the government level:

- The National Institute of Statistics of Mozambique (INE) is working in coordination with several institutions to enhance the collection of sectorial data.

- The e-Government Network is now in effect and is creating several electronic public services. Progressively the e-Government Network infrastructure is being extended to more remote areas, such as districts and localities.

- The Ministry of Agriculture of Mozambique applies the National Agriculture Survey (TIA) [23] every year to collect information concerning crops, livestock, production, farms, infrastructure, occurrences of natural disasters (flood, pest, drought, etc.), socioeconomic data on households, commercialization and other information.

- The Ministry of Health collects and maintains data on diseases, patients, outbreaks, births, deaths and other information.

- The Government of Mozambique is implementing the Government Data Centre.

- The National Institute of Meteorology of Mozambique (INAM) collects information on the weather at the national level.

- The Ministry of Finance of Mozambique runs the Financial Management System of the State (SISTAFE), which executes all the financial transactions of Government institutions.

- As is happening worldwide, the use of mobile technology and its applications are increasing, becoming one of the main data generation tools, in particular in relation to the personal location [2, 6].
These data collected from internal and external sources are increasing daily and are kept by organizations in different formats: in hard copy, or in files or relational databases. Usually, a small part of structured data are analysed using statistical techniques for reporting or research purposes, while the other parts of the data, including the unstructured and semi-structure data, tend not to be analysed. Information contained in unanalysed data remains unknown. The reason for not analysing all data is that the existing tools are based on traditional statistical techniques, which have limitations in terms of analysing large amounts of data and unstructured data. The use of data mining techniques extends the capacity of existing data analysis, making possible the analysis of large amounts of data.

Data mining is a new technique in Mozambique and its use still limited to a very small group of operators [22]. There are some theoretical applications in the area of agriculture [1], electricity [24, 25] and malaria [26], but no case of deployment was found during the literature review. As noted in the introduction, the successful implementation of data mining depends on several factors [14, 17, 18]. However, the literature review reveals that there are several lists of factors and no previous study has indicated which factors might pertain to Mozambique. The countries studied, such as Germany [14], China [16], Malaysia [20], Korea [13] and South Africa [9, 12] have different socioeconomic structures, which means that the results may not be applicable to Mozambique. This study aims to determine factors affecting the use of data mining in Mozambique with the objective of establishing which elements must be taken into consideration to ensure the success of data mining. The study also will determine the level of awareness and use of data mining. The level of awareness and use indicates how important data mining is to the stakeholder and possible attitudes to data mining. Stakeholders with greater awareness of data mining and who are active users tend to have positive attitudes, promoting the success of data mining.

There is limited literature on how data mining can be implemented successfully in organizations, in particular in countries such as Mozambique which experience several constraints in relation to data quality and availability as well as skills. Most of the available studies in the area of data mining are related to description of its advantages, the implementation process, the development of applications and the enhancement of algorithms. This study will contribute information on how to enhance the success of the application of data mining in organizations.

1.2 Research questions
To address the issue presented above, this thesis will answer the following questions:

Question 1: What is the level of awareness and adoption of data mining in Mozambique?

Question 2: What are the factors affecting the use of data mining in Mozambique?
1.3 Objectives

**General objective**
The general objective of this study is to identify factors affecting the use of data mining in Mozambique to enhance the success of its implementation.

**Specific objectives:**

a. To explore the perceptions of potential practitioners in data analysis, IT and managers in relation to the awareness, use and factors affecting the use of data mining.

b. To observe in two organizations the attitudes of organizational stakeholders’ that constitute enablers and impediments for the application of data mining.

c. To develop applications using the data existing in two organizations to gain an understanding through practice of issues related to the quality of data, the integration of data, access to data, the support of stakeholders and interest in the results of data mining.

1.4 List of included publications

This section presents the list of papers published in the proceedings of international conferences. All papers were peer-reviewed.

**Paper I**

This paper responds to research objectives b) and c) and identifies through observation the level of awareness and use of data mining and the factors affecting its use in a studied company. The paper also demonstrates that there are problems in the area of electricity that could be solved with data mining. Using data mining techniques, a short-term forecast model is developed. The observations made during the development of the model show that the quality of data is limited, the data are not integrated, there are problems in accessing some data, there are concerns with privacy, there is no use of data mining or data mining technology and there is low awareness of data mining.

The contributions of each author are as follows: Constantino Sotomane participated in the data collection, pre-processing, development of the model and development of the article. Constantino Sotomane also maintained continuous contact with the company experts for feedback on models and collection of additional information. Lars Asker participated in the data collection, development of the model and reviewed the article. Venâncio Massingue reviewed the article and gave feedback for its improvement.
Paper II

This paper responds to research objectives b) and c) and identifies through observation the level of awareness and use of data mining and factors affecting its use in a studied company. The paper also demonstrates that there are problems in the area of agriculture that could be solved with data mining. Using data mining techniques, association rules between the socioeconomic characteristics of small farmers and yields of maize were extracted. The observations identified that the quality of data is limited, the agricultural database is rich in information on several agricultural issues but there is no integration of information from other sectors, there are problems accessing some data, there are concerns with privacy, there is no use of data mining or data mining technology and there is low awareness of data mining.

The contributions of each author are as follows: Constantino Sotomane participated in the data collection, pre-processing, development of the model and development of the report. Constantino Sotomane also maintained contact with the company experts for feedback on the models and collection of additional information. Jordi Gallego-Ayala analysed the results of the data mining from an agricultural perspective and contributed to the discussion of the paper. Lars Asker and Henrik Boström gave supervision during the experiments and development of the article and reviewed the article. Venâncio Massingue reviewed the paper and contributed with aspects related to the agriculture’s development policies in Mozambique.

Paper III

This paper responds to research objective a) and describes the results of focus group discussions and inquiries that explored the perceptions of potential practitioners and informed respondents concerning their awareness and use of data mining and factors affecting its use in Mozambique.

The contributions of each author are as follows: Constantino Sotomane designed the data collection instruments, conducted the data collection process which involved a survey, the interviews and focus group discussion. He also analyzed the data and developed the article. Lars Asker, Henrik Boström and Venâncio Massingue gave the supervision during the development of the work and corrected the reports.
Paper IV

This paper responds to research objectives b) and c) and identifies through observation the level of awareness and use of data mining and factors affecting its use in a studied company. The paper also demonstrates that there are problems in the field of the electricity market that could be solved with data mining. Using data mining techniques, a short-term forecast model is developed and used to determine the amount of energy to be sold to the international market. The observations demonstrate that most of the factors identified in the first paper still exist. However, more data were available for analysis with data mining.

The contributions of each author are as follows: Constantino Sotomane collected the data, realized the pre-processing, modeling and the developed the article. Henrik Boström, Lars Asker and Venâncio Massingue gave the supervision during the development of the work and corrected the article.

1.5 Organization of the thesis
The thesis is organized in six chapters as follows. Chapter 2 provides an overview of socioeconomic and IT development in Mozambique. Chapter 3 presents the concept and applications of data mining, as well as describing the techniques and tasks of data mining used in the studies included in this thesis. Chapter 4 presents the research methodology and addresses philosophical assumptions, the research approach and strategy and choices of methods adopted in the study, as well as the selection criteria for data mining applications and samples used in the studies. Chapter 4 also describes the specific data mining methods used in the thesis, the research process and ethical aspects. Chapter 5 presents the main contributions of the research. Finally, Chapter 6 presents the conclusions and suggests future directions for research. Figure 1 summarizes the organization of the thesis.
Figure 1. Structure of the Thesis

Chapter 1: Introduction
- List of publications
- Research Questions
- Objectives
- Area of Motivation and Problem

Chapter 2: Overview of Mozambique
- The role and status of ICT
- Socio-economic situation
- Overview of Geographic, Political and

Chapter 3: Concept of Data Mining
- Data Mining in Organizations
- Data Mining Methods

Chapter 4: Research Methods
- Ethical issues
- Techniques
- Data collection, processing and analysis
- Selection of data mining applications
- Strategy and choices
- Philosphophical assumptions, approach,

Chapter 5: Contributions
- Overview of Contributions
- Analysis of contributions per study
- Overview of Contributions

Chapter 6: Conclusions
- Future Works
- Conclusions
- Conclusions

Future works
Chapter 2: Overview of Socioeconomic and ICT Development in Mozambique

2.1 Overview of geographic, political and socioeconomic situation

Mozambique is located in Southern Africa and borders the Indian Ocean, Tanzania, Malawi, Zambia, Zimbabwe, Swaziland and South Africa (see Figure 2). It is a young country, gaining its independence in 1975 after 10 years of struggle [27]. At the time of obtaining independence, the country inherited several problems, in particular the scarcity of qualified human resources and undeveloped infrastructure [27]. Two years after independence, a destabilizing civil war began that lasted 16 years and destroyed almost the entire basic infrastructure and killed millions of people [27]. The rebuilding of the infrastructure and socioeconomic system took several years and in some areas is still continuing.

Figure 2: Map of Mozambique, bordering countries and the Indian Ocean
Mozambique has an area of 799,390 km$^2$ and a population of 23 million [28]. It is endowed with rich and extensive natural resources [29]. The country’s economy is based largely on agriculture [30], but with an extractive industry that has been growing in the last few years due to the discovery of several natural resources such as gas and coal [31] [32]. The growth in the mineral sectors is contributing to the growth of foreign direct investment, which might accelerate the socioeconomic development of the country [31]. Mozambique has been making significant progress in socioeconomic development [30]; for example, in the period 2001–2011 the average growth rate of GDP per capita was 7.9% [28], which is among the highest in the world.

Despite its enormous natural resources and high growth in GDP, the country is still one of the least developed in the world. The human development report of 2013 ranks Mozambique as 185 out of a total of 187 countries with a human development index (HDI) of 0.327 [33]; this indicates that significant work still needs to be done for the development of the country. The country is also vulnerable to natural disasters, in particular floods, droughts and tropical depressions, which hit the country almost every year affecting its development and nullifying the effects of investment [34].

ICT is seen as one of the levers of development of the country [2, 35]. Mobile technology, in particular, constitutes a considerable opportunity for economic development and the promotion of entrepreneurship due to its rapid and wide coverage in the country. Similar to other African countries, more people have access to mobile phone than to basic infrastructure and services, such as water, electricity, financial banking and medical services [2]. This opens up the opportunity to provide innovative services to this group of the population and generate revenue. A report on the transformation in the use of ICT in Africa [2] indicates that ICT has contributed 7% of GDP growth in Africa in the last few years. The next section summarizes the status and the role of ICT in Mozambique.

2.2 The role and status of ICT

The development of the country is the primary agenda in Mozambique and ICT is considered one of the tools that can leverage development [35, 36]. In 2000, the government approved ICT policy and in 2002 the respective implementation strategy, which defined thirty-seven projects in the following areas: development of human capacity, infrastructure, content and applications, governance, policy and regulation, entrepreneurship development and development of provinces [37]. These projects generated several other programmes that increased the use of ICT at organizational and individual levels. An example of this is the implementation of the e-Government project, which generated the following results: the national broadband backbone connects all capitals of provinces and districts [38]; the electronic government network links government institutions up to the provincial level; at least hundred and forty government agencies share internet access; the government service portal provides several public services and information to citizens; the government has centralized the e-mail system,

---

1 Calculated by the author based on the projection of the population for 2013.
providing more than ten thousand public officers and government leaders with e-mail; the government website hosts more than forty websites of government agencies. As a result of the implementation of the e-government project, the following applications are hosted by government: land management, criminal register, driving licensing, mining register, business entity register. These applications are installed at the central level and can be accessed at the provincial level for the provision of services to citizens [39]. The state financial administration system (SISTAFE) and the single window system are also in operation. There are examples of several other developments in the area of ICT in several other sectors in addition to the thirty-seven projects included in the ICT strategy, e.g. the implementation of a payroll system in the central bank [40], the use of computers to input data in the field during agricultural surveys [23], the computerization of the banking system and introduction of automated teller machines [36], and the computerization of the voting registration system from 2008 [36]. Many more examples of the implementation and use of ICT in Mozambique can be given.

As in other African countries, the mobile telecommunications sector is growing rapidly. Mozambique has three mobile operators and approximately eleven million subscribers, covering all the capitals of districts [38]. The widespread access to mobile technology, combined with the continuous expansion of smartphone use, constitutes a source of growth in personal location data [6].

Globalization, increased competition in business and the attraction of large multinational companies due to the natural resources of the country is affecting the rapid growth of the ICT sector in the country and consequently the amount of data stored in organizations.
Chapter 3: The concept of data mining

Data mining is the process of extracting knowledge from large amounts of data [7]. The term is synonymous with knowledge discovery in databases (KDD) which, according to Fayyad et al. [41], is a non-trivial process of identifying valid, novel, potentially useful and ultimately understandable patterns in data. It is the entire process of data access, data exploration, data preparation, modelling, model deployment, and model evaluation [41, 42]. Figures 3 and 4 present two examples of KDD process models. The term “data mining” is used interchangeably both as synonym for KDD and as a step in the KDD process, in which algorithms are applied to extract patterns [5].

![Diagram of KDD process](image)

Figure 3: Overview of steps in the KDD process developed by Fayyad et al. [41]

The concept of KDD is relatively new and was originally discussed in the first workshop on KDD in 1991 [43]. Since then, several process models have been proposed in an attempt to standardize the KDD process and allow a means of solving practical business problems [43]. A survey undertaken by Lucasz [43] found twelve KDD process models proposed by several researchers. Fayyad et al.’s [41] model and the CRISP-DM model [44] are those more frequently referenced [43].

Fayyad et al.’s [41] model, presented in Figure 3, constitutes the first approach and is comprises the following nine iterative steps:

- **Step 1 – developing an understanding of the application domain and identifying the goals:** determining with the domain experts the problem to be solved through data mining, the necessary knowledge concerning the business that will allow understand of the problem and enable it to be solved.

- **Step 2 – creating a target dataset:** selecting and building the dataset on which the discovery is to be performed. This can involve collecting and integrating data from other sources.

- **Step 3 – cleaning and pre-processing the dataset:** performing operations to improve the quality of data by removing errors and noise, fixing missing values and dealing with outliers.
• **Step 4 – reducing and projecting the dataset:** transforming the data for data mining. This includes reducing dimensions, selecting features, transforming attributes and other operations. The transformation is undertaken in accordance with the goal.

• **Step 5 – matching the goals of the KDD process to a particular data mining method:** selecting the appropriate method for the data mining step. There are several methods, such as classification, clustering and regression, among others.

• **Step 6 – performing exploratory analysis and selecting models and hypotheses:** based on the choice in step 5, the algorithm that will facilitate the implementation of the goal is selected. For example a neural network can be selected rather than a decision tree to perform a classification task because it is more important to obtain good precision than a sound understanding of the phenomena.

• **Step 7 – data mining:** applying the algorithm to find patterns. The parameters of the algorithm can be adjusted to improve the quality of the patterns.

• **Step 8 – interpreting the mining patterns:** interpreting the meaning of patterns found in relation to the goal defined in step 1. The interpretation can be undertaken through the use of visualization techniques, using domain experts or other means. Steps 1 to 7 can be repeated if the goal is not fully met.

• **Step 9 – reporting and using the information discovered:** documenting the results and applying them to solve the problem identified in step 1.

The arrows in the model show that the process is iterative, i.e. the user can move forward and backward to refine the results of each step and the patterns discovered [41].

The CRISP-DM process model (Figure 4) is more industry-oriented as it was created by a consortium of companies – NCR Systems Engineering Copenhagen (USA and Denmark), DaimlerChrysler AG (Germany), SPSS Inc. (USA) and OHRA Verzekeringen en Bank Groep BV (The Netherlands) [43] – to respond to business needs.

![Figure 4: Overview of steps in KDD process using CRISP-DM methodology [42, 44]](image-url)
The CRISP-DM process model comprises the following six iterative steps [42, 44]:

- **Step 1 – business understanding**: understanding the data mining project in terms of the business perspective and based on this, defining the problem and planning the data mining solution.
- **Step 2 – data understanding**: exploring the data to understand and identify data quality problems and gain insights into the data that can help in the data mining process.
- **Step 3 – data preparation**: building the final dataset for modelling. This includes cleaning, selecting attributes and undertaking all other necessary transformations.
- **Step 4 – modelling**: selecting and applying data mining algorithms. This step iterates with step 3 in the process of experimenting with different algorithms as different algorithms can require different formatting of data.
- **Step 5 – evaluation**: reviewing the model to ensure that it satisfies all the business objectives defined in step 1. This step determines the use of the results of the data mining in the organization.
- **Step 6 – deployment**: applying the model within the business. Depending on the organization, the deployment can be implemented in different forms, e.g. the generation of a report, the implementation of repeatable processes of data mining across the organization, or other kinds of use.

To ensure the quality of the model or pattern to be discovered, there are iterations between several steps: business understanding and data understanding, data preparation and modelling. If the results of the evaluation are not satisfactory, the process returns to the initial step; otherwise, the process continues to the final step.

Table 1 maps the steps in Fayyad et al.’s and the CRISP-DM process models. It can be observed that some of the CRISP-DM model steps correspond to more than one of Fayyad et al.’s steps. Step 1 is similar for both process models: developing understanding of the domain area and setting the goal of the data mining project. Step 2 in both process models establishes the dataset for mining, but the CRISP-DM model also uses this step to understand the data. Steps 3 and 4 in Fayyad et al.’s model correspond to step 3 in the CRISP-DM process model in which the data are cleaned and transformed. Steps 5, 6 and 7 in Fayyad et al.’s model correspond to step 4 in the CRISP-DM model, which includes the selection of the data mining task, the selection of the algorithm and its application. Step 8 of Fayyad et al.’s model corresponds to step 5 of the CRISP-DM model and addresses the evaluation of the results. Step 9 of Fayyad et al.’s model corresponds to step 6 in the CRISP-DM model; both steps deal with the reporting and application of the results within the organization. In this thesis, Fayyad et al.’s process model is employed in Study 2 [1] and the CRISP-DM model in Studies 1 [25] and 4 [24].
3.1 Data mining methods

In this section, the term “data mining” refers to the step in the KDD process in which algorithms are applied to extract patterns [5, 7]. The goal of data mining can be verification or discovery [5, 41, 45]. Verification aims to confirm or falsify a hypothesis; it is also called top-down or deductive analysis and represents the classic form of analysing data [42]. On the other hand, discovery aims to find new patterns in the dataset. The pattern discovered can be used directly as new knowledge to improve decision making or as a new hypothesis for further confirmation [41, 45]. The discovery approach is also called bottom-up analysis or inductive analysis [42].

