Data Mining Process Models:
A roadmap for knowledge discovery

Armando B. Mendes, Universidade Açores, Portugal, amendes@uac.pt
Luís Cavique, Universidade Aberta, Portugal, lcavique@univ-ab.pt
Jorge M.A. Santos, Universidade Évora, Portugal, jmas@uevora.pt

1. Introduction

Extracting knowledge from data is the major objective of any data analysis process, including the ones developed in several sciences as statistics and quantitative methods, data base \ data warehouse and data mining.

From the latter disciplines the data mining is the most ambitious because intends to analyse and extract knowledge from massive often badly structured data with many specific objectives. It is also used for relational data base data, network data, text data, log file data, and data in many other forms.

In this way, is no surprise that a myriad of applications and methodologies have been and are being developed and applied for data analysis functions, where CRISP-DM (cross industry standard process for data mining) and SEMMA (sample, explore, modify, model, assessment) are two examples.

The need for a roadmap is, therefore, highly recognised in the field and almost every software company has established their own process model.

The data mining community and allied fields went through a long process of decomposing the tasks involved and many alternatives were developed in the process. These resources can be best utilized when we systematically analyse which of the tasks should be accomplished and which methods are best for each task and for each phase.

Traditional statistics companies tend to overvalue the analysis phase. This is obviously one of the core phases with a large number of methods developed to handle many search and model building tasks. But, hypothesis test and inference classification and forecast are only a part of the whole process of knowledge discovery, and often a short one of that. The entire process must include many more phases as is detailed in this chapter. We will walk through

each phase, outlining typical tasks and methods involved and present cases of application.

Many of these methodologies recognize the iterative foundation of the process with many loops connecting the phases through the same application domain and feeds back on earlier findings. This adaptive development process is not new to data analysis literature.) describes continuous actions cycles that involved significant user participation. As each cycle is completed, the system gets closer to its established state like an evolution spiral ().

In this chapter we present in some detail the CRISP-DM process model. The Cross Industry Standard Process for Data Mining (CRISP-DM) methodology, working over the successful results of those earlier experiences developed a very well specified process model where even the user intervention or domain knowledge was not forgotten. In fact, as noticed, data warehouse and data mining development are dominated by IT departments. As IT professionals have little experience with quantitative modelling, some basic concepts of data models have recently being rediscovered like evolutionary development.

On the other hand, note that, compared to the traditional manual analysis, the process models supply a much higher degree of system autonomy, especially in processing large hypothesis spaces. However at the current state of the art, a human analyst still makes many decisions in the course of a discovery process. Is also important to note that, using the decision support paradigm, the intention is not to automate the process but to help the analyst and decision maker to use their intuition in understanding the dynamic, using domain knowledge and knowledge extract from data for ultimately making decisions and manage problems.

The chapter is organized in the following sections. In section 2 CRISP-DM Process Model is presented. In section 3 we point some authors' studied cases. In section 4 the EDA (Electricidade dos Açores) cases are developed. Finally, in section 5 we draw some conclusions.

2. CRISP-DM Process Model

Several authors have been suggesting process models for knowledge extraction from data bases (*e.g*. Klösgen and Żytkow, 2002; Hand *et al.*, 2001; Fayyad *et al*., 1996). In spite of that, the CRISP-DM has been progressively becoming more visible, as the users recognize it as well structured and practical. It has also being validated by successful stories in projects of substantial dimension. The initiative that lead to CRISP-DM was conducted by companies in software development, consulting firms, as well as clients of data mining, with the objective of becoming independent of the business sector as well as of the software application (Clifton and Thuraisingham, 2001; Lavrač, *et al*., 2004).

In Figure 1 are depicted the six phases of the CRISP-DM process model version 1.

1. Business Understanding
2. Data Understanding
3. Data Preparation
4. Modelling
5. Evaluation
6. Deployment

This process model is described in detail by the authors in Chapman *et al*. (2000) or by one of the companies involved in the development of the process model). An explanation of each phase is expressed in the following paragraphs.