The discovery task is the most common in data mining and is subdivided into descriptive and predictive models [5, 41]. Descriptive models seek to find patterns that characterize the data [5, 41, 42]. Descriptive tasks include: clustering, summarization, developing association rules and sequence discovery [5, 42]. Predictive models seek to predict unknown or future values of a variable of interest in the presence of other variables. Predictive models include classification, regression, time series analyses and numerical prediction [41, 42]. There is no clear boundary between the predictive and descriptive models: some predictive models can be descriptive because they provide the characterization of the dataset and vice-versa [45]. An example of this is the decision tree classification algorithm which is able to predict and describe a given phenomenon.

In the three studies presented in this thesis, both predictive and descriptive models were employed. Study 1 [25] uses clustering, classification and regression models; Study 2 [1] uses classification and association rule models and Study 4 [24] uses clustering and regression models. The following sub-sections describe the algorithms used in the studies included in the thesis.

### Clustering

Clustering is the process of grouping objects or data based on similarities. Given a dataset $D$ with $n$ objects, the clustering algorithm finds $k$ group of objects ($k \leq n$), where each group represents a cluster [7]. Objects within a cluster are more similar than objects in different clusters and objects in different clusters are more different than objects in the
same cluster [5, 7]. There are several means of measuring similarity between objects depending on the type of data. Many clustering algorithms determine similarity based on Euclidian or Manhattan distance [7]. There are several clustering methods, some of the most popular of which are described below:

- **Partitioning**: given dataset \( D \) with \( n \) objects, partition clusters algorithms find \( k \) partitions \( (k \leq n) \), each partition representing a cluster. The similarity between objects is measured based on distance. Two well-known examples of the clustering partition method are K-means and k-medoids [7].

- **Hierarchical**: the clusters are formed using the bottom-up (agglomerative) or top-down (divisive) approach. The agglomerative approach starts with each cluster formed by one object and the most similar objects are successively merged into one cluster until the termination condition is satisfied or all objects are assigned to one cluster. The divisive approach starts with all objects assigned to one cluster and they are successively split into smaller clusters until the termination criteria are satisfied or each cluster is formed by one object. Examples of algorithms for agglomerative and divisive clustering are AGNES (AGlomerative NESting) and DIANA (Divisive ANAlysis) respectively [7].

- **Density-based**: the clusters are formed based on the density of the object. Areas with a high density of objects form one cluster and areas with sparse objects are considered noise or outliers. The density-based method can find clusters of arbitrary shapes. An example of a density-based clustering algorithm is DBSCAN [7].

- **Model-based**: each cluster is characterized using a mathematical model; this method assumes that the dataset is generated by a mixture of underlying probability distribution models. One example of a model-based clustering method is the expectation and maximization (EM) algorithm [7] summarized below, which is used in Studies 1 [25] and 4 [24].

Clustering is an aspect of descriptive data mining tasks and has several applications, such as characterizing the buying patterns of customers in market research, classifying documents on the Web and detecting outliers or noise in the data.

**The expectation maximization (EM) algorithm**

The EM clustering algorithm finds the most likely set of clusters for the dataset [42]. It is based on the concept of the finite mixture model, which considers that observations comprise a set of models of \( k \) probability distributions, each distribution representing a cluster [7, 42, 46].

Given a dataset \( D \) composed of two mixtures of normal probability distribution represented by Gaussian distribution \( g(\mu_1, \sigma_1) \) and \( g(\mu_2, \sigma_2) \) respectively, the EM algorithm can be summarized using the iterative steps below [7]:

- **Step 1**: randomly select the initial parameters: \( k \) objects representing a cluster centre \( (\mu_1, \ldots, \mu_k) \), standard deviation \( (\sigma_1, \ldots, \sigma_k) \) and other additional parameters.

- **Step 2**: iteratively refine the clusters based on following two steps:
  - **Expectation step**: assign each object \( x_i \) to cluster \( C_k \) with the probability:
This step calculates the probability of cluster membership of object \( x_i \), for each cluster. These probabilities are the expected cluster memberships for object \( x_i \).

- **Maximization step:** using the estimated cluster probability to re-estimate the model parameters. For example:

\[
\mu_k = \frac{1}{n} \sum_{i=1}^{n} \sum_{j} x_i p(x_i \in C_k) \]

Figure 5 is an example of a plot resulting from the clustering of an iris dataset using the WEKA workbench [47]. The dataset is comprises 150 instances, four attributes and a class label. The attributes are sepal length, sepal width, petal length and petal width, which characterize three types of class (type of plant): Iris setosa, Iris virginica and Iris versicolor. Before the clustering, the class label was removed and the parameter number of clusters in the EM algorithm was set to three. The results illustrate that the data were partitioned in three groups corresponding to each type of plant.

![Figure 5: Example of clustering of an iris dataset in relation to the types of plant](image)
Classification

Classification assigns an object to a certain predefined category or class based on its features or attributes \([5, 7]\). Let us take database \(D\) with \(T = \{t_1, t_2, ..., t_n\}\) tuples, where each tuple is composed of attributes \(A = \{A_1, A_2, ..., A_p\}\) and a class \(c_i\), where \(c_i \in C = \{c_1, c_2, ..., c_m\}\). The classification problem is to define the mapping \(f : D \rightarrow C\) where each \(t_j\) is assigned to one class \(c_i\) \([5]\).

The construction of a classifier has two phases:

- the learning phase in which the training data, with a known class label, are used to build the classification model;
- the test phase in which the accuracy of the classifier is measured.

The accuracy of the classifier is the percentage of the test data correctly classified. If the accuracy is satisfactory, the classifier is used to classify a new dataset without a class label.

There are several techniques and algorithms for classification, the most common of which include: decision tree, Bayesian classification, rule-based classification, neural networks, support vector machines and k-nearest neighbour. Detailed descriptions of each of these techniques can be found in the literature \([5, 7, 42, 46]\) and other publications related to machine learning and statistics.

To enhance the efficiency and scalability of the classifier, it is important to perform pre-processing of the data before training the algorithm. The pre-processing consists of cleaning the data to reduce error and noise, handling missing values, treating outliers, removing irrelevant attributes, generalizing or reducing dimensions, or undertaking other required transformations.

Figure 6: Example of a decision tree classifying plants
Classification relates to predictive data mining tasks and has several applications, such as in medical diagnosis, loan approval, pattern recognition and classifying financial market trends [5]. Figure 6 presents an example of the application of a decision tree classification algorithm [48], used to classify plants based on sepal and petal characteristics. This is discussed in further depth in the following sub-section.

**Decision tree**

The decision tree is one of the most popular classification techniques, the output of which is a tree-like structure. In the decision tree, the nodes represent the attributes, the edges the values of the attribute and the leaves the classes. The topmost node is called the root node and represents the best splitting attribute, which means this is the attribute that will generate partitions with a high degree of purity. A partition is pure when all the instances in it are of the same class [7, 49].

The decision tree algorithm splits the dataset into groups recursively, based on an attribute value test, until the termination criteria are reached. The termination criteria can be any of the following: all the instances belong to the same class; there are no remaining attributes in which the instances may further be partitioned; there are no instances for a given branch [7]. The splitting attribute is chosen based on splitting criteria which vary according to the algorithm; for example, the C4.5 algorithm [48] uses the information gain ratio [5, 7] and the CART algorithm [5, 7, 42] uses the Gini index [5, 7].

Figure 6, developed using data from the WEKA workbench [47] dataset, is an example of a decision tree which classifies three types of plants – Iris sesota, Iris versicolor and Iris virginica – based on four numeric attributes: sepal length, sepal width, petal length and petal width. In Figure 6, the attributes are represented by the ellipses and constitute the decision point; the values of the attributes are given beside the directional arrows; the classes are represented in rectangles. As can be seen from the tree in Figure 6, if the petal width is less than or equal to 0.6 cm the plant is Iris sesota, if the petal width is greater than 0.6 cm or greater than 1.7 cm, the plant is Iris virginica. The numbers in the classes show how many instances were assigned to the class; for example, 50 examples that were assigned to the class Iris sesota were correctly classified and 0 instances were incorrectly classified.

**Association rules**

The association rules is a data mining method used to discover interesting relationships in a large database [5, 7]. Given a database $D$ of transactions $T$, in which each transaction is composed of a set of items $I = \{I_1, I_2, ..., I_m\}$, the association rule is an implication of the form $X \Rightarrow Y$, where $X, Y \subseteq I$ and $X \cap Y = 0$. The strength of association rules is given by support and confidence [5, 7]. The support of an association rule $X \Rightarrow Y$ is the percentage of transactions in the database that contain $X \cup Y$ [5]. This can be represented by expression 3 [7] as follows:

$$Support(X \Rightarrow Y) = P(X \cup Y)$$

The confidence of association rules is the ratio of the number of transactions that contain $X \cup Y$ to the number of transactions that contain $X$ [5], as expressed by equation (4) [7]:

$$Confidence(X \Rightarrow Y) = \frac{P(X \cap Y)}{P(X)}$$
Good association rules are those that satisfy the minimum support threshold and the minimum confidence threshold, which are determined by the user or domain expert [7, 50]. The objective in association rule mining is to find rules that satisfy the minimum support and minimum confidence levels, i.e. strong rules.

Expressions (5) and (6) are examples of association rules. The association rule of expression (5) says that customers who buy computers tend to buy antivirus software at the same time and expression (6) indicates that customers who buy peanut butter tend to buy bread [5, 7]. Expression (5) has a support of 2% and a confidence of 60% and expression (6) has a support of 60% and a confidence of 100%. In general, the best rules have low support and high confidence.

\[
\text{Computer} \Rightarrow \text{antivirus software}[\text{Support} = 2\%, \text{confidence} = 60\%][7] \tag{5}
\]

\[
\text{PeanutButter} \Rightarrow \text{Bread}[\text{Support} = 60\%, \text{confidence} = 100\%][5] \tag{6}
\]

There are several algorithms for extracting association rules. The most popular is the Apriori algorithm [5, 7, 51]. In Study 2 [1] included in this thesis, an association rule algorithm called hotspot [52] is employed to extract the relationship between the level of productivity of maize (high or low) and the characteristics of small farmer households. The hotspot algorithm is summarized in the next sub-section.

Association rule mining can be applied in several areas. The area of market basket analysis is the most common; for example, the result of the two association rules (given by expressions (5) and (6)) can be used by marketing analysts to design a marketing campaign for targeted customers, or store managers can use the results to plan the placement of their products. Other areas of application of association rules include telecommunications, web mining, bioinformatics [5], agriculture [1], medicine [53] and insurance [52, 54].

The hotspot algorithm
The hotspot is an association rule algorithm which analyses an attribute of interest [53]. It performs analyses of a segment of a dataset [52]. In the example given in Figure 7, it is developed based on a labour negotiation dataset derived from the WEKA workbench [47], in which the interest lies in analysing reasons for bad contracts. Bad contracts are those that are not accepted by the employee or employer. The decision tree in Figure 7 generated by the hotspot algorithm shows four reasons for a bad contract, one of which is when the wage increase in the first year is less than 4% and the statutory holiday allocation per year is less than 12 days (see the top branch of Figure 7).

The hotspot algorithm uses a combination of greedy and depth-first search algorithms to construct the tree of rules. It starts with the entire dataset at the top and goes down using the depth-first search combined with the greedy (best-first) approach. This means that the hotspot method explores recursively using the depth-first search approach for all attributes starting with those that represent greater confidence in relation to the target value [53].
Figure 7: Results of the hotspot algorithm for the analysis of labour negotiation data [47]

**Numerical prediction**

Numerical prediction employs the function \( y = f(X) \) that models the relationship between the input value \( X \) and the output value \( y \), which must be continuous. When the value of \( y \) is categorical, the predictive task becomes classification. Similar to classification, the construction of predictive models encompasses two phases:

i. the learning phase in which the training data, with known output value \( y \), are used to build the predictive model;

ii. the test phase in which the test data, independent of the training data, are used to measure the accuracy of the predictive model.

In real life, there are many applications of numerical prediction. Some example include weather forecasting, electrical consumption forecasting, the forecasting of production values based on historical data and the forecasting of values that the costumer will spend based on the price of products. There are several techniques that can be employed for numerical prediction, such as neural networks, support vector machines (SVMs), moving average techniques and regressions. The last of these is discussed in the following subsection.

**Regression models**

Regression is one of the most popular techniques used in numeral prediction [7, 42]. Studies 1 [25] and 4 [24] in this thesis use regression models.

Given that input \( X^T = (X_1, X_2, ..., X_p) \), also called independent variables, with regression coefficients \( B_j = (\hat{B}_0, \hat{B}_1, ..., \hat{B}_j) \), the regression model estimates output \( Y \), also called the dependent variable using expression (7) below [46]:

\[
Y = f(x) = B_0 + \sum_{j=1}^{p} x_j B_j
\] (7)
A training set, composed of a set of inputs and outputs, is used to estimate the regression coefficient $B_j$. The most popular approach for this is the least squares method, which determines the coefficients $B_j = (B_0, B_1, ..., B_p)$, minimizing the residual sum of square (RSS) given by expression (8) below [46]:

$$\text{RSS} = \sum_{i=1}^{N} (y_i - f(x_i))^2$$  \hspace{1cm} (8)

$$\text{RSS} = \sum_{i=1}^{N} \left( y_i - B_0 - \sum_{j=1}^{p} x_i B_j \right)^2$$  \hspace{1cm} (9)

Using the matrix notation, expression (9) can be re-written as in expression (10):

$$\text{RSS}(B) = (y - XB)^T (y - XB)$$ \hspace{2cm} (10)

Differentiating expression (10) with respect to $B$, expressions (11) and (12) are derived:

$$\frac{\partial \text{RSS}}{\partial B} = -2X^T (y - XB)$$ \hspace{1cm} (11)

$$\frac{\partial^2 \text{RSS}}{\partial B \partial B^T} = -2X^T X$$ \hspace{1cm} (12)

Assuming that $X^T X$ is positive and setting the first derivate to zero in expression (11), expression (13) is obtained (for further detail please see Hastie et al. [46]):

$$X^T (y - XB) = 0$$ \hspace{1cm} (13)

which yields expression (14) to estimate the regression coefficients:

$$B = (X^T X)^{-1} X^T y$$ \hspace{1cm} (14)

In practice, the amount of data is large and software such as MS-Excel, Matlab and SPSS are used to build the regression models [5].

There are numerous applications of linear regression. In this thesis, linear regression is used in Study 1 [25] and Study 4 [24] to predict electricity consumption. Figure 8 is the result of the regression model for the prediction of electricity consumption for Maputo city. The real consumption is represented by the continuous line and the prediction is represented by the dashed line. Figure 8 shows that the line representing real consumption is very close to that representing predicted consumption and at some points they overlap, which means that the values predicted by the model approximate the real values.
Figure 8: Example results of a regression model for the prediction of electricity consumption

3.2 Evaluation of data-mining methods

This section describes how the results obtained using the application of each data-mining method were evaluated.

Evaluating the results of clustering

In general, it is difficult to evaluate clustering. The practical way is to verify whether the result is useful for the application [49]. This principle was followed in evaluating the results of clustering in Study 1 and Study 4. In Study 1, the predictive model was built using the result of running the EM clustering algorithm with 1, 2, 3, 4, 5 and 6 clusters. The prediction based on clustering with 3 clusters provided greater accuracy [25]. In this case, the result of 3 clusters is more useful. In Study 2, the 3 clusters determined in the first study were evaluated and improved by domain experts.

Evaluating the output of classification

Two classification methods were used: the decision tree and rule-based classification (the JRip algorithm). In both cases, a 10-fold cross-validation method was used; this method divides the dataset into 10 parts, each of which has approximately the same proportion of classes as in the full dataset. Interactively, one part is held for testing while the rest (nine parts) are used to train the classification model. After 10 interactions, all errors are averaged to generate the overall error [49].
Evaluating association rules
The strength of association rules is given by support and confidence. The best association rules are those that satisfy the minimum support and confidence thresholds and have meaning in relation to the objective of the data mining. This principle was used as the basis for evaluating the association rules resulting from the application of the hotspot algorithm in relation to an agricultural dataset. Experiments with different levels of support were undertaken and the level of support that generated interesting rules was 6%; this was selected to build the association rules used in Study 2. The association rules presented in the form of a tree satisfies the thresholds of minimum support. The confidence level is given for each node of the tree.

Evaluating numerical prediction
The evaluation of the numerical regression was undertaken by testing the model with separate test sets. The accuracy of the predictive model is obtained by calculating the error, which is the difference between the predicted value and the real value. In Study 1, the mean absolute percentage error (MAPE) was used and in Study 2 the mean absolute error (MAE) was used. In the first study, because the electricity consumption forecasting Electricity of Mozambique (EDM) model was not available, the result was evaluated by comparing it with those of similar studies. The MAPE error obtained using the developed model was in the range of other similar works [55, 56]. On the other hand, visualization method was used to establish the closeness of the graphics of real and predicted electricity consumption. In Study 4, the MAE prediction error of the EDM model was compared with the MAE prediction error of the developed model.

3.3 Data mining in organizations
Data mining can be applied in several areas, such as agriculture, telecommunications, health, retail, insurance, science and energy, to name only a few. With the spread of the “big-data” phenomena, data mining is becoming increasingly important due to its ability to analyse large amounts of data. The KDnuggets poll of 2012 [57] lists more than 29 fields of application of data mining, with the five top areas including customer relationship management (CRM), health care, retail, banking, and education. In 2010, Annti and Jari [58] produced an edited volume containing several articles on data mining applications for the public and private sectors. The application of data mining in organizations is still in its early stages and faces several challenges, such as the difficulty of integrating data, the volume of data, the availability of skills, the cost and complexity [59]. Recently, several studies have attempted to understand these challenges and propose solutions to mitigate them [19] [18]. This thesis includes the results of a study of the level of awareness and use of data mining and the factors affecting its use in Mozambique [22], most of which are similar to those indicated by studies related to other countries.
Chapter 4: Research Methodology

4.1 Philosophical assumptions, approach, strategy and methodological choices

Figure 9 presents the general approach, philosophical assumptions, strategies, data collection methods and data analysis techniques applied in this thesis.

![Diagram showing predominant approaches, philosophical assumptions, strategies and choices adopted in this thesis]

The philosophical assumption defines how knowledge is obtained. Mayer [60] considers three categories, namely: positivism, interpretivism and critical. Positivist assumptions view reality as objective and able to be described by measurable properties that are independent of the researcher and her/his instruments [60]. Interpretivism considers that reality is a social construction, such as language, consciousness, shared meaning and instruments, which means that understanding of phenomena is attained through the values and meanings that people assign to them [60-62]. Critical is based on the notion that social reality is historically constructed and is produced and reproduced by people [60]. In general, this study is guided by an interpretative philosophical approach because the aim is to understand the level of use and the factors affecting the use of data mining in Mozambique, based on the perceptions of the participants in the research.

The study uses an inductive approach, which is characteristic of the interpretive philosophical assumption, as shown in Figure 9. In the study, no hypotheses are established in advance, rather the results are obtained from the analysis of evidence collected during the research process using different methods [63]. The deductive approach is the opposite, beginning with a hypothesis which is accepted or rejected through analysis of the data.
As a research strategy, the study uses exploratory multiple case studies. Two institutions, one in the area of electricity and the other in the area of agriculture, were selected for observation. Maputo city, the capital of Mozambique, was selected as the locus in which to employ questionnaires and focus group discussions. This triangulation was intended to corroborate the results as there was the expectation of a lack of information on the topic. The use of an exploratory approach was selected because the area of data mining is new in Mozambique: few people know about the topic and no previous study in this area has been undertaken.

Several studies on the success factors related to data mining have used deductive, positivist and quantitative approaches [14, 16, 17]. They have tended to employ a theoretical model to derive hypotheses that are verified by surveying a selected sample of potential informed respondents. For example, Hilbert [14] used a critical success factor (CSF) framework, Nemat and Barko [15] used a square route (TSR) framework, an extension of the iron triangle (IRT) employed by Cho et al. [13], and Huang et al. [64] used the technology adoption model 3 (TAM3). The use of a framework requires that the respondents have sound knowledge of the topic to give their genuine perceptions concerning the relationship between the dependent and independent variables in the investigation. In the case of the data mining in Mozambique, the number of informed respondents was insufficient to verify a hypothesis using quantitative means.

In this study, quantitative methods were applied during the development of data-mining applications in the area of agriculture and electricity. The qualitative data were transcribed and categorized (coded). The categories were defined based on the extensive list of factors affecting data mining extracted from the literature. Based on the categories, the qualitative data were labelled and sorted to extract the findings. The quantitative data were analysed using descriptive statistics. During the development of the application, an experimental strategy was used.

### 4.2 Selection of data-mining cases and samples

The selection of the organizations in which data mining was applied was based on an exploratory visit made to several organizations located in Maputo city during the second week of June in 2008 with the objective of finding research cases in the area of data mining. During the visit, the top managers of each organization were contacted. In this period, the concept of data mining was almost unknown to the people who were contacted, but there was a general interest in having better tools for data analysis. Despite the interest in experimenting using the data-mining application, most of the organizations were not prepared to provide their data for the study due to privacy issues. Only two organizations provided data for the application of data mining, one of which suggested a problem to be solved using this method. The data provided by the two organizations were employed in the application of data mining and developed to comprise three of the studies contained in this thesis. The databases of the two organizations were generated using different methods: one was obtained using automatic measurement, whereas the other was obtained through a national survey, using methodologies previously established and with intensive human involvement in the process of collection and data input.
The focus group discussion and questionnaires were applied in Maputo city, selected because it is the capital of the country. In the period of study, it was valid to consider systems and applications at the provincial level, as small-scale replications of the systems existing in Maputo, in particular in public institutions.

4.3 Data collection and processing and analytic techniques

**Study 1: ICT for automated forecasting of electrical power consumption: A case study in Maputo**

The data used in this study were provided by Electricity of Mozambique (EDM) and correspond to the consumption of electricity in Maputo from January 2003 to December 2009. Temperature data for the years 2007 and 2008 were also provided. The temperature data for the other years were not available at the required level of breakdown. To solve this problem, the temperature data and the humidity data were downloaded from the internet. After completing the dataset, the CRISP-DM data-mining process model was used to build a short-term load-forecasting model. Clustering and visualization techniques were used to understand and prepare the data and also for linear regression in the modelling step. The following tools were used to process and analyse the data and build the model: Microsoft Excel, WEKA workbench and Matlab. This study is reported in Paper 1 [25] included in this thesis and the summary of results is presented in Chapter 5.

**Study 2: Extracting patterns from a socioeconomic database to characterize small farmers with high and low corn yields in Mozambique: A data-mining approach**

The data used in this research were provided by the Directorate of Economics of the Ministry of Agriculture. The dataset contains information concerning livestock and agricultural production, as well as socioeconomic and demographic information concerning the farmers, information on the use of technology, community infrastructure, loss of production and calamities for the years 2007 and 2008. This dataset is part of the database of the Annual Agricultural Survey (TIA [23]), which is implemented every year. The data collection process follows the methodology established by the National Institute of Statistics of Mozambique. One crop (maize) and one province (Zambezia) were selected for analysis. For this study, the following tools were used: Cygwin scripts for data handling, the WEKA workbench, Matlab and Microsoft Excel. The details of this study are reported in Paper 2 and the summary of results is presented in Chapter 5.

**Study 3: Factors affecting the use of data mining in Mozambique**

The data for this study were collected using a survey. Questionnaires and focus group discussions were conducted to gain the perceptions of data analysts, ICT practitioners and managers working in different institutions in Maputo. An on-line form using LimeSurvey [65] was developed and the link distributed to potential respondents. Physical questionnaires were distributed to those respondents who preferred this format. In total,
110 respondents replied. However, approximately 46 (41.8%) responses were considered invalid because they did not answer the questions related to data mining. This is an indication that although the survey was distributed to potentially knowledgeable respondents, most were not aware of data mining. A focus group discussion with experts on data analysis and IT and with managers was organized. The participants discussed the topic over a period of two hours. The qualitative and quantitative data were analysed using Microsoft Excel, and compared. This study is reported in detail in Paper 3 [22] and the summary of results is presented in Chapter 5.

**Study 4: Short-term load forecasting of electricity consumption in Maputo**

The data used in this study were provided by the EDM and correspond to the consumption of electricity of Maputo for the period January 2003 to October 2012. The weather data were provided by the National Institute of Metrology (INAM) and comprised maximum and minimum temperatures in Maputo city for the period January 2003 to October 2012. Other data, such as price and working procedures, were also provided by EDM. Based on the information available, a short-term forecasting model was developed with aim of estimating the energy to be sold to the Southern Africa Power Pool. The details of this study are presented in Paper 4 [24] and the summary of results is given in Chapter 5. Microsoft Excel, Matlab and the WEKA workbench were the main tools used for the processing and analysis of the data.

### 4.4 Specific methods

This section presents the specific data-mining techniques used during the development of the application of Studies 1 [25], 2 [1] and 4 [24]. All three applications followed the KDD methodologies, also called data mining [7, 42, 45]. Study 1 [25] and Study 4 [24] used the CRISP-DM process model [42, 44] and Study 2 [1] used the Fayyad et al. process model [41, 45] (cf. Chapter 3).

### 4.5 Research development process

Figure 10 depicts the flow of the research aimed at determining the awareness and level of use of data mining and the factors affecting its use in Mozambique. The research started with a visit to several institutions in Mozambique with the objective of finding applications for data mining to serve as case studies. The result of the exploratory visits was the identification of two institutions in which to apply data mining, based on the availability of data and problems to be addressed through data mining. The exploratory visit also found that the concept of data mining was almost unknown and there was no application of data mining in the institutions visited. Thus, the exploratory study on the awareness and use of data mining and influential factors was conducted in Maputo city and in two institutions through the development of the application in the identified institutions to learn from experience what factors affect the use of data mining. The factors identified are later used to develop a framework facilitating the use of data mining in Mozambique.
4.6 Ethical issues

The study was conducted using data obtained from Electricity of Mozambique and Ministry of Agriculture. These two organizations agreed that their data can be used for this research purpose. Meetings with top managers of the two organizations were undertaken and each identified key contact person for further interaction. The results of the study were shared with the institution through the contact person before publication.

The participants of the focus group were asked to sign a consent letter, authorizing that the information collected during the discussion could be used for research purposes. The names of the participants in the focus group and survey and their respective institutions are not given. After the analysis of the data, a report was sent to all the participants in the focus group discussion for comment before publication. It was not possible to send the report to all participants of the survey because most participated anonymously.

4.7 Soundness of the research

In the interpretive paradigm, the soundness of research is measured by trustworthiness, which means that the confirmability, dependability, credibility and transferability of the research must be ensured [62]. This requires that the research is well described and that all the data and respective analyses are well documented and kept for future verification [62]. The research presented in this thesis is documented through papers, peer-reviewed and published in the proceedings of international conferences. The respondents, focal points and other participants in the research were involved in verifying the description of the findings and giving feedback before the publication of each study/paper. The study
also used triangulation by collecting empirical evidence using different methods. This
was intended to ensure the credibility of the study. The papers can be used by other
researchers to determine whether and to what extent the research is transferable to their
own settings.
Chapter 5: Contributions

5.1 Overview of the contributions

This research makes two main contributions:

i. The first contribution is the identification of the level of awareness and use of data mining in Mozambique.

ii. The second contribution is the identification of the factors affecting the use of data mining in Mozambique.

Other contributions of this study are as follows:

- Increasing awareness of the topic among the participants of the study.
- The development of applications in the area of agriculture and electricity.
- Contributing information concerning the level of use of data mining and the factors affecting its use in Mozambique.

5.2 Study 1: ICT for the automated forecasting of electrical power consumption: A case study in Maputo

The main output of this study is the short-term forecasting model for electricity consumption. EM clustering was used to describe consumption data and identify clusters that were used to build the forecasting model. The forecasting model comprises three "sub-models", each model corresponding to a cluster. The details of this study are presented in Paper 1 [25]. Below some of the key results are presented and Figures 11–14 show the results of the exploratory data analysis and describe the electricity consumption data.

Figure 11 shows the characteristic of electricity consumption during the day and year. Peak consumption is in the period 19:00–21:00, which is when most people are at home and before they go to sleep. This means that the main consumers of electricity are domestic users. Also, during the year, there are periods of high consumption and of low consumption. Furthermore, there are outliers, or errors (some strange peaks) in the data.
Figure 11: Three-dimensional view of electricity consumption in Maputo

Figure 12 shows the smooth quadratic long-term trend, which expresses how the consumption of electricity develops over the years. This is useful for the long-long term prediction of the consumption of electricity.

Figure 12: Long term trend of electricity consumption in Maputo

Figure 13 compares the temperature and electricity consumption. Both variables are correlated, which means that the temperature can be used as a variable for the short-term prediction of electricity consumption.

Figure 13: Temperature and electricity consumption in Maputo
Figure 13: Comparision of temperature and electricity consumption in Maputo

Figure 14 compares the variation in electricity consumption and the GDP per capita in the period 2003–2008 and shows that a similar increasing trend. This result means that there is a strong relationship between socioeconomic development and electricity consumption.

Clustering was then used to describe the data. The result of the clustering is illustrated in Figure 15 and indicates that the electricity consumption in Mozambique is characterized by three classes: cluster 0 with observations distributed along the year (holidays and weekends); cluster 1 with observations corresponding to the summer (January to March
and September to December); cluster 2 with observations corresponding to the winter (March to September). Each point of Figure 15 represents the daily average of electricity consumption and also identifies the class to which the day belongs in accordance with the result of the EM clustering algorithm.

Based on the results of exploratory analysis, an electrical consumption forecasting model composed of three sub-models, each corresponding to one class, was developed. Considering that elements of same class are similar [7], it was assumed that each sub-model could be developed using linear models. This approach is expressed using equation (15):

\[ L_{ci}^{d+1} = b_1 L_{ci}^d + b_2 T^{d+1} \]

where \( L \) is electricity consumption; \( d + 1 \) is the next day (day to be forecasted); \( d \) is the actual day; \( ci \) is the cluster number (in this case there are three clusters: \( c0 = \) holidays, \( c1 = \) summer working days and \( c2 = \) winter working days); \( b_1 \) and \( b_2 \) are regression coefficients; \( T \) is the average temperature.

The application of the forecasting model given in expression (15) generated the results presented in Figures 16–19. Figure 16 presents the real and the forecasted electricity consumption over five summer working days; Figure 17 presents the real and the forecasted electricity consumption over six winter working days; Figure 18 presents the real and the forecasted electricity for six Saturdays; finally, Figure 19 presents the real and the forecasted electricity for five Sundays. All the days presented in Figures 16–19 were selected randomly.

Fig. 15: Clustering of electricity consumption data of Maputo City
Figure 16: Comparision of forecast and real electricity consumption over five summer working days

Figure 17: Comparision of forecast and real electricity consumption over six winter working days
Lessons from Study 1

This study responds to research objectives b) and c). Furthermore, it demonstrates the potential of data mining in the area of the prediction of consumption of electricity. The factors affecting data mining as observed during the development of this application are summarized in Table 2.
Table 2: Lessons learned during Study 1 [25]

<table>
<thead>
<tr>
<th>NR</th>
<th>ASPECT/ FACTOR</th>
<th>FINDING OF OBSERVATION</th>
<th>COMMENT</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Awareness of data mining</td>
<td>weak</td>
<td>Available partially, in different granularity, with missing and anomalous values</td>
</tr>
<tr>
<td>2</td>
<td>Use of data mining</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Data issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Quality of data</td>
<td>Acceptable</td>
<td>Available partially, in different granularity, with missing and anomalous values</td>
</tr>
<tr>
<td>3.2</td>
<td>Data integration</td>
<td>Not integrated</td>
<td>available in different files, in different formats and from different formats</td>
</tr>
<tr>
<td>3.3</td>
<td>Age of data</td>
<td>9 years</td>
<td>the access of the data is not immediate. Need to be retrieved by the expert</td>
</tr>
<tr>
<td>3.4</td>
<td>Sufficiency of data</td>
<td>Acceptable</td>
<td>Data of operation, weather not available as required. The volume of data is increasing.</td>
</tr>
<tr>
<td>4</td>
<td>Privacy issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>4.1</td>
<td>Privacy</td>
<td>high concern on privacy</td>
<td>Avoid to provide institutional data</td>
</tr>
<tr>
<td>5</td>
<td>Skills and Human resources</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>Teamwork</td>
<td>do not exist in the area of data mining</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>Skills</td>
<td>do not exist in the area of data mining</td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>Understanding of data mining and</td>
<td>Limited</td>
<td></td>
</tr>
<tr>
<td></td>
<td>techniques</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Stakeholder support issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.1</td>
<td>Champion</td>
<td>do not exist in the area of data mining</td>
<td></td>
</tr>
<tr>
<td>6.2</td>
<td>Commitment of top management</td>
<td>Interested to see the result</td>
<td></td>
</tr>
<tr>
<td>6.3</td>
<td>Change management</td>
<td>Not observed</td>
<td></td>
</tr>
<tr>
<td>6.4</td>
<td>Organizational acceptance</td>
<td>Data mining is not known and is not used</td>
<td></td>
</tr>
<tr>
<td>6.5</td>
<td>Use of data mining results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Organizational issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1</td>
<td>Functional integration</td>
<td>No observed</td>
<td></td>
</tr>
<tr>
<td>7.2</td>
<td>Alignment of IT and business</td>
<td>Not Observed</td>
<td></td>
</tr>
<tr>
<td>7.3</td>
<td>Strategy of outsourcing</td>
<td>Not observed</td>
<td></td>
</tr>
<tr>
<td>7.4</td>
<td>Interdisciplinary learning</td>
<td>Not Observed</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Financial issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.1</td>
<td>costs</td>
<td>Not Observed</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Technology issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.1</td>
<td>Software and tools</td>
<td>Not specific for data mining</td>
<td></td>
</tr>
<tr>
<td>9.2</td>
<td>Existence of database</td>
<td>yes</td>
<td></td>
</tr>
<tr>
<td>9.3</td>
<td>Existence of data warehouse</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>

As can be seen from Table 2, some of the issues identified are as follows:

- **Problems to be solved with data mining**: there are problems to be solved with data mining in the field of electricity in Mozambique, for example the prediction of electrical consumption.
- **Quality of data**: the data have errors and outliers. However the quality is improving year by year. In the data set, it was observed that the quality of data in 2008 were better than in 2003.
- **Integration**: the data are not integrated. Data are stored by each department in different files.
- **Access to the data**: to access the data, as they are not integrated, multiple authorizations are required. Thus, it is necessary to go to several departments to obtain sectoral data.
• **Availability of the data**: some data are not available and where available they are not available at the required level of breakdown. For example, the weather data were available aggregated by month, but the historical prediction value was not available.

• **Privacy**: there are privacy concerns and most of the data were not available for this reason.

• **Data analysis tools**: the organization uses Microsoft Excel to analyse their data.

5.3 Study 2: Extracting patterns from a socioeconomic database to characterize small farmers with high and low corn yields in Mozambique: a data mining approach

The second study was also undertaken to address two objectives: i) to develop a data mining case study in the area of the agriculture; ii) to observe the level of awareness and use of data mining and the factors that may affect its use in this specific field.

The agricultural data were mined to extract rules describing the relationship between the level of productivity (high or low) and socioeconomic characteristics of small farmer households. The JRip and hotspot algorithms were used to extract the rules. The decision rules extracted using the JRip algorithm are presented in Table 3 and the association rules extracted by hotspot are presented in the form of a tree in Figures 20 and 21. The results indicate that households with high yields of maize are characterized by the commercialization of their products and relative wellbeing. Households with low maize yields are characterized by hunger and loss of production before and after the harvest.

Table 3: JRip decision rules characterizing small farmer households with high yields of maize

<table>
<thead>
<tr>
<th>Rule</th>
<th>Condition</th>
<th>Class</th>
<th>Support</th>
<th>Confidence</th>
</tr>
</thead>
<tbody>
<tr>
<td>R1:</td>
<td>(SoldQty &gt;= 38.556) &amp; (SoldQty &gt;= 278.399994) &amp; (SoldQty &gt;= 650.325012) =&gt; Class=H (221.0/52.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R2:</td>
<td>(SoldQty &gt;= 20) &amp; (SoldQty &gt;= 174) &amp; (NrChikens &gt;= 7) &amp; (EatReserve = No) &amp; (NrSecondaryAct &gt;= 1) &amp; (NrGoats &gt;= 3) =&gt; Class=H (30.0/5.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R3:</td>
<td>(SoldQty &gt;= 50.025002) &amp; (SoldQty &gt;= 150.075012) &amp; (LifeImproved = Better) &amp; (RegularTransport = No) &amp; (SoldQty &gt;= 278.399994) =&gt; Class=H (87.0/33.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R4:</td>
<td>(LifeImproved = Better) &amp; (LossBeforeHarvest = No) &amp; (WillSell = Yes) =&gt; Class=H (46.0/21.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R5:</td>
<td>(SoldQty &gt;= 50.025002) &amp; (SoldQty &gt;= 180.524994) &amp; (LossBeforeHarvest = No) &amp; (HasBike = No) =&gt; Class=H (68.0/27.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>R6:</td>
<td>=&gt; Class=L (8369.0/1891.0)</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Table 3 contains the decision rules generated by the JRip classification algorithm. In general, the rules indicate that small farmers with high production are those who have commercialized part of their production and have relative wellbeing.
Figure 20: Association rule tree for small farmer households with high yields of maize

Figure 20 presents the association rules that characterize small farmers with high yields of maize; for example, the topmost branch represents a rule that indicates that farmers with high yields sell part of their products and have some cashew trees. Figure 21 is the tree characterizing small farmers with low maize yields with each branch representing a rule, the topmost branch, for example, illustrating the rule that indicates that farmers with low yields eat more maize in the hungry period, have no surplus and no part-time workers to help on the farm. This means that these farmers suffer hunger, which results in them eating the surplus that was to be used for sowing in the next season.