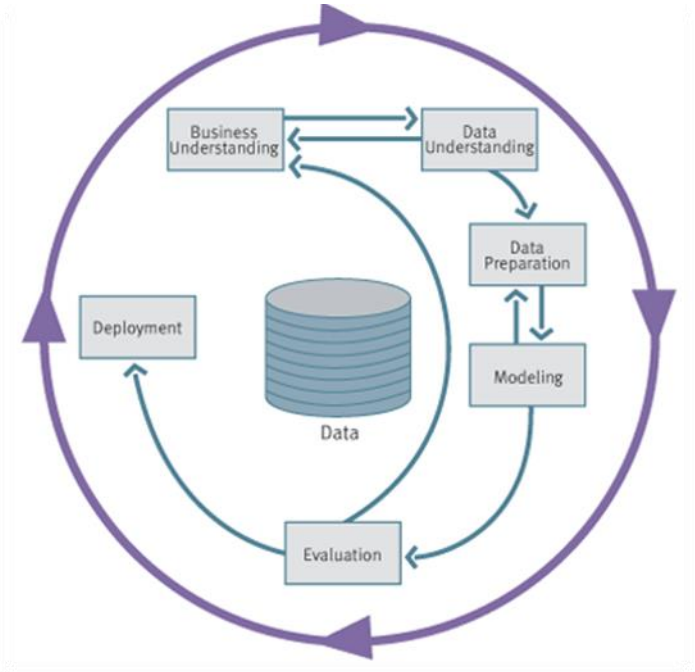

Figure 1: CRISP-DM v. 1.0 process model (reproduced with authorization from www.crisp-dm.org).

In the Business understanding phase the analyst should understand the business purposes, define and evaluate the problem, and establishes purposes for the data mining project. The main output is a project plan.

Strong competitive pressure, the often saturated market, and maturity of products cause a higher demand for information and quantitative modelling of the behaviour of customers and competitors. These mean a big variety of problems to tackle in business. Some examples are market basket analysis, portfolio's return of investment, determine the credit worthiness of a costumer or identify fraudulent use of credit cards. Controlling and scheduling of production is another application field. Understanding of dynamic behaviour is another common business concern in many areas, such as identification of faults in telecommunications networks or analysing costumer interaction with a Web site (Klösgen and Żytkow, 2002).

The problems can often be divided in two big sets. The classification and forecasting problems where an especial variable, often called decision, target or dependent variable has to be predicted by using the information contained in a set of explanatory or predictor variables. The intention in this type of problems is to set a model using data over both types of variables and use the

model to make forecasts or classifications when the value of the dependent variable is unknown. These are called supervised problems.

In the case of unsupervised problems, there are no specially variables and all the data is treated equally. In that case the intention is to divide the data in previously unknown classes (clustering modelling) or construct association rules like in market basket analysis.

The Data understanding phase includes activities as data collection, exploration and data quality verification. The main purposes are to get familiar with the data, to detect especially interesting subsets and to discover first insights. Data screening is very relevant. If we seek predictors for one variable, we need this variable in the data as well as a number of potential predictor variables whose values are known before the value of the variable can be predicted. This means understanding very well the variables which are available and have a good insight over how they are associated. We must also be sure that data cover a broad range of values. In particular data should be available for contrasting various generalizations.

Data preparation includes ETL (Extract, Transformation and Loading) and pre-processing activities, as selecting data, data cleansing, data fusion and integration, and all other activities needed to prepare data for modelling. It is an iterative process including feedback loops, as the modelling process may generate insights in the domain that require additional data preparations. New variables may be derived if they are more suited to the analytical tasks, like scale reduction as discretization or dichotomization. This process is usually called feature extraction in data mining literature. If the size of the data is too large for efficient analysis, data reduction techniques can be used, either by feature selection, which consists on selecting subsets of relevant variables, and sampling methods for selecting subsets of records or data table lines.