Lessons from Study 2

This study responds to research objectives b) and c). Furthermore, it demonstrates the potential of data mining in the area of agriculture. Certain factors affecting the data mining observed during the development of this application using data from the agricultural sector are identified as follows:

- **Problems to be solved with data mining**: there are problems to be solved with data mining in the area of agriculture.

- **Quality of data**: the data have errors and outliers. However, this is improving yearly. In the data set, it was observed that the quality of the data in the agricultural survey in 2008 was better than that in the survey of 2002.

- **Integration**: the dataset is one of the most complete. It includes several aspects relating to agriculture, livestock and the socioeconomic life of farmers. This database is not integrated with other related databases, such as those concerning climate, commercialization/price, natural disasters, disease outbreaks, transport infrastructure, communication infrastructure and others.

- **Access to the data**: to access the data, special authorizations is required. Also, after authorization, it usually takes some time to obtain the data.

- **Availability of the data**: available upon request. The data available on the internet are aggregated.

- **Privacy**: there are privacy concerns. Most data are not available for this reason. For example, there are concerns in terms of providing the raw data of the survey because they contain the coordinates of the residences of the farmers.

- **Data analysis tools**: organizations predominantly use statistical methods.

These findings comprises some of the factors affecting the use of data mining in Mozambique. The summary of the findings is provided in Table 4.
<table>
<thead>
<tr>
<th>NR</th>
<th>ASPECT/FACTOR</th>
<th>FINDING OF OBSERVATION</th>
<th>COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Awareness of data mining</td>
<td>weak</td>
<td>They use of other statistical technique to analyses the dataset</td>
</tr>
<tr>
<td>2</td>
<td>Use of data mining</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Data issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Quality of data</td>
<td>Limited quality</td>
<td>there are some problem of accuracy, completeness and validity of the data as result of typing or wrong provision of information</td>
</tr>
<tr>
<td>3.2</td>
<td>Data integration</td>
<td>Not integrated database</td>
<td>some data are available from different institutions in different formats</td>
</tr>
<tr>
<td>3.3</td>
<td>Age of data</td>
<td>6 years</td>
<td>Annual measurements. The access of data is not immediate</td>
</tr>
<tr>
<td>3.4</td>
<td>Sufficiency of data</td>
<td>enough data available</td>
<td>The volume of data are increasing</td>
</tr>
<tr>
<td>4</td>
<td>Privacy</td>
<td>high concern on privacy</td>
<td>Avoid to provide institucional data</td>
</tr>
<tr>
<td>5</td>
<td>Skills and Human resources</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>Teamwork</td>
<td>do not exist in the area of data mining</td>
<td>There are considerable skills on statistics and other data analysis techniques. Limited skill in data mining</td>
</tr>
<tr>
<td>5.2</td>
<td>Skill</td>
<td>do not exist in the area of data mining</td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>Understanding of data mining and techniques</td>
<td>Limited</td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Stakeholder support issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.1</td>
<td>Champion</td>
<td>Do not exists in the area of data mining</td>
<td></td>
</tr>
<tr>
<td>6.2</td>
<td>Commitment of top management</td>
<td>Not observed</td>
<td></td>
</tr>
<tr>
<td>6.3</td>
<td>Existence of change management</td>
<td>Not observed</td>
<td></td>
</tr>
<tr>
<td>6.4</td>
<td>Organizational acceptance</td>
<td>Data mining is not known and is not used</td>
<td></td>
</tr>
<tr>
<td>6.5</td>
<td>Use of data mining results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Organizational issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1</td>
<td>Functional integration</td>
<td>No observed</td>
<td>The data mining is not used and the awareness among the observed people is weak. It make difficult to observe this aspects</td>
</tr>
<tr>
<td>7.2</td>
<td>Alignment of IT and business</td>
<td>Not observed</td>
<td></td>
</tr>
<tr>
<td>7.3</td>
<td>Strategy of outsourcing</td>
<td>Not observed</td>
<td></td>
</tr>
<tr>
<td>7.4</td>
<td>Interdisciplinary learning</td>
<td>Not observed</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Financial issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.1</td>
<td>costs</td>
<td>Not Observed</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Technology issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.1</td>
<td>Software and tools for data mining</td>
<td>do not exist specific software for data mining</td>
<td></td>
</tr>
<tr>
<td>9.2</td>
<td>Existence of database</td>
<td>yes</td>
<td>Aggregated data were available from web and downloadable</td>
</tr>
<tr>
<td>9.3</td>
<td>Existence of data warehouse</td>
<td>No</td>
<td></td>
</tr>
</tbody>
</table>
Fig. 21: Association rules tree for the small farmers households with low yield of maize [1]
5.4 Study 3: Factors affecting the use of data mining in Mozambique

The main result of this study was the identification of the level of awareness and use of data mining and the factors affecting its use in Mozambique. The findings indicate that the level of awareness and use of data mining are low. Only a limited number of practitioners in the area of IT know the technique or have used it. A significant number of respondents (41.8%) did not answer questions included in the questionnaire to test the level of knowledge. These questions aimed, for example, to provide a basic definition of data mining and a self-evaluation of the level of understanding of the topic. Those questionnaires that included respondents who answered but did not know what data mining was were considered invalid. It was assumed that the respondents were not sufficiently familiar with the topic to provide the answers required for the study. The factors that could influence the use of data mining in Mozambique as determined by the study are summarized in Table 5.

Table 5: Factors affecting the use of data mining

<table>
<thead>
<tr>
<th>STAKEHOLDERS SUPPORT ISSUES</th>
<th>DATA ISSUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existence of business champion</td>
<td>Quality of data</td>
</tr>
<tr>
<td>Commitment of top management</td>
<td>Availability of data warehouse</td>
</tr>
<tr>
<td>Existence of change management</td>
<td>Data integration</td>
</tr>
<tr>
<td>Organizational acceptance</td>
<td>Age of data</td>
</tr>
<tr>
<td>Use of data mining</td>
<td>Insufficient data</td>
</tr>
<tr>
<td>User request</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>ORGANIZATIONAL ISSUES</th>
<th>SKILLS AND HUMAN RESOURCE ISSUES</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional integration</td>
<td>Existence of teamwork</td>
</tr>
<tr>
<td>Alignment of IT and business</td>
<td>Existence of skill</td>
</tr>
<tr>
<td>Strategy of outsourcing</td>
<td>Understanding of data mining and techniques</td>
</tr>
<tr>
<td>Interdisciplinary learning</td>
<td>Awareness of data mining</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>FINANCIAL TECHNOLGY ISSUES</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>costs</td>
<td>Existence of adequate software and tools</td>
</tr>
<tr>
<td></td>
<td>Existence of database</td>
</tr>
<tr>
<td></td>
<td>Existence of data warehouse</td>
</tr>
</tbody>
</table>

This study responds to research objective a). It was intended to establish the level of awareness and use of data mining and factors affecting its use in Mozambique.

5.5 Study 4: Short-term load-forecasting model for Maputo

The objective of this study was to verify the progress of enablers for the application of data mining at EDM, in particular in relation to issues related to the availability, integration and quality of data. The results of Study 1 [25] were presented to the EDM team and during the presentation, the need to apply the forecasting model developed in Study 1 for the estimation of electricity to be sold on the day-ahead market (DAM) in Southern Africa Power Pool (SAPP) was raised. More data on electricity consumption,
prices and procedures of consumption forecasting were made available. The results of the clustering of electricity consumption data (see Figure 15) were analysed and improved by the domain expert, enhancing the accuracy of the model developed in Study 1 [25].

Figures 22–24 illustrate selected results from Study 4. Figure 22 presents the DAM estimated using the EDM model (DAMedm), the DAM estimated using the RR model developed in Study 4 (DAMrr), the real demand (Dr), the total consumption for the EDM model (Dr+DAMedm), the total consumption for the RR model (Dr+DAMrr) and the available electricity (Pk). The graph in Figure 22 shows that the area between the total consumption of the RR model and the available electricity is smaller than the area between the total consumption of the EDM model and the available electricity. This means that the RR model is able to forecast the DAM with greater accuracy than the EDM model. The accuracy of the RR model is also confirmed by the results of the MAE shown in Table 6. The MAE of the RR model is 13.02%, whereas the MAE of the EDM model is 31.20%.

![Figure 22: Comparison of the results of the RR model and the EDM model](image)

<table>
<thead>
<tr>
<th></th>
<th>Model EDM %</th>
<th>Model RR %</th>
</tr>
</thead>
<tbody>
<tr>
<td>MAE</td>
<td>31.2</td>
<td>12.02</td>
</tr>
</tbody>
</table>

Table 6: Comparison of the mean absolute error (MAE) of the EDM model and the RR model
Figure 23: Comparison of underestimated electricity using the EDM and RR models

Figure 24: Comparison of overestimated electricity using the EDM and RR models

Figure 23 compares the amount of electricity underestimated by the EDM model (coloured black) and the RR model (not coloured) per month. In this figure, it should be
noted that the EDM model shows a greater loss of electricity than the RR model. In Figure 24, which compares the amount of electricity overestimated by the EDM model (coloured black) and the RR model (not coloured), it can be observed that the RR model overestimates more electricity for the DAM than the EDM model, but the overestimated amount is insignificant compared to the electricity saved with the deduction of the underestimation. The advantage of the RR model can thus be demonstrated by the financial estimations presented in Table 7.

Table 7 shows that by using the RR model, the EDM will save approximately USD 8,652,883.20 in a period of 10 months. This saving has several positive implications, such as the company being able to improve its service and reduce its tariff, the ability to invest in new infrastructure, enabling increased access to electricity, reducing the adverse impact on the environment created by the use of firewood energy source, and so on.

Table 7: Comparison of the financial implications of the EDM and RR models

<table>
<thead>
<tr>
<th>Month</th>
<th>Revenue DAMedm</th>
<th>Cost OVEedm</th>
<th>Revenue DAMrr</th>
<th>Cost OVErr</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>jan</td>
<td>2,625,799.30</td>
<td>127,063.00</td>
<td>3,434,371.10</td>
<td>247,866.84</td>
<td>687,767.96</td>
</tr>
<tr>
<td>feb</td>
<td>2,369,497.20</td>
<td>112,581.43</td>
<td>3,101,807.60</td>
<td>223,413.99</td>
<td>621,477.84</td>
</tr>
<tr>
<td>mar</td>
<td>1,970,771.30</td>
<td>173,993.23</td>
<td>2,636,892.20</td>
<td>342,749.96</td>
<td>497,364.17</td>
</tr>
<tr>
<td>apr</td>
<td>2,671,978.80</td>
<td>92,806.40</td>
<td>4,044,773.20</td>
<td>211,835.41</td>
<td>1,253,765.39</td>
</tr>
<tr>
<td>may</td>
<td>3,052,263.80</td>
<td>88,934.37</td>
<td>4,629,429.50</td>
<td>139,700.00</td>
<td>1,526,400.07</td>
</tr>
<tr>
<td>jun</td>
<td>3,063,529.90</td>
<td>178,699.01</td>
<td>4,711,075.00</td>
<td>236,965.64</td>
<td>1,589,278.47</td>
</tr>
<tr>
<td>jul</td>
<td>4,831,463.40</td>
<td>250,340.74</td>
<td>6,356,750.60</td>
<td>332,323.02</td>
<td>1,443,304.92</td>
</tr>
<tr>
<td>aug</td>
<td>4,683,128.40</td>
<td>311,088.03</td>
<td>5,399,073.10</td>
<td>492,586.49</td>
<td>534,446.24</td>
</tr>
<tr>
<td>sep</td>
<td>2,701,306.40</td>
<td>60,413.27</td>
<td>3,248,548.20</td>
<td>260,537.72</td>
<td>347,117.35</td>
</tr>
<tr>
<td>oct</td>
<td>1,766,400.50</td>
<td>129,504.08</td>
<td>2,093,872.30</td>
<td>305,015.08</td>
<td>151,960.80</td>
</tr>
<tr>
<td>Tot</td>
<td>29,736,139.00</td>
<td>1,525,423.55</td>
<td>39,656,592.80</td>
<td>2,792,994.15</td>
<td>8,652,883.20</td>
</tr>
</tbody>
</table>

Lessons from Study 3

This study responds to research objectives b) and c). Furthermore, it demonstrates the potential of data mining in the area of the prediction and commercialization of electricity. Some of the factors affecting the use of data mining observed during the development of this application are as follows:

- **Problems to be solved with data mining:** there are problems in the field of electricity market that can be solved with data mining.
- **Quality of data:** the data show some errors and outliers. However their quality is improving yearly. In the data set, it was observed that the quality of the data for 2011 was better than the quality of data in 2003.
- **Integration:** data are not integrated and are stored by each department in different files,
• **Access to the data:** data are not integrated and multiple authorizations are required from several departments.

• **Availability of the data:** there was progress in terms of the availability of data. The weather data were available at the required specificity. The data related to price and processes of forecasting electricity were available.

• **Privacy:** there were privacy concerns. Most of the data were not available for this reason.

• **Data analysis tools:** for the most part, organisations use Microsoft Excel to analyse the data.

These findings comprise some of the factors affecting the use of data mining in Mozambique. The summary of the findings is given in Table 8.

*Table 8: Lessons learned from Study 4 [24] on the level of adoption and factors affecting the use of data mining*

<table>
<thead>
<tr>
<th>NR</th>
<th>ASPECT/FACTOR</th>
<th>FINDING OF OBSERVATION</th>
<th>COMMENTS</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Awareness of data mining</td>
<td>weak</td>
<td></td>
</tr>
<tr>
<td>2</td>
<td>Use of data mining</td>
<td>None</td>
<td></td>
</tr>
<tr>
<td>3</td>
<td>Data issues</td>
<td>Available partially, with missing and anomalous values, quality is improving</td>
<td></td>
</tr>
<tr>
<td>3.1</td>
<td>Quality of data</td>
<td>Acceptable</td>
<td></td>
</tr>
<tr>
<td>3.2</td>
<td>Data integration</td>
<td>Not integrated</td>
<td>available in different files, in different formats and from different formats</td>
</tr>
<tr>
<td>3.3</td>
<td>Age of data</td>
<td>9 years</td>
<td>available immediately</td>
</tr>
<tr>
<td>3.4</td>
<td>Sufficiency of data</td>
<td>Acceptable</td>
<td>The volume of data is increasing</td>
</tr>
<tr>
<td>4</td>
<td>Privacy</td>
<td>high concern on privacy</td>
<td>Avoid to provide institutional data and information</td>
</tr>
<tr>
<td>5</td>
<td>Skills and Human resources</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.1</td>
<td>Teamwork</td>
<td>do not exist in the area of data mining</td>
<td></td>
</tr>
<tr>
<td>5.2</td>
<td>Skill</td>
<td>do not exist in the area of data mining</td>
<td></td>
</tr>
<tr>
<td>5.3</td>
<td>Understanding of data mining</td>
<td>Data mining is not known</td>
<td></td>
</tr>
<tr>
<td>5.4</td>
<td>Techniques</td>
<td></td>
<td></td>
</tr>
<tr>
<td>5.5</td>
<td>Awareness of data mining</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6</td>
<td>Stakeholder support issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>6.1</td>
<td>Champion</td>
<td>do not exist in the area of data mining</td>
<td></td>
</tr>
<tr>
<td>6.2</td>
<td>Commitment of top management</td>
<td>Interested to see the result</td>
<td></td>
</tr>
<tr>
<td>6.3</td>
<td>Change management</td>
<td>Not observed</td>
<td>The data mining is not used and the awareness among the observed people is weak. It make difficult to observe this aspects</td>
</tr>
<tr>
<td>6.4</td>
<td>Organizational acceptance</td>
<td>Data mining is not known and is not used</td>
<td></td>
</tr>
<tr>
<td>6.5</td>
<td>Use of data mining results</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7</td>
<td>Organizational issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>7.1</td>
<td>Functional integration</td>
<td>No observed</td>
<td></td>
</tr>
<tr>
<td>7.2</td>
<td>Alignment of IT and business</td>
<td>Not Observed</td>
<td>The data mining is not used and the awareness among the observed people is weak. It make difficult to observe this aspects</td>
</tr>
<tr>
<td>7.3</td>
<td>Strategy of outsourcing</td>
<td>Not observed</td>
<td></td>
</tr>
<tr>
<td>7.4</td>
<td>Interdisciplinary learning</td>
<td>Not Observed</td>
<td></td>
</tr>
<tr>
<td>8</td>
<td>Financial issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>8.1</td>
<td>costs</td>
<td>Not Observed</td>
<td></td>
</tr>
<tr>
<td>9</td>
<td>Technology issues</td>
<td></td>
<td></td>
</tr>
<tr>
<td>9.1</td>
<td>Software and tools for data mining</td>
<td>Not specific for data mining</td>
<td>Not specific for data mining</td>
</tr>
<tr>
<td>9.2</td>
<td>Existence of database</td>
<td>yes</td>
<td>It is not relational database. The measurement are kept in excell files</td>
</tr>
</tbody>
</table>
Chapter 6: Conclusions and Future Directions

6.1 Conclusions

This thesis presents the results of a study to identify the level of awareness and use of data mining and factors affecting its use in Mozambique. The results show that data mining in Mozambique is in its infancy, with low levels of use and low awareness. However, there are problems that can be solved with data mining. The factors affecting the use of data mining in Mozambique are as follows:

- **Data issues** – low quality of data, limited integration of data, limited availability of data and lack of accessibility in a timely fashion.
- **Skills and human resources** – limited availability of human resources with skills in the area of data mining.
- **Stakeholder issues** – limited use of the results of the data analysis, lack of data mining champions and lack of awareness of data mining on the part of users and managers.
- **Organizational issues** – alignment of IT and business, functional integration.
- **Financial issues** – cost of implementing data mining.
- **Technological issues** – most companies do not have data warehouses or databases; the software for data mining is not yet common in companies.
- **Privacy issues** – organizations avoid providing data and information due to privacy concerns.

These factors were both observed and indicated by experts who participated in the focus group and by the participants in the survey.

Paper 1 presents a short-term electricity consumption model based on data mining using the data available at a studied company in which short-term forecasting was a problem that needed a solution. During the development of the forecasting model, a number of problems were identified, such as the quality of data, the unavailability of some data, limited integration of data, and the difficulty in providing some information due to privacy issues. It was also observed that the access to information was not immediate even when available. Issues concerning a lack of skills were evident as the experts contacted in the company were not of data mining. They were using Microsoft Excel to predict electricity consumption, and this activity was being undertaken by experts in the electricity market, whereas the IT department was addressing other issues, such as billing systems. The observations in Study 1 are similar to those in Study 4 at the same company. The slight difference was the increase in the level of the interest concerning the results and availability of more data.

During the development of Paper 2, which was based on the study in the area of agriculture, it was observed that the agricultural dataset is rich, with information on
different agricultural issues. The agricultural database also represents an opportunity for solving different problems with data mining. The application developed was able to extract association rules between the productivity of maize and the socioeconomic characteristics of small farmers. As in Study 1, during the development of the data mining application, problems were identified in terms of the quality of data and the integration of the dataset with other related data, such as climate data and price data. There were also privacy issues in terms of providing fields that could facilitate the identification of the farmer, for example through the use of coordinates. Several studies are being conducted using this dataset, but they all use statistics and econometric techniques. There are no skills in data mining.

Study 3 confirmed the observations made in Studies 1, 2 and 4. Study 3 employed focus group discussion and inquiry of experts in the area of ICT and data analysis, as well as organization managers. Study 3 also integrated the factors identified in the three other studies and identified additional factors that were not observed in those three studies, such as stakeholder issues, organizational issues, financial issues and technological issues.

The list of factors identified in this study is more extensive than presented in most other studies accessed during the literature review, even though some are similar to those determined by other studies. The extensive of the factors in these lists means that in the case of Mozambique, it is necessary to pay attention to more issues to ensure success of data mining projects. For example, Sim [17] only considers data issues in terms of the success of data mining, Bole et al. [18] consider organizational issues, data issues, skills and the quality of data important, Hilbert[14] considers data issues, organizational issues and technology issues important. All the reviewed studies highlight data issues, in particular the quality of data, as important factors for data mining, followed by organizational factors, skills and stakeholder interests. Technological and financial aspects are mentioned by a few studies as important factors. Privacy is a factor mentioned with less frequency in the reviewed studies.

What has become evident during the development of the application employed here is the need to improve the quality, integration, access and management of data to improve the quality and the result of data mining. The quality of data is also important for the validity of the models; for example, the electricity model is highly sensitive to changes in temperature, which means that the results of forecasting can be affected to a considerable extent by an error in the prediction of the temperature.

6.2 Future research directions
This thesis presents factors affecting the use of data mining in Mozambique. Action needs to be taken to minimize the effect of these factors and facilitate the implementation and use of data mining in Mozambique. This includes awareness campaigns, the development of human capacity, the development of infrastructure to support the management and analysis of data, the implementation of policies to promote data management and analysis, and support for studies to enhance data mining targeted at specific sectors.
For future research, a framework would be developed to facilitate the successful implementation and use of data mining in Mozambique. The framework will contain two components: i) a maturity model to measure the capability of organizations to use data mining and to provide guidelines to improve its capability for the successful use of data mining; and ii) a risk analysis model to measure the risk of organizations in using data mining based on the existing factors.
References


[40] E. Macome, "The Dynamics of the Adoption and Use of ICT-based Initiatives for Development: Results of a Field Study in Mozambique," PhD, School of Information Technology, University of Pretoria, Johanesburg, 2002.


APPENDIX A: Paper I
ICT for Automated Forecasting of Electrical Power Consumption: A Case Study in Maputo

Constantino SOTOMANE¹², Lars ASKER¹, Venâncio MASSINGUE²
¹DSV, Forum 100, SE-164 40 Kista, Sweden
Tel: +46 086747002, Email:{asker,Sotomane}@dsv.su.se
²MCT, Av. Patrice Lumumba n.º 770, Maputo Moçambique
Tel:+258 21 352800, Fax: +258 21 352860/80
Email:{venancio.massingue,constantino.sotomane}@mct.gov.mz

Abstract: Accurate short term load forecasting is crucial for efficient operations planning of electrical power systems. We present a model for automatic forecasting of the short term (24 hours) electrical power consumption in Maputo, Mozambique. The proposed model is based on analysis of historical records of power consumption combined with information about additional factors that influence the consumption. The data is clustered into segments with the objective of identifying similar consumption patterns. These consumption patterns are then correlated with weather conditions and used to construct an automated prediction model for load forecasting. Today these forecasts are made manually by experts at Electricidade de Moçambique (the local power company) using conventional methods. The automated prediction model that was developed in this project presents an accurate and consistent complement to manual prediction and is currently being evaluated for the possibility of augmenting the manual forecasts with additional information.

Keywords: Electrical forecasting, clustering, decision tree, regression, data mining

1. Introduction

Accurate forecasting of electrical power consumption allows for efficient planning of production and distribution of electrical power. Load forecasting can be categorized into three types based on the period of prediction: short-term forecasts range from one hour to one week, medium-term forecasts range from a week to a year, and long-term forecasts are those that stretch for more than a year [1-3]. The different types of forecasts are relevant for different types of decisions in a power company [3]. The long and medium term forecasts are used to determine the capacity of generation, transmission, distribution, scheduling of annual maintenance, and affect future expansion plans, etc. [3]. Short-term forecasts are used for adequate scheduling and operations of power systems and provide input data for load flow and contingency analysis [2, 4].

The electrical demand is influenced by several factors such as: factors related to economic development and demographic grow, time factors related to seasonal, weekly and holiday effects; and weather factors [5]. The current and predicted price of electricity has also influence on the electrical demand [4].

Forecast of electrical demand is done using a wide variety of methods and techniques. Medium and long term forecasting is generally done using econometric and end-use based approaches [3, 6] while short term forecasting use different methods such as similar day
approach [7-9], regression models [3, 6], time series [8], neural networks [2, 4], expert systems [3, 6], fuzzy logic [9], statistical learning algorithms and data mining [7-10]. In recent years there has been an increasing use of Data Mining approaches, as part of efforts to increase transparency and interpretability of the prediction models [7, 8, 10-12]. Such approach give the possibility of discovering patterns and hidden dynamics that describe the behavior of the electrical load and related factors [8, 10, 12] and based on that, build models which are easily interpreted and that can thereby contribute to an increased understanding of the underlying factors that affect the power consumption [7].

Several data mining based models uses advanced techniques such as neural networks [8], fuzzy logic, wavelet transformation [8, 12], support vector machines [11] and others for detecting patterns and modeling, making the model difficult to implement and understand in particularly in situations characterized by lack of expert with knowledge on this kind of techniques. In our work, we use data exploration analysis with graphic representation to facilitate the understanding of the patterns and the structure of the models. Other aspect in our work is the use of simple linear models. The combination of understandable model structure and simplicity of the model makes it easy to implement in easy available application such as Excel.

Analysis of historical records of electrical consumption in Maputo, Mozambique indicate that the short term consumption is above all characterized by three factors: (1) type of day (workday, weekend or holiday), (2) temperature, and (3) season. The automatic prediction model that we present in this paper thus takes as input these three factors and produces as output a forecast of the consumption per hour for the day in question. In the experimental evaluation we used historical data and thus the actual temperature while the intended input for the final model will be based on meteorological forecasts of the temperature.

2. Objective

The objective of this work is to develop an automated prediction model that can be useful as an accurate and consistent complement to the daily manual 24 hour forecast of electrical power consumption that is currently done at EDM.

3. The Data Mining Process

The data mining process can be divided into problem definition, data exploration, data preparation, modelling, evaluation and exploration [13]. These are described below.

**Problem definition:** Mozambique Electricity’s company (EDM), needs to forecast electrical load one day ahead in order to optimize energy production.

**Data exploration** - EDM has access to records of electricity consumption in different regions of which Maputo area is one which one was selected for the study. The data correspond to hourly measurements from January 2003 to December 2009. Because the weather data was not available per hour from EDM for the corresponding period, we downloaded it from Internet [14]. Data exploratory analysis was performed in order to gain insight of the data [15, 16]. We present below selected graphics that explain the characteristic of the electricity load in Maputo.
Figure 1 is a three-dimensional plot of electrical load; it shows the variation of the consumption of electricity per hour over each day and during the five year period. Figure 2 is a two-dimensional plot of the electrical load showing the seasonality and the long-term trend. Figure 3, shows the correlation between the long term electrical demand trend and the economic development, expressed in GDP per capita.

Figure 4 shows the normalized electrical load curve after removing the long-term trend; it shows that the means is zero but with the seasonal variability remaining.

Figure 5 is a comparison between the electrical load and the temperature over the year. There is a strong correlation (Correlation Coefficient is 0.93) between the two variables. The high consumption and temperature corresponds to the months of January, February,
March, October, November, December (summer – hot and humid period in Mozambique). The rest of period is considered the spring season (cooler and dry). A similar comparison was made between load and humidity, but it was observed that the correlation is not strong (Correlation Coefficient is 0.35).

**Data Preparation**: The data preparation stage, consisted on cleaning the electricity load data and the weather data. In electricity load data we found several abnormal load curves which were removed from our database, there was no missing values in the electrical load data. The weather data had several missing values, which were replaced by the result of linear interpolation between the precedent and the antecedent of each missing data point [10]. After the correction of the missing data we represent the electrical load in a matrix format with the instances corresponding to days and the columns to hours.

4. **Clustering**

Clustering is an unsupervised learning method that partitions a data set into (possibly overlapping) groups [18, 19]. In order to cluster the load forecasting dataset we first removed the long-term trend to make all data comparable. The long-term trend was removed by fitting the data to a curve expressed by the quadratic polynomial equation (1).

\[ p = 14.45 \times 10^{-6} x^2 + 1.8 \times 10^3 x + 119.7 \]

For clustering we used the EM (Expectation Maximization) algorithm embedded in the machine learning workbench available on the WEKA [20, 21]. We tried clustering with 2, 3, 4, 5 and 6 classes. The clustering with 3 classes provided better results. After clustering the load forecasting dataset we build a separate model for each group. We used the C4.5 decision tree algorithm, also from WEKA [20, 21], to train a classifier that assigns a new example to one of the clusters. Figure 6 shows the resulting classifier for 3 classes: class0 corresponds to summer workdays, class1 corresponds to winter workdays, and class2 corresponds to weekends and holidays.

![Decision Tree](image)

*Figure 6: Decision Tree which describe the Relation between the Clusters and Time Factors Affecting the Electrical Demand*

In the next section we use the output of this stage to build 3 models one for each class. Then these models will serve to forecast the electricity consumption.

5 **Modeling and Forecasting the Electrical Load**

Once our data are grouped in classes representing seasons, workdays, weekends and holidays, we built models for each class using the multiple regression equation[6] given below.

\[ L_{c+1}^d = b1 \times L_c^d + b2 \times T_c^d \]

Where \( L \) the electrical load, \( d \) is the index of the day, \( c \) is the class of the day and \( T \) is the temperature. We are using only the temperature as the weather factor, because during
the exploratory data analysis we observed that the humidity is less correlated with the Electricity consumption.

The model was trained using the data from January 2003 to December 2008, and the data of year 2009 were used as test set. Some results of the modelling are presented in figures 7, 8, 9 and 10 and in Table 1.

Figure 7. Sample of Comparison between the Predicted and Real Electricity Consumption for the Working Days of Summer 2009.

Figure 8. Sample of Comparison between the Predicted and Real Electricity Consumption for the Working Days of Winter 2009.

Figure 9. Sample of Comparison between the Predicted and Real Electricity Consumption for the Saturdays of 2009.
Figure 10: Sample of Comparison between the Predicted and Real Electricity Consumption for the Sundays of 2009.

Table 1 below presents the average of the Mean Absolute Percentage Error (MAPE), obtained for each class. For the weekends the largest MAPE was obtained in Sundays and for workdays the largest MAPE was obtained during summer. The MAPE is calculated using equation 3.

\[
MAPE = \frac{1}{N} \sum_{i=1}^{N} \left| \frac{\text{RealLoad} - \text{ForecastedLoad}}{\text{RealLoad}} \right|
\]

Where \( N \) is the number of data point and \( i \) the index of the data point.

<table>
<thead>
<tr>
<th></th>
<th>Average MAPE</th>
<th>Maximo MAPE</th>
<th>Minimo MAPE</th>
<th>Standard deviation</th>
</tr>
</thead>
<tbody>
<tr>
<td>WorkDays (summer)</td>
<td>3.03</td>
<td>12.58</td>
<td>0.013</td>
<td>2.45</td>
</tr>
<tr>
<td>WorkDays (winter)</td>
<td>2.23</td>
<td>10.15</td>
<td>0.15</td>
<td>1.82</td>
</tr>
<tr>
<td>Saturdays</td>
<td>2.75</td>
<td>7.02</td>
<td>0.1</td>
<td>2.05</td>
</tr>
<tr>
<td>Sundays</td>
<td>3.53</td>
<td>13.07</td>
<td>0.09</td>
<td>2.83</td>
</tr>
<tr>
<td>Average</td>
<td>2.9</td>
<td>13.07</td>
<td>0.013</td>
<td>2.3</td>
</tr>
</tbody>
</table>

The average of MAPE of the 337 days forecasted during the years 2009 is 2.9% which is a promising result. The prediction of the public holidays is not considered in our approach because each public holiday has its own characteristics and can occur in different seasons. We do not have enough data to train a model for each specific holiday; moreover the model is dependent to the accuracy of weather forecast.

6 Business Benefits

Incorrect estimation of electrical demand will result in underestimation or overestimation. Underestimation will lead to production below required capacity, which might result in poor quality of service including localized brownouts, blackouts and even penalties for using beyond predetermined levels [22]. Overestimation will lead to waste of energy (energy produced and available but not consumed). A precise forecast will without doubt contribute to minimise this problem, which in turn leads to lowered end user costs and a contribute to environmental protection.

Mozambique is one of few countries in southern Africa that is self sufficient in electricity. This is mainly due to the electric power generated by the Cahora Bassa hydroelectric power plant. A major part of the surplus production is sold to South Africa. Thus, accurate demand forecasting is important to cope with the increasing demand for domestic consumption and well as export [23]. The strategy of the government is to increase the access to electricity (currently only 11% [24] of rural areas have access) by
expanding the infrastructure for electricity to the whole country [25]. An accurate prediction of the electrical demand will help to reduce cost related to the loss of the electricity which is not used or related to the payment of penalties for using more than the predetermined level, these savings can be redirected to the investment in the expansion of the electrical network. The Data Mining process used to build the present forecast model, results in a simple and understandable model (linear regression model) which can be implemented in a simple application such as excel.

7. Conclusions

In this paper we presented research aimed at building models for short term forecast of electrical consumption. Based on historical data of electrical consumption we identified a set of models, one for each class of similar days, to forecast electrical consumption. The initial data exploration was helpful in order to better understand the factors that influence electrical consumption in Maputo. It also facilitated the data correction and the selection of relevant attributes for automatic classification. We observed that the clustering method was able to identify similarities in the database, which contributed to the accuracy of model. The partitioning of the data into different groups of similar days increased the efficiency of the implemented model. The process model contributes not only to electrical demand forecasting but also provides a tool for better understanding of power consumption behaviour which in turn will help for better market and risk analysis in the area of electricity consumption and forecasting. The developed model is currently in the process of evaluation by the personnel of EDM for possible adoption. The personnel of EDM were also actively involved in research process which resulted on the forecast model.

One problem encountered during this research was the unavailability of weather data from official sources and also other data related to the operation of the electricity network that can affect the consumption pattern such as interruption of the network for maintenances. The availability of different data types and of good quality will enable the development of more efficient forecasting models; it is important that the various sectors that generate and own data, such as power consumption data, meteorological data, social and economic development data, etc., are able to collaborate and share this in an efficient and standardised way. With access to such data, we believe that automated prediction models, such as the one developed in this project, will soon prove to be an even more useful support in the process of efficient short term load forecasting. The result of the evaluation will be used to improve the forecasting model and implement for the application at the EDM.

Acknowledgement

The authors would like to thank the Mozambique electricity company for all the support and collaboration which make possible the realization of the research. Special thank to Mr. Sandro Chiang for their commitment and support.

References

APPENDIX B: Paper II
Extracting Patterns from Socioeconomic Databases to Characterize Small Farmers with High and Low Corn Yields in Mozambique: a data mining approach.

Constantino Sotomane\textsuperscript{1,2}, Jordi Gallego-Ayala\textsuperscript{2}, Lars Asker\textsuperscript{1}, Henrik Bostrom\textsuperscript{1}, and Venancio Massingue\textsuperscript{2}.

\textsuperscript{1}Dep. of Computer and Systems Science, Stockholm University
Forum 100, SE-164 40 Kista, Sweden
E-Mail{sotomane,asker,henke}@dsv.su.se
\textsuperscript{2}Ministry of Science and Technology
Av. Patrice Lumumba, N 770, Maputo, Mozambique
E-mail{constantino.sotomane,jordi.gallego-ayala,venancio.massingue}@mct.gov.mz

Abstract. Mozambique is mainly a rural country. Agriculture is a pillar of the Mozambique economy and is the main source of income for 80\% of the population living in rural areas. One of the major problems in the agricultural sector is low productivity, which for most crops is the lowest in Africa. The main food crop cultivated in Mozambique is maize. This research aims to characterize households with high and low maize yields based on the National Agricultural Survey Data from 2007 and 2008 using a data mining approach. To this end, we used: a) decision trees, b) association rules, and c) classification rules. The results show that households with high maize yields are those with the capacity to generate income through the commercialization of their production and agricultural assets. Households with low maize yields are associated with production loss before harvest which results in food insecurity.

Keywords: Data mining, Maize, Decision Trees, Association Rules, Classification Rules, Mozambique.

1 Introduction

The agricultural sector is one of the main pillars of Mozambican socioeconomic growth; in 2010 agriculture contributed 23\% of GDP, and this sector employs approximately 80\% of the active population living in rural areas, some 70\% of Mozambicans [1-3].

Rural areas are highly affected by poverty, food insecurity, low income and unemployment as a result of low agricultural productivity [4]. These problems could be solved by increasing production and productivity. The agricultural sector is not only
important for rural areas; it also provides crops for export and food to the urban popu-
lation.

One of the main problems in agriculture in Mozambique is productivity, which is
one of the lowest in Africa. Today, increasing production is related to increasing the
area of production and labour because agriculture is predominantly subsistence and
rainfed [4]. There are efforts at both continental and national levels to solve the prob-
lems of agricultural development. At the continental level, the New Partnership for
Africa's Development is implementing the Comprehensive Africa Agricultural Deve-
lopment Programme which aims to eliminate hunger and reduce poverty through agricu-
ture, focusing on land and water management, market access, food supply, and
hunger and agricultural research [5].

In Mozambique, the scientific community has undertaken extensive studies using
National Agricultural Survey (TIA) data to analyse socioeconomic aspects of rural
areas of Mozambique which potentially affect agricultural productivity, including
poverty, adoption of technology, and farm household efficiency [2, 3, 6, 7]. These
studies predominantly used econometric tools in their analyses. In contrast, we pro-
pose to use a data mining approach in order to characterize households with high and
low yields of maize. The use of data mining has two objectives: a) to provide an ana-
lytical tool which would be an alternative to the econometric methods commonly used
to analyse TIA data, b) to introduce the use of this method in this field of study to
enable better exploration of the large volume of data which the agricultural sector is
continuously gathering. The volume of data in this sector is set to increase over time
and the technology for collecting and storing data will evolve. Thus, data mining be-
comes an effective tool for extracting information from these data. Some of the data
mining techniques have the advantage of representing the results in a more intuitive
and easy form for assimilation. Therefore the main objective and novelty of this study
is the use of data mining techniques to characterize households with high and low
maize yields in rural areas of Mozambique. The results of this study are intended to
provide a support tool to practitioners and decision makers for the planning and de-
velopment of agricultural policies in Mozambique.

The remainder of this paper is structured as follows: this introductory section is
followed by a description of the main undertaking of this study which was to integrate
data mining techniques in the analysis of the agricultural sector. Section three de-
scribes in detail the methodology used in this work: the knowledge discovery from
data (KDD) process. Section four presents the main results of this study. Finally, sec-
tions five and six present the discussion and conclusions derived from this research.

2 Overview of data mining in agriculture

The field of data mining is evolving: nowadays this branch of science is applied in
several disciplines [8-10].

The application of data mining in agriculture and related fields is still emerging
[10]; nevertheless, several studies using this method have been conducted to date. The
area of agriculture is multi-disciplinary and the application of data mining can be
found in each of the subfields, which include farming, water management, environmental management, livestock, soil, weather, and agro-processing, among others. Some of the works related to the application of data mining in agriculture and subfields are described in literature [10-12]. One important subfield is related to crop yield prediction. There is a growing literature on the study of crop yield prediction using statistical methods, most recently using machine learning and data mining techniques. Drummond et al. used an artificial neural network (ANN) to build a model to forecast yield based on soil and site properties [13]; Serele used ANN to build a model to predict corn yield from topographical features, vegetation and texture indices [14]; Ruß used data from 2003–2006 collected by GPS and sensors placed on three fields in Germany to forecast yield using the support vector machine (SVM), ANN and regression trees [15]. Ruß used attributes related to the amount of fertilizer applied, vegetation, and electrical conductivity. Comparing the tree techniques, Ruß concluded that SVM provides a more accurate model but with limited understandability, while regression trees provide a model easy to understand but with poor accuracy. The problem of model understandability was proposed as future work. In line with this, Ruß and Kruse applied a two-step hierarchical clustering method to delimit the field in spatial contiguous clusters [16]. In the first phase, Ruß and Kruse used the K-means algorithm for tasselling and in the second phase they used hierarchical clustering to merge the small clusters created in the first phase. The objective of this algorithm was to determine the so-called management zones for better fertilization, given that different parts of fields may require different amounts of basic fertilization [16], since better fertilization can result in better yields. Waheed applied the CART decision tree to classify plots according to irrigation and weed and nitrogen levels, achieving an accuracy ≥75% [17]. The practical application of this work was rapid mapping of the effects of irrigation, and weed and nitrogen management, which in turn can be useful in planning in-season irrigation and weed and nitrogen management for maize.

The works mentioned above and most of the recent trends in this research area are focused on precision agriculture and the use of geo-referenced data collected using novel equipment such as ground-based sensors, aerial and satellite imagery, and soil sampling. However, within this field of research there are few works that use socioeconomic data to model crops. Ekasingh et al. combined socioeconomic data and biophysical characteristics to simulate farmer crop choice in the watershed in northern Thailand using decision trees [18]. The decision tree model was built using four socioeconomic variables, namely: land unity, estimated cost of production, the labour ratio, and the estimated profit level. The approach used by Ekasingh et al was replicated by Ngamsomsuke and Ekasingh to model crop choice in two other watersheds in Thailand, one in the north and another in the north-east, with the objective of verifying whether the model was applicable more widely [19]. Ekasingh and Ngamsomsuke then developed a simplified decision tree model in order to increase its understandability for policy makers, as well as for agricultural and natural resource scientists [20]. The simplified model incorporated new variables such as land type, water availability, tenure, capital, labour availability, non-farm income, and livestock. The resulting model was simpler but with a small reduction in accuracy.
Taking into consideration the state of the art in relation to the application of data mining in agriculture, this study uses data from the National Agricultural Survey to extract patterns that characterize households with low and high maize yields. This characterization could be useful for practitioners and policy makers in developing plans and drawing up policies to increase yields. The main contribution of this work is the use of data mining to characterize households based on maize yields. We did not find any similar work during the literature review.

3 Methodology

We followed the knowledge discovery from data (KDD) process [8, 9, 21]. KDD is the process of discovering interesting information from large amounts of data stored in databases, data warehouses, or other information repositories [9, 21, 22]. Data mining is the use of algorithms to extract information and patterns derived from the KDD process; it is a step in the entire KDD process [21]. The term data mining is used interchangeably to refer to the KDD process and the data mining step respectively. KDD is developed according to the following iterative steps: a) understanding the application domain and identifying the goals; b) selecting the dataset on which the discovery is to be performed; c) cleaning and pre-processing the dataset; d) reducing and projecting the dataset; e) matching the goals of the KDD process to a particular data mining method; f) exploratory analysis and selection of models and hypotheses; g) data mining; h) interpreting the mining pattern; i) reporting and using the information discovered. In the following paragraphs, we describe the application of the KDD process in the analysis of the National Agricultural Survey Data.

The understanding of the domain and identification of the goal for data mining were performed by reviewing documents related to agricultural development such as strategies and plans [1, 4] and studies of agriculture in Mozambique [2, 3, 6, 7], and also through interaction with domain experts.

For the analysis we used the TIA database for the years 2007 and 2008. The TIA database contains data on livestock and agricultural production, socioeconomic and demographic aspects relating to the farmers, use of technology, community infrastructure, loss of production, and calamities. This information is collected in national representative samples of small-, median-, and large-scale farms (see TIA [23] for more information on the classification of farms and survey sample). We selected data on small farming households related to maize production for the years 2007 and 2008. Small-scale farmers are the majority in Mozambique and maize is the main food crop [7, 24, 25]. The selected dataset after the cleaning process was composed of 8821 instances and 109 attributes including the class attribute. Most of the attributes were sparse (with several missing values) and unbalanced (with one category more common than others). The class attribute was binary and also unbalanced, with Class L (low yield) proving more frequent than Class H (high yield).

During the exploratory analysis of the dataset, we found that it contained outliers, errors and missing values. We did not correct for the outliers and the errors, but we handled the missing values using two approaches, by either a) removing instances
with missing values for the attribute maize yield, or b) replacing the missing values with a question mark (?) or a new value. The question mark replaced values that existed but were unknown due to some reason not given in the database. The new values were for those attributes where values were not given as a consequence of a previous answer, but where it was possible to establish their value. For example, for households that responded that they did not have a loss before harvest, we completed the empty value related to the attribute “reason for loss” using the value NotLoss. Where a loss had been suffered before harvest but no value was given for the attribute “reason for loss”, we completed the empty value using a question mark (?).

We transformed our target attribute for analysis (yield) into two categories or classes: Class H (high yield) and Class L (low yield). Households in Class H were those with a yield above 75% (Upper Quartile), corresponding to 961.21 kg/ha, and households with a yield less than this threshold were classified as Class L. The final data set comprised a table with instances representing each household and columns with variables (attributes) which characterized the instances.

The final dataset was put through the data mining step. We used Weka, a data mining and machine learning software package [26]. We selected Weka because it has a collection of machine learning algorithms for data mining tasks and pre-processing tools. In the data mining step, we applied tree algorithms, a decision tree, association rules, and classification rules, because these are suitable for descriptive tasks and generate outputs that are easy to understand. A detailed description of decision trees, association rules and classification rules can be found in the literature [9, 21, 22].

We gave the rules generated to a domain expert to establish their usefulness which also helped us with interpretation.

4 Results

In this section we present the results of the application of the KDD process using TIA data from 2007 and 2008.

Fig. 1 shows the classification rules generated by the JRip algorithm in Weka. At the end of each rule, the numbers of instances covered and the numbers of those incorrectly classified are presented in brackets; for example in rule R1, the numbers (221.0/52.0) mean that 221 instances are covered by this rule and 52 are classified incorrectly. The rules are labelled R1, R2, R3, R4, R5 and R6 and other variables in the rules have the following meanings: SoldQty is the quantity of the maize sold by the household; NrChikens is the number of chickens that the household had; NrGoats is the number of goats that the household had; EatReserve indicates if the household consumed maize reserved for seed during the period of hunger; LifeImproved indicates if the life of the household had improved compared to three years previously; NrSecondAct is the number of members of the household engaged in agriculture as a secondary activity; LossBeforeHarvest indicates if the household lost production before harvest; WillSell indicates if the household planned to sell more from their production; HasBike indicates if the household had a bicycle or not.
Fig. 1. Decision rules extracted from the dataset using the JRip algorithm

R1: (SoldQty >= 38.556) & (SoldQty >= 278.399994) & (SoldQty >= 650.325012) => Class=H (221.0/52.0)

R2: (SoldQty >= 20) & (SoldQty >= 174) & (NrChikens >= 7) & (EatReserve = No) & (NrSecondarAct >= 1) & (NrGoats >= 3) => Class=H (30.0/5.0)

R3: (SoldQty >= 50.025002) & (SoldQty >= 150.075012) & (LifeImproved = Better) & (RegularTransport = No) & (SoldQty >= 278.399994) => Class=H (87.0/33.0)

R4: (LifeImproved = Better) & (LossBeforeHarvest = No) & (WillSell = Yes) => Class=H (46.0/21.0)

R5: (SoldQty >= 50.025002) & (SoldQty >= 180.524994) & (LossBeforeHarvest = No) & (HasBike = No) => Class=H (68.0/27.0)

R6: => Class=L (8369.0/1891.0)

Fig. 1. Decision rules extracted from the dataset using the JRip algorithm

Fig. 2. Tree of association rules for Class H households

Fig. 2 presents the association rules in the form of a tree for Class H generated by the HotSpot algorithm [27] in Weka. Each branch represents rules that describe the association between Class H and related variables. For example, the topmost branches correspond to the rule Class H \(\rightarrow\) SoldQty\(\geq\)100 \& NrCashewTree \(\leq\) 28, which we can interpret as an association between the Class H yield and households who sold more than 100kgs of their products and had 28 cashew trees or fewer. This association tree was extracted using a support of 6%. The confidence is given by attribute in respective node of the tree. We noted that as we increased the support, the tree became simpler until we reached a rule Class \(\rightarrow\) H. Those variables on the tree and not defined previously have the following meanings: NrCashewTree is the number of cashew trees owned by the household; HHAge is the age of the head of the household; SellProd indicates whether the household commercialized part of their production.

Fig. 3 is the tree of association rules of households with a low maize yield. This rule was also generated with a support of 6%, which can be adjusted in accordance with the level of characterization that is required. Fig. 3 contains variables not defined previously which have the following meanings: EatMore indicates the food crops they consumed most during the period of hunger; HasSurplus indicates if the household had a surplus during the period of hunger; Surplus6 indicates if the household had a surplus in June (the index is the number of the months); Hunger indicates if the household passed a period of the year without food; ReasonLoss (1 or 2) indicates the first or second reason for loss of production; PartTimeWorker indicates if the house-
hold employed a part-time worker in agriculture; HHEdLevel is the education level of the household; NrCocoTree is the number of coconut trees owned by the household; and Km_TaredRoad indicates the distance in Km to the tarred road.

Fig. 3. Tree of association rules for the Class L households

The algorithm C4.5 generated rules similar to the association rules and classification rules presented above; however, its decision tree is complex and is therefore excluded from this report due to limitations of space, but will be the focus of further work.

5 Discussion

We extracted from the National Agricultural Survey Database rules that characterize households with high and low maize yields. The rules were generated using three different methods which serve to complement each other.

From Fig.2 it is possible to generate three distinct rules which associate the households with high yields with the commercialization of maize, the existence of agricultural assets, and household members with a capacity to generate income from off-farm activities (more young heads of the households).
The importance of commercialization and income for the development of agriculture is described by several actors [2, 4].

The association rule tree in Fig. 3 describes households with a low maize yield. This tree represents eight rules which associate Class L with a lack of surplus, loss of production, limited agricultural assets and low commercialization. In other words, households with low yields are those with low income from agricultural assets and production. Without income, the farmers are not able invest in agricultural technology, resulting in loss of production and food insecurity, forming a vicious circle between agricultural production and poverty [1, 4].

6 Conclusions and Future Work

In this study we used data mining to extract rules from the National Agricultural Survey database to characterize households with high maize yields and low maize yields. We used three algorithms: the decision tree, the association rule and the classification rule. The rules indicate that households with high yields are characterized by having the capacity to generate income through commercialization, agricultural assets, and to participate in off-farm activities. Commercialization was the most highlighted aspect in all the rules extracted. Households with low maize yields are the majority class and are strongly associated with food insecurity and loss of production. Low maize yield households do not commercialize their product and have limited agricultural assets.

One of the practical applications of these findings could lie in the development of policies to improve agricultural markets, promotion of off-farm activities, and the generation of cash crops and agricultural assets that would result in income to the households, which consequently would contribute to productivity. Experts in the area will be able to exploit the rules for further meanings and applications.

Another important output of this work is the representation of the results. Both decision trees and rules are easy to understand, giving the possibility to visualize how the variables and their values are associated with each specific class.

The C4.5 decision tree algorithm generated a complex tree which we decided not include in this work. Improving the representation of the results for the decision tree is proposed as one aspect of future research work.

Despite the fact that it was not the objective of this work, the rules extracted can serve to predict the class of the households given the respective attributes. The result of the evaluation of the models shows that Class L is predicted with high accuracy and Class H with low accuracy. Improving the accuracy of the model and finding representations to enhance their understandability are important tasks to be considered in future work. We also found during this work that the data contain outliers and some errors which we did not correct for the modelling because we wanted to see if those outliers represented interesting patterns. Improving the quality of the data, and integrating other related data in the model, such as weather data, price information, and other additional socioeconomic variables, could yield better results.
Notwithstanding the above and the fact that there is room for further developments, there is strong potential for the results obtained from this research to prove useful to both practitioners and policy makers.

Acknowledgments:
The authors thank the reviewers for their valuable suggestions on earlier drafts. The authors are also thankful to the Ministry of Agriculture Directorate of Economics for the TIA data and to all farmers who spent their precious time in responding to the survey. Special thanks also to Ms Ellen Payongayong for her valuable support.

References

3. Rafael N. Uaiene and Channing Arndt, FARM HOUSEHOLD EFFICIENCY IN MOZAMBIQUE, in International Association of Agricultural Economists. 2009: Beijing, China. p. 20.
APPENDIX C: Paper III
Factors Affecting the Use of Data Mining in Mozambique

Constantino SOTOMANE¹, Lars ASKER¹, Henrik BOSTRÖM¹, Venâncio MASSINGUE³
¹Stockholm University, Forum 100, SE-164 40 Kista, Sweden
Email: {sotomane, asker, henrik.bostrom}@dsv.su.se
²Ministry of Science and Technology, Maputo, Mozambique
Email: constantino.sotomane@mct.gov.mz
³Eduardo Mondlane University, Maputo, Mozambique
Email: venancio.massingue@uem.mz

Abstract: We present a study aimed at finding important factors that affect the acceptance and use of data mining in Mozambique. Input from potential users has been collected and analysed using a mix of qualitative and quantitative methods. The findings indicate that the level of adoption of data mining in Mozambique is primarily affected by poor quality of data, limited skills and human resources, limited support of stakeholders, organizational issues, limited financial resources and lack of adequate technology. These factors are similar to those identified in other studies.

Keywords: Data mining, success factors, Mozambique, data quality.

1. Introduction

Data Mining is an automatic or semi-automatic process of discovery of non-trivial and meaningful patterns in large volumes of data [1, 2]. The discovered patterns can lead to better decision-making in the organization [3]. With the advances of technology for computing, collection and storage of data, organizations are enabled to store data from a wide range of sources at a high rate, which leads to fast growth of volumes of stored data [4]. Data mining is used in many different domains, e.g., market analysis, fraud detection, customer retention, banking, retail, sales, manufacturing, health care, agriculture, and scientific exploration, to name just a few [4, 5].

The use of the data mining in Mozambique is infrequent, but there are good perspectives of increasing the use due to the need of the organizations to use decision making systems to enhance their competitiveness and productivity. For example, as part of Mozambique’s initiative to implement an E-Government strategy [6], several data-rich initiatives, such as the Government Data Centre, the State Financial Management System (SISTAFE) and the integrated platform of service to the citizen are in implementation. Those initiatives will increase the amount of on-line electronic data and in turn trigger the need for efficient tools for analysis and prediction. Data mining could provide such tools. The successful application of data mining is however not only depending on the quality of its algorithms but also influenced by several other factors [7-10].

The implementation of an emergent technology can be expensive and risky, and needs to be accompanied with adequate studies of key factors that are critical for its success [8, 11]. Knowing these factors helps to determine when and under which conditions it should be implemented. This study, aims to explore the level of knowledge and awareness of data mining in Mozambique and to identify critical factors affecting the successful application of
techniques and tools from the field of data mining. The factors to be identified are those that must be given special and continual attention throughout the whole process.

The rest of the paper is organized as follows: Section 2 presents related work, Section 3 presents the methodology used in the research, Section 4 presents the results from the study, Section 5 provides an analysis of the results, Section 6 presents business benefits and Section 7 finally, gives conclusions and outlines directions for future work.

2. Related Work

The study of critical factors and adoption of data mining is rather new and there are comparatively few studies targeting conditions relevant to Africa. This section presents a review of some of the few studies that have been presented on this topic.

Hilbert [8] carried out an empirical study involving 145 companies in Germany, and determined five success factors of data mining in organizations. These were: i) commitment by the top management, ii) existence of change management, iii) existence of financial resources, iv) alignment of IT with the business and v) quality of data. A team lead by Hart [12] investigated the issues affecting the adoption of data mining in South Africa and studied fifteen companies located in the Western Cape Province. They concluded that the adoption is affected by: i) management support, ii) cost, iii) availability of data warehouses, iv) data integrity, v) age of accumulated data, vi) teamwork, vii) skills, viii) integration of functions in the organization, and ix) company infrastructure. Four years later, Hart [7] studied the progress of data mining in South Africa by conducting three qualitative studies in large companies that had been involved in data mining and he found an increased awareness of data mining. He identified the following factors influencing the implementation of data mining: i) lack of organizational acceptance, ii) lack of understanding of data mining techniques, iii) insufficient data, iv) no use of data mining, v) lack of skill and vi) poor quality of data. Chen [13] studied critical factors for the implementation of data mining in China, and interviewed 86 enterprises and some graduates from science, engineering and IT. The following factors were found to be important: i) limited awareness and knowledge about data mining, ii) limited skills and iii) data issues. Sim [10] analysed data from 56 online respondents and found that the success of data mining was affected by the i) dataset factor (recovery of database, easy access to data, low complexity of data, scalability of database and maintenance of dataset). Cho et al [14] investigated data mining in Korean organizations. Based on an e-mail survey sent to 90 professionals in the field of data mining or data analysis, he concluded that the factors affecting data mining in Korea are i) data quality, ii) process documentation, iii) understanding of data, iv) data integration, v) data mining outsourcing strategy, vi) understanding of organizational data mining and vii) end-users request. Nemati and Barko [3] in their study found that the key success factors of the organizational data mining are: i) data quality, ii) data integration, iii) technical integration and expertise, iv) outsourcing strategy of data mining, v) skills, vi) time, vii) resource and viii) quality of the output. In a study by Nie [11], the following factors were identified as important: i) the quality of data, ii) existence of a multidisciplinary team, iii) existence of a sponsor, iv) financial resources and v) support of the top executives. A recent study conducted by Bole [15] found the following critical success factors: i) stakeholder support, ii) data quality, iii) focus on problem solving, iv) interpersonal skills, v) existence of business champions and vi) interdisciplinary learning.

To our knowledge, no study has focused on the conditions in Mozambique. The factors identified in different studies above may not be useful for explaining the success or failure of data mining adoption in Mozambique, because they are from different cultural and organizational contexts [14, 16]. Given that the application of data mining is still in its
infancy in Mozambique, it is of great importance to get an understanding of the specific conditions in the country to further guide the adoption process.

3. Methodology

This study uses a mix of qualitative and quantitative research methods with the objective to do triangulation [17, 18]. The qualitative data was collected using a focus group discussion and by open-ended questions in a questionnaire. The quantitative data was obtained by collecting fixed alternative answers to questions within the same questionnaire.

The respondents of the questionnaires were ICT users, data analysts and managers. The selection of the respondents was made by combining purposive and snowball sampling [19]. We selected people who are potential users or informed on data mining from public, private, education and research institutions in Mozambique. Each respondent was asked to forward the questionnaire to other informed or potential users of data mining. The questionnaire was distributed physically, by e-mail and by web, providing to the respondent a link of the online questionnaire, which was developed using limeSurvey [20]. In total, we received 110 responses of which 46 were considered non-valid due to insufficient data.

The questionnaire had both open and closed questions organized in three sections: a section with questions to collect information about the respondents and their organizations; a section with questions related to concepts, adoption and factors affecting the adoption of data mining in Mozambique and a section with questions about data quality.

The focus group discussion involved twelve people selected based on their experience from data analysis, ICT and management. The following topics were discussed in the focus group: i) Concept and level of adoption of data mining; ii) Factors influencing the adoption of data mining by organizations, and iii) Quality of data for data mining.

4. Findings of the Study

This section presents the results of the study. It is organized into four subsections: knowledge and awareness of data mining, the adoption of data mining in Mozambique, the factors influencing the adoption of data mining and quality of data.

4.1 Knowledge and Awareness of Data Mining

The participants of the focus group showed through their involvement in the discussions that in general they understand many of the data mining concepts. One participant asked for explanations about the concept of data mining and the meaning of the term “patterns”. This participant was from the area of data analysis.

From the perspective of the participants of the focus group, data mining is a technique of data analysis which goes beyond basic statistics and can discover patterns that cannot be found by standard statistical techniques. Useful patterns discovered by data mining must be interesting and original, as stated by one participant: “...good pattern must be something that nobody started thinking about it...”. The importance of access to massive amounts of good quality data was stressed. The quality of data is critical because, as stated by one of the participants, “...bad data generates bad results”, independently of the efficiency of the algorithm. Other characteristics of data mining considered by the participants were transparency and openness as well as representation of the input data and results. The process of sharing the result (patterns) with a domain expert in order to interpret and give meaning was considered to be an important step. The transparency and openness guarantees the reliability of the patterns through the evaluation. The representation of the input data means choosing good features to describe the data. The representation of the result means presenting the output patterns in a way that facilitates the interaction, understanding and interpretation by the users. One of the participants of the focus group, from the industry,
considered data mining as a tool for business intelligence. He defined it as a process of looking for information from the data to influence effective decision making in organizations.

The result of the survey shows that 57.8% of the respondents consider themselves knowing the concept of data mining and 48.4% were able to define it. It was typically defined as a process of finding patterns in (large) databases. The word "pattern" was associated with different forms of result, such as, association rules, clusters, classification models, or other representations of information or knowledge obtained depending on the algorithm used.

Some of the respondents obviously had their own, non-standard, interpretation of the concept, such as considering data mining to be any data analysis process, including what is done manually. For example, one of the respondents said: "I use data mining because always, I do manual comparison between documents looking for consistencies".

4.2 Data Mining Adoption in Mozambique

The participants of the focus group expressed that a majority of organizations in Mozambique are small and perform only basic data analysis. The potential number of data mining users is large and they may be found at well-established companies, such as banks and telecommunications companies, due to the complexity of their businesses and the large volumes of data that they collect and maintain. The public sector was considered by the participants as one area in which data mining could be used for improving decision making and management. In general, the participants felt that the use of data mining is likely to increase in the future, because of the rapid increase of data combined with harder competition on the market. One participant said: "... if in the future a company does not do data analysis, it will surely die ..."

The result of the survey indicates that only 12.5% respondents had experience of data mining, which varied from 6 months of experience to 5 years. All these 12.5% respondents were from the area of IT and had at least a BSc degree.

4.3 Factors Influencing the Adoption of Data Mining

The participants of the focus group were unanimous in stating that data mining is hardly used in Mozambique. The following reasons for this were identified:

- Limited knowledge of data mining – Experts in data mining are scarce and expensive.
- Weak awareness about data mining – Weak awareness about the potential of data mining and of data as source of knowledge for decision making. One of participants highlighted: "when people realize the importance of data, then they will start taking care the data and doing appropriate analysis…"
- Resistance to change – some people do not want to adapt to the new technology.
- Complexity and terminology – data mining requires expertise and knowledge of complex algorithms, e.g., neural networks and support vector machines.
- Lack of technology – data mining requires specialized software. Some organizations have limited access to databases, data warehouses and tools for data analysis.
- Data quality and availability – in general the quality of data is not good and do not integrate information from other areas. It is difficult to get access to external data.
- Sizes of the companies – most organizations are small and with limited capabilities for data analysis.
- Use of the data mining result – there is no motivation for complex data analysis because results of existing data analyses are not used. To highlight the weak culture of making
decisions based on data one of participant said “how many of us sees weather forecast before leaving home? Nobody…”

The perception of the participants of the focus group is that the use of data mining will grow due to the increased amounts of data that is collected and stored in organizations, needing proper analysis to extract information to support efficient decision making.

The respondents of the survey in general considered that the factors that affect the adoption of data mining are the following: limited knowledge (42.19%), lack of human resource (26.56%), no need for data mining application (17.19%), insufficiency of data for data mining (7.81%) and acceptance in organization (7.8%). Those 12.5% of respondents that use data mining consider as problems encountered when developing data mining projects the followings: Poor quality of data (62.5%), insufficient data for data mining (62.5%), inadequate technology (50%), cost of the data mining (37.5%), limited knowledge about the data (37.5%), lack of human resources in the area of IT and statistics (25%), limited knowledge about the organization (25%), limited support of top management (25%) and non-acceptance of data mining in the organization (12.5%).

4.4 Quality of Data

The participants of the focus group were unanimous that the quality of data in Mozambique is low. The reasons for this are described below:

- Data collection process – The data collection process does not follow good practice in terms of methodology, sampling, coding and naming. Referring to this, one of the participants mentioned that some organizations have data from several years that are not useful because of bad quality due to bad collection process.

- Historical situation of the country – The shortage of qualified resources after the independence, the impossibility of collecting data along the country during the 16 years of civil war, unavailability of resources for data collection and other limitations affects the quality and availability of data.

- Technology for data collection and storage – not all organizations have databases, even fewer have a data warehouse. Some information is kept on paper only. Lack of appropriate equipment for accurate data collection. Another related problem is the limited knowledge about methods and techniques for correction and quality control of data.

- Social aspects – The data collectors are rewarded or penalized in accordance to the amount of data collected, and as a consequence, there are cases of “invention of data” to reach the target. Some data collectors do not have enough time to collect data because of other work.

- Human resources – There is a shortage of human resources with skills for data collection and analysis. As a consequence, the data collection is not preceded with the appropriate design of instruments, the collection process does not follow good practises and there is not adequate quality control. The lack of skilled human resources leads to inadequate data collection.

- Uses of data – People, in particular nationals, do not use data for analysis. For this reason, the data management departments do not know the user demand in terms of quantity, quality and level of aggregation. Since people do not use the data, they do not know if the quality is good or not.

Related to the quality of the data, 65.62% of the respondents of the questionnaire were undecided, 31.2% considered it to be good and 3.1% considered it to be very good. None of the participants considered the quality of data to be bad.
5. Discussion

Data mining is still relatively unknown in Mozambique. This is especially true for areas outside of IT. Some of the participants in the study believed data mining to be any data analysis process, including manual ones, with the aim of finding interesting patterns in data. Only 12.5% of the participants of the study had any experience from data mining. The perception that data mining is more commonly used by big companies, such as banks or telecom companies, was not confirmed in any of the six big organizations which we contacted. These findings are not surprising but instead in line with what can be expected to be the case for most African countries. Bole [15] explain this as an effect of the large distance between data mining and business.

Table 1 below, presents the factors affecting the adoption of data mining as identified by previous studies, grouped into six clusters (factor groups), to facilitate the comparison. In Table 2, the factor groups are related to each study, where Stdy [N] refers to reference no. N in the list of references of this paper.

Table 1: Data Mining Factors Organized in Data Mining Factors Dimensions

<table>
<thead>
<tr>
<th>Stakeholder support issues</th>
<th>Data issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Existence of business champion</td>
<td>Quality of data</td>
</tr>
<tr>
<td>Commitment of top management</td>
<td>Availability of data warehouse</td>
</tr>
<tr>
<td>Existence of change management</td>
<td>Data integration</td>
</tr>
<tr>
<td>Non organizational acceptance</td>
<td>Age of data</td>
</tr>
<tr>
<td>Not use of data mining</td>
<td>Insufficient data</td>
</tr>
<tr>
<td>User request</td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Organizational issues</th>
<th>Skills and human resource issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>Functional integration</td>
<td>Existence of teamwork</td>
</tr>
<tr>
<td>Alignment of IT and business</td>
<td>Existence of skill</td>
</tr>
<tr>
<td>Strategy of outsourcing</td>
<td>Understanding of data mining and techniques</td>
</tr>
<tr>
<td>Interdisciplinary learning</td>
<td>Awareness of data mining</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Financial issues</th>
<th>Technology issues</th>
</tr>
</thead>
<tbody>
<tr>
<td>costs</td>
<td>Existence of adequate software and tools</td>
</tr>
<tr>
<td></td>
<td>Existence of database</td>
</tr>
<tr>
<td></td>
<td>Existence of data warehouse</td>
</tr>
</tbody>
</table>

Factors from all groups were mentioned in the data collected by the quantitative and qualitative methods employed in the current study which means that all are relevant also for Mozambique. It is evident that the quality of data and the human skills are very important factors for the success of data mining as also indicated by [11] and [15] respectively, likewise, the support of the project by stakeholder is important. Many projects terminate without success because of absence of this kind of support, in particular the organizational and top management support.

Table 2: Result of Similar Studies (Study). Adopted from[15]

<table>
<thead>
<tr>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>4</td>
<td>Data issues</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>3</td>
<td>Skills &amp; human resource issues</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>1</td>
<td>Stakeholder support issues</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>6</td>
<td>Organizational issues</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>2</td>
<td>Financial issues</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
<tr>
<td>5</td>
<td>Technology issues</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
<td>x</td>
</tr>
</tbody>
</table>

Participants of the focus group discussion were unanimous considering the issue of poor data quality. This perception is supported by several studies on the quality of data in Mozambique, e.g., [21-23], and applications of data mining in Mozambique [24, 25]. Most respondents were however neutral and some considered that the quality of data was good.
This controversial perception of the respondents indicates their limited knowledge about the data, as mentioned by one of the participants of the focus group “… most of people do not know the situation of the data because they do not use…”

The result of this study is based on the perception of a limited group of experts located in Maputo, which possibly makes it non-representative for the rest of the country. It is however our opinion that it is likely that most of the findings are applicable to any region of the country.

6. Business Benefits

The benefit of this research is identification of the main factors affecting the adoption of data mining in organizations in Mozambique. These factors can be used to ensure the success of data mining in organizations and to decide about the viability of implementing it, reducing the risk of failure and financial loss. Data mining is an important tool to leverage the productivity of organizations [3], which can result in socio-economic benefits. If an application of data mining can result in an accurate forecast model of electrical consumption, it will provide financial, social and environmental benefit for the country.

7. Conclusions and Future Work

This study explored the perception of potential practitioners around four issues related to data mining in Mozambique: level of knowledge and awareness, level of adoption, factors of adoption and the quality of data. The main finding is the level of knowledge, awareness and adoption of data mining are very low. Few people use data mining and those that do are mainly working within the IT sector. The participants of the study perceive that the following factors influence the adoption of data mining in Mozambique: poor quality of data, limited skills and human resources, limited support of stakeholders, organizational issues, limited financial resources and lack of adequate technology. The quality of data is considered to be low and represents one of the main inhibitors for the adoption of data mining. Despite all, the future perspective of adoption of data mining is good, given the increase of data and the need for better decision making. There are several areas were data mining can make a contribution in Mozambique. Examples include flood forecasting, yield prediction for various crops, disease outbreak prediction, malaria prevalence and prediction, HIV prevalence and prediction and electricity forecasting. To ensure an effective adoption of data mining, it is necessary to work towards improvement of the quality of data and its availability in integrated databases. It is also important to create awareness of the potential of data mining in order to enable stronger support from stakeholders, acceptance in the organization and allocation of resources. Another important recommendation is the development of human resources in the area of data analysis and data mining.

To better enlighten the issues related to the adoption of data mining in Mozambique, we recommended that studies are conducted within each specific sector in order to better explore the factors of the adoptions. It is also important to realize studies to measure the readiness and evaluate the risk of the implementation of data mining for each organization.

Acknowledgement: We would like to thank the respondents of the questionnaire and the participants in the focus group for their invaluable collaboration. We thank STIFIMO for organizing the focus group as part of their weekly research event. We thank Ander Moreno for helping in the development of online questionnaire, and finally we wish to extend our gratitude to Dr. Erkki Sutinen, Dr. Gil Gonçalves, Dr. Getrudes Macueve and Dr. Jordi Gallego-Ayala for reviewing and commenting the early version of this paper. We thank SIDA/SAREC program at UEM for supporting this research.

References


APPENDIX D: Paper IV
Short-term Forecasting of Electricity Consumption in Maputo

Constantino Sotomane\textsuperscript{1,2}, Lars Asker\textsuperscript{1}, Henrik Boström\textsuperscript{1}, Venâncio Massingue\textsuperscript{3}

\textsuperscript{1}Stockholm University, Department of Computer and Systems Science, Sweden  
\textsuperscript{2}Ministry Of Science and Technology, Mozambique  
\textsuperscript{3}Eduardo Mondlane University, Mozambique

Sotomane@dsv.su.se, Asker@dsv.su.se, henrik.bostrom@dsv.su.se, Venancio.massingue@uem.ms

Abstract. We present a short-term load forecasting model for Maputo. The model is based on the concept of multiple models. A clustering method is combined with expert’s knowledge to identify sub-models. The resulting model, which is the combination of several sub-models, is evaluated and compared to the model currently used by the Electricidade de Moçambique E.P (EDM). The results show that the developed model performs better accuracy than the one currently used by EDM. The results obtained by the application of the model when translated into financial figures demonstrate significant economic advantages. The social and environmental implications of the model are also analysed.

Keywords: Short-term load forecasting, Day-Ahead-Market, robust regression, multiple models, clustering, Mozambique.

I. INTRODUCTION

Load forecasting is important for management and operation in the electricity industry [1-3]. There are three types of load forecasting models: long-term, mid-term and short-term. Long-term load forecasting corresponds to predicting load one year ahead and is used for planning, e.g., the planning of investments in new capacity [4, 5]; mid-term load forecasting corresponds to a period of months to days and is used to estimate the medium-term load requirements, such as peaks during summer or winter periods [4]; while short-term load forecasting corresponds to the period from hours to days and is important for management and operation of power systems and the electricity market [6].

Short-term load forecasting is the basis for several operations activities such as scheduling the capacity of generation, reliability analysis, market assessment and maintenance plans [1, 4, 6]. With the deregulation of the electricity sector, short-term load forecasting has become more critical due to its importance for planning of electricity transactions in the competitive market [1, 3, 4, 6]. The importance and complexity of short-term load forecasting will continue to increase, due to supply-demand fluctuation, changes in prices of electricity, changes in weather and the high financial penalties resulting from forecasting errors [1, 5].

The structure of the electricity sector has been changing in the last decades all over the world, from monopoly to deregulated and competitive market structures [5]. As part of the restructuring process, the member states of the Southern African Development Community (SADC) created in 1995 the Southern Africa Power Pool (SAPP) “with primary aim to provide reliable and economical electricity to supply to the consumers of each of the SAPP members, consistent with reasonable utilization of natural resources and the effect on the environment” [7].

Since 2009, SAPP is operating a competitive energy market for SADC in the form of a Day-Ahead-Market (DAM), where bids for electricity trading are submitted one day before [5, 8]. The Electricidade de Moçambique E.P (EDM), the main supplier of electricity in the country, is member of SAPP and also must send its bids one day in advance in order to participate in the market. To be able to send the bids a day in advance, it is necessary to forecast the demand of the internal consumers and based on that, to infer the amount of electricity for DAM.

EDM forecasts the hourly maximum consumption for the next day and based on that, estimates the amount of electricity for the DAM. This procedure is not optimal since it underestimates the electricity for DAM, resulting in loss of electricity that is not consumed neither sold. The aim of this work is to build a short-term load forecasting model to improve the accuracy of electricity estimated to DAM.

The electricity forecast is done traditionally using statistical methods such as linear or multiple regression, box Jenkins and others, but it has been observed that these techniques are deficient when there are abrupt changes in environmental or sociological variables [1, 2, 9]. In this work, we use the concept of multiple-models where the forecasting model is composed of a set of sub-models. Each sub-model is used to forecasts the consumption’s demand of certain type of days, for example the first sub-model can be used to forecast working days of summer, the second sub-model working days of winter and so on, the overall model is the combination of all sub-models. The concept of multiple models assumes that the data to be modelled are generated by the mixture of models and has been used successfully to minimize the effect of abrupt change of the values of the attributes of the dataset [10-12]. Clustering is combined with domain knowledge in the identification of sub-models.

The rest of the paper is organized as follows. Section 2 describes the data and methods utilized in this study. The results are presented in section 3 and are discussed in section 4. Finally, the conclusions and directions for future work are outlined in section 5.

II. DATA AND METHODS

This section describes the dataset employed in the study, how the short-term load forecasting model was built, the method currently used by EDM to forecast electricity consumption and how to estimate DAM. The section also
Short-term Forecasting of Electricity Consumption in Maputo describes how the amount of electricity is converted into financial value.

A. The dataset

To build the forecast model we used information from several sources:

- Electricity consumption of Maputo – Measurements of hourly electricity consumption of the period from January 2003 to October 2012;
- Historical temperature in Maputo - The maximum and minimum daily average temperature in Maputo city for the period of January 2003 to October 2012;
- Season - Mozambique has mainly two seasons: the dry and cold season (winter) from April to September and wet and hot season (summer) from October to March;
- Type of day: Holiday, working day, Saturday or Sunday.

This information was integrated, resulting in a dataset with 29 attributes and 3592 instances. Each instance represents a single day. The attributes are date, hourly consumption (24 attributes), maximum temperature, minimum temperature, season and type of day. The dataset was split into a training set, corresponding to January 2003 to December 2011 (3287 instances) and a test set corresponding from January to October 2012 (305 instances).

To understand the data, following techniques were used:

- interaction with expert of the area of the electricity;
- Plotting (visualization) the electricity data and temperature data to determine the trends, behaviours and relation between attributes;
- Clustering the electricity data with the objective to describe them;

Both electricity and temperature data contained errors. The electricity consumption data were characterized by defective load curves due to high or low consumptions values. The defective days were replaced with the average of the past three similar days of the week, for example a defective Monday was replaced by the average of the last three Mondays. The same procedure was used for other days of the week.

The temperature data were characterized by missing values. The missing temperatures values were replaced by average temperatures of the month. For example if the minimum temperature of 15th of August was missing, the average minimum temperature of the month August was used as the temperature of 15th of August. The same procedure served to handle missing values for maximum temperatures.

The training and test set were corrected separately.

B. Short Term Load Forecasting Model

The consumption of electricity is non-linear; it changes with weather, season and type of day. This means that the forecast model designed for specific condition could not be accurate for different condition. Based on this assumption, five sub-models were designed, each one to predict electricity consumption for a specific type of the day. This approach is called multiple models [10-13]. The idea of using the multiple models is to decompose a complex system in simple sub-systems to facilitate the modelling and improve accuracy.

The five types of days are: working days of summer, working days of winter, Saturdays, Sundays and Holidays. They were identified by a combination of clustering of electricity consumption data and the knowledge of the experts of EDM. For clustering, the EM algorithm of Weka’s workbench was used with the parameter number of cluster set to 3 and the electricity consumption as input data. The number of clusters was determined after several iterations.

The clustering identified three types of days: working days of summer, working days of winter and non-working days. The experts of the EDM suggested to split the cluster non-working days in three types of days: Saturdays, Sundays and Holidays. The combination of clustering and domain knowledge resulted in five types of days, early mentioned, each one corresponding to one sub-model. Other approaches for identifying the sub-models could be used, see e.g., [10-13].

Once each sub-model is composed by the same type of days, is assumed to be linear and possible to mode with linear model. The robust regression model was used to build each sub-model. Other algorithms such as artificial neural networks [14, 15] and SVMs [15] could be used to build the sub-models. The robust regression model [6] was selected due to its relative simplicity and robustness in the presence of outlier.

Equation 1 shows the demand forecasting model based on the concept of multiple models. Df is the demand to forecast, t+1 is the day to forecast, ci is the type of the day (where i is the index of the sub-model), D is the actual demand, t is the actual day, b1, b2 and b3 are the regression coefficients, Tm is the prediction of maximum temperature of the day to forecast and Tm is the prediction of minimum temperature of the day to forecast.

$$D_{c_i}^{t+1} = b_{1c_i}D_{c_i}^{t} + b_{2c_i}T_{M}^{t+1} + b_{3c_i}T_{m}^{t+1}$$  (1)

Data from January 2003 to December 2011 were used to train the sub-models of equation 1 and data from January 2012 to October 2012 were used to test the models.

The accuracy of the model is evaluated by the Mean Absolute Error (MAE) given by the equation 2.

$$MAE = \frac{1}{N}\sum_{i=1}^{N}|P_k - (D_r + DAM)|$$  (2)

Where N is the number of data points, Pk is the available energy, Dr is real demand and DAM is forecasted DAM. Ideally Pk must be equal to Dr+DAM.

C. Short-term load forecasting: The EDM Model

Currently EDM forecasts electricity for the next day by calculating the maximum consumption for each hour based on the previous 12 days of the week. For example, if the day to be forecasted is Monday, the 12 previous Mondays are used to find the maxims value of each hour.

D. Estimation of DAM

Equations 3 and 4, provided by EDM, describe the process of estimating the electricity to be sold in the Day-Ahead-Market (DAM). The amount of electricity to be sold to DAM should be what remain after satisfying the internal needs.

$$DAM = round(P_k - D_f) \times 0.90$$  (3)

$$DAM = 0 \text{ if } (P_k - D_f) \leq 0$$  (4)
In equation 3, \( P_k = 350\text{MW} \) is the available electricity, acquired from suppliers based on contracts. \( D_f \) is the forecasted demand for the next day. \( \text{round()} \) is a function that rounds down (e.g. 3.4=3 and 2.8=2) the difference between \( P_k \) and \( D_f \). The coefficient 0.90, was defined by EDM, serves to reduce the result of the difference between \( P_k \) and \( D_f \) in order to minimize the risk of committing to DAM more energy than available. Equation 4 shows that when the forecasted demand exceeds the available electricity, no energy will be sold to EDM.

E. Estimation of financial values

The financial values equivalent to the electricity available for DAM were obtained by multiplying the amount of electricity for DAM with the average market clearing price (MCP) given in USD. Because the values of MCP were obtained as monthly average, the values of DAM were aggregated per month before the multiplication. The values of MCP for the period January to October 2012 were obtained from the report of SAPP [8, 16].

The financial values equivalent to the underestimated electricity for DAM are calculated by multiplying the tariff at which the EDM would buy the electricity at that particular time of day, by the amount of electricity underestimated. The tariff varies according to season and time of day (peak hour, standard hour and off-peak hour). The season of June and August is the most expensive because has the highest demand [16]. The peak hours are the most expensive and off-peak hours are the cheapest. The tariffs used in this study were obtained from the catalogue of ESKOM for years 2012/2013 [17].

The tariff for the calculation of the financial values equivalent to the overestimation depends on the amount of electricity overestimated. If the amount overestimated is above 15% of the available electricity (\( P_k \)), the emergency tariff is used. The emergency tariff is the highest tariff and varies in accordance to the period and the time of day. If the overestimated amount is less than 15% of the available electricity, is used the same tariff that EDM would buy electricity, multiplied by two. This tariff depends also on the time of day. For this study the emergency tariff was provided by EDM.

III. RESULTS

This section presents the results of estimating the DAM (see section D) using the two different models for demand forecasting; the model currently used by EDM (described in section C), and the new model presented in this paper (described in section B). For reference, in this paper we will call them EDM and RR (robust regression) respectively. Both models were trained using data from January 2003 to December 2011 and tested using data from January 2012 to October 2012.

Figure 1 shows the result of the automatic clustering of the electricity consumption data with weka’s workbench, this result was obtained during the pre-processing with the objective to describe the electrical consumption data. The clustering of figure 1 was used by the domain expert to determine the five types of days, which were used to build sub-models of Model RR: sub-models for winter working day, summer working days, Saturdays, Sundays and Holidays.

Figure 2 shows the comparison between the electricity consumption and the temperature during the year. It is noted that the consumption of electricity increase when the temperature increases and decrease when the temperature decrease. This correlation determined the selection of the temperature as the variable for the forecasting model.

Figure 3 presents the DAM estimated using model EDM (DAMedm), DAM estimated using the model RR (DAMrr), the real demand (Dr), the total consumption for the EDM (Dr+DAMedm), the total consumption for the model RR (Dr+DAMrr) and the available electricity (Pk). The graph in Figure 3 shows only the first week of June and was selected arbitrarily from the result of applying both models to the test set. The stacked bar chart of figure 4 compares the underestimated electricity by the model EDM (black portion) and by the Model RR (not colored portion) per month. The stacked bar chart of figure 5 compares the overestimated electricity by the model EDM (black portion) and by the Model RR (not colored portion).

Table 1 presents the Mean Absolute Error (MAE) for the evaluation of Models EDM and RR respectively. Table 2 presents the monthly revenues obtained by applying the models DAMedm and DAMrr, together with the costs resulting in the overestimation by the model DAMedm and DAMrr respectively. The last column of Table 2 is the
Short-term Forecasting of Electricity Consumption in Maputo

difference between the revenues by the model DAMrr and DAMedm after deducting the cost related to overestimation. For example the difference of the month of January is calculated as following:

\[
(3,434,371.10 - 247,866.84) - (2,625,799.30 - 127,063.00) = 687,767.96.
\]

The model RR loss less electricity than the model EDM.

### Table 1: Mean Absolute Error (MAE) of models EDM and Model RR respectively

<table>
<thead>
<tr>
<th>Month</th>
<th>Revenue EDM</th>
<th>Cost EDM</th>
<th>Overestimate EDM</th>
<th>Revenue RR</th>
<th>Cost RR</th>
<th>Overestimate RR</th>
<th>Difference</th>
</tr>
</thead>
<tbody>
<tr>
<td>Jan</td>
<td>2,625,799.30</td>
<td>127,063.00</td>
<td>2,792,994.15</td>
<td>2,792,994.15</td>
<td>8,652,883.20</td>
<td>8,652,883.20</td>
<td>151,960.80</td>
</tr>
<tr>
<td>Feb</td>
<td>2,369,497.20</td>
<td>112,581.43</td>
<td>2,482,078.63</td>
<td>2,482,078.63</td>
<td>8,482,078.63</td>
<td>8,482,078.63</td>
<td>497,364.17</td>
</tr>
<tr>
<td>Mar</td>
<td>1,970,771.30</td>
<td>173,993.23</td>
<td>2,144,764.53</td>
<td>2,144,764.53</td>
<td>8,344,764.53</td>
<td>8,344,764.53</td>
<td>497,364.17</td>
</tr>
<tr>
<td>Apr</td>
<td>2,671,978.80</td>
<td>92,806.40</td>
<td>2,764,785.20</td>
<td>2,764,785.20</td>
<td>8,564,785.20</td>
<td>8,564,785.20</td>
<td>497,364.17</td>
</tr>
<tr>
<td>May</td>
<td>3,052,263.80</td>
<td>89,834.37</td>
<td>3,142,098.17</td>
<td>3,142,098.17</td>
<td>8,342,098.17</td>
<td>8,342,098.17</td>
<td>497,364.17</td>
</tr>
<tr>
<td>Jun</td>
<td>3,063,529.90</td>
<td>178,699.01</td>
<td>3,242,228.91</td>
<td>3,242,228.91</td>
<td>8,242,228.91</td>
<td>8,242,228.91</td>
<td>497,364.17</td>
</tr>
<tr>
<td>Jul</td>
<td>4,831,463.40</td>
<td>250,340.74</td>
<td>5,081,804.14</td>
<td>5,081,804.14</td>
<td>8,881,804.14</td>
<td>8,881,804.14</td>
<td>497,364.17</td>
</tr>
<tr>
<td>Aug</td>
<td>4,683,128.40</td>
<td>311,088.03</td>
<td>4,994,216.43</td>
<td>4,994,216.43</td>
<td>8,894,216.43</td>
<td>8,894,216.43</td>
<td>497,364.17</td>
</tr>
<tr>
<td>Sep</td>
<td>2,701,306.40</td>
<td>60,413.27</td>
<td>2,761,719.67</td>
<td>2,761,719.67</td>
<td>8,561,719.67</td>
<td>8,561,719.67</td>
<td>497,364.17</td>
</tr>
<tr>
<td>Oct</td>
<td>1,766,400.50</td>
<td>129,504.08</td>
<td>1,895,904.58</td>
<td>1,895,904.58</td>
<td>8,495,904.58</td>
<td>8,495,904.58</td>
<td>497,364.17</td>
</tr>
<tr>
<td>Total</td>
<td>29,736,139.00</td>
<td>1,525,423.55</td>
<td>31,261,562.55</td>
<td>31,261,562.55</td>
<td>8,716,562.55</td>
<td>8,716,562.55</td>
<td>497,364.17</td>
</tr>
</tbody>
</table>

The results show that the RR model is better than the EDM model. The RR model allocates more electricity for DAM and reduces the amount of electricity lost due to underestimation. From table 2 it is observed that the application of the RR model increase the revenues in USD 8,652,883.20 in relation to the EDM model during the period January 2012 to October 2012.

In Fig. 3, it is noted that the area between the total consumption of the RR model and the available electricity is smaller than the area between the total consumption of the EDM model and the available energy. This means that the RR model is able to forecast the DAM with more accuracy than the EDM model. The curacy of the RR model is also confirmed by the results of the MAE. The MAE of RR model is less than of EDM Model as shown in table 1.

In Fig. 4, is observed that the RR model overestimates more electricity for DAM than the EDM model. This result is not desirable, because the overestimation is expensive, since the electricity for replacement is purchased at high price because is non-planned purchase. For the present case, the effect of this unwanted result is small and is compensated by the good estimation of DAM by the RR model. The results presented in table 2 show that the good estimation of DAM by RR model compensates the slight increase of overestimation of electricity.

The use of clustering to describe the data, provided insight to the domain expert, and facilitated the selection of type of days for the construction of sub-models. This is an example of how data mining can contribute to better decision making.

**IV. EVALUATION AND DISCUSSION**

**F. Socio-economic implications**

The good accuracy obtained by the model RR contributes with socioeconomic benefits. The first benefit is for the electricity company which increases the revenue by selling more electricity and reduces the loss due to the lower underestimation. The increase of income of the company enhances its capability to invest in electrical infrastructure to serve more efficiently and effectively. The expansion of the electricity infrastructure is a key objective in Mozambique, because despite a huge potential in energy resources, only 18% of the population has access to electricity [18]. The reduction of losses can potentially influence the reduction of electricity prices for end-users, making it more affordable.

The result also contributes to the efforts to reduce adverse impact on the environment created by the use of...
firewood based energy and charcoal due to limited access and high cost of electricity. Cuvilas et al [19] indicate that 81% of the energy used in Mozambique is wood fuel and there are increasing use of charcoal both in rural and in urban area which contributes to the deforestation.

One of the implications of the results is the development of price policies and awareness campaigns to promote the use of electricity more rationally and encourage more use of electricity during the time of day when it is cheaper. This type of actions can reduce the over-consumption of energy during peak hours and make more reserves for DAM.

G. Limitations

Despite the above reported advantages, the model RR has the drawback of increasing the level of overestimation. Compared to the overall contribution of the model RR, the effect of the increase of the overestimation is small; nevertheless, improvement of the model RR is desired to reduce this drawback and to increase its accuracy.

V. CONCLUSION AND FUTURE WORK

This study presents a new short-term load forecast model for the city of Maputo, Mozambique. The model is compared with the one currently used by the EDM. The results shows that the new model (RR) is more accurate than the one used by EDM, allowing for more electricity to be allocated to DAM, which reduces the loss of electricity due to underestimation. The model RR has the drawback of slightly increasing the amount of overestimated electricity, but this small negative effect is compensated by the increases in revenue created by the model RR. The analysis of the results indicated that the increase in the amounts of electricity to sell and the savings of electricity obtained by applying the model RR can contribute to effort for widening the access to electricity and consequently reducing adverse effects on the environment created by the use of energy from charcoal and wood fuel.

The results show how data mining can benefit organizations, in this case, the clustering method helped to determine sub-models used to build a more accurate load forecasting model which in turn served to decide the amount of electricity to sell to DAM. The results may contribute to the increase in revenue and reduction of the adverse effects on the environment. Future work includes improving the model in order to further enhance its accuracy and solve the problem of increasing the over-consumed electricity. The integration of more data such as environmental, socioeconomic and regarding the operation may help to increase the accuracy of the future forecasting models.

VI. ACKNOWLEDGEMENTS

The first author acknowledges the support of the SIDA/SAREC at UEM, the EDM in special Mr. Sandro Chiang for valuable support and advices during the development of this work and the national institute of metrology for providing the temperature data.

REFERENCES


2013 International Conference on Advances in ICT for Emerging Regions (ICTer) 12th & 13th December 2013