The necessary data commonly is spread over various relational tables, databases or even information systems. Data fusion and integration is so a very relevant issue as we must find out how data from different sources can be pulled together, because the preponderance of quantitative modelling tools apply to single tables.

ETL, data cleaning, selection and reduction are some of the key steps in creating a data warehouse, which is a repository of data that summarizes the history of business operations and is created for the propose of analysis.

Several modelling techniques are applied to the data set and modelling parameters are optimized. The models are built by searching in a combinatory space where many possibilities are explored (Klösgen and Żytkow, 2002). Huge heuristics for model construction have been developed in the latter years. When the explanatory variables and also the dependent variable are categorical the methods for classification trees and induction of rules are adequate. In the case of numerical variables, some examples are neural networks and linear regression, but rules and trees, namely regression and model trees, can also be used. Other methods like support vector machines and nearest neighbour method can also produce good results. Models can be expressed as equations, logical rules, visual trees or even sets of nearest neighbours. Choosing the set of right methods is not a simple task, and can be modelled as a multicriteria decision problem, so that the choice must include external judgment on the importance of the different criteria for the particular application. As is recognized, each heuristic method searches for the best model in a particular hypothesis space. So, is necessary to evaluate and compare models by applying the model to new data or data set apart from the modelling data.

Before accepting the model or models resulting from the last phase, a thorough evaluation is imperative. mention the case of survey analysis. Sometimes, some unexpected and not substantively explained results are apparent. Most of them led to the identification of some tricky behaviour of interviewers trying to speed up their interview efforts and other types of deficiencies or errors in data. This evaluation uses not only verification against new data, but mainly domain knowledge and knowledge generated by the process, especially in business and data understanding. A key task is to verify if the model solves the initial problem in a suitable way.

Another serious problem that requires knowledge refinement stems from the abundance of results normally easily obtained using data mining software. Such abundance is one of the main challenges in many business applications.

Performance measures comparisons are the major technique to deal with the refinement of results. Is also relevant to ensure robustness of results according to several small data variations. Results from many instances of search process can be generalized, for example, by identifying the main influence and their conditions. Visualization is also an important tool in this phase. Visualization enhances understanding of data, models and knowledge, allowing the analyst to redirect and focus the search process.

The deployment are ways to spread the models and\or the knowledge generated by the project, and make sure that the appropriate decision maker or user who understands and uses it. This process model recognizes that users can only be in charge of knowledge applications if they understand the knowledge discovered. But users are often not familiar with tools and different data and knowledge formats that those tools use. The understandable presentation of results is critical. For that means everything from writing a report to implementing an application for model renewal whenever needed is possible. The process starts from business problem and data understanding, and ends with actionable conclusions from the discovered knowledge. After and action is deployed, practical evaluation of results is possible.

Authors like note that is difficult to find documented and detailed information about deployment of results in the corporate world. The same authors claim that KLM estimated that the implementation of projects including a substantial element of discovery from historical data, had a pay-back time less than one year and lead to a two percent reduction in total human resource costs.

This process model can be compared with the OR decision methodology () or general methodologies for solving problems and it’s easy to show the high similitude between them being the latter manly an evolution from the first with some adaption for data rich environments.

Several authors, like Poon and Wagner (2001), recognize as a major critical factor the executive and operational sponsorship in the adoption of systems based in business models.

The following sections will present the application of this methodology. At the end, the results and conclusions will be presented.

3. Process Model Cases of Application

The authors have been applying CRISP-DM process model to quantitative modelling management problems, mainly problems tackled by data mining procedures. In the following table two of the cases are summarised.

Table 1: CRISP-DM process model and cases of application.

Business understanding	understanding commercial market and the itemset bought by each customer	the proverbs are a manifestation of traditional culture. Knowing a proverb is an indicator of the cultural reference region a person lived
Data understanding	big database with records of distribution of frozen food items throughout Portugal by the Nestle enterprise and bought together by customers	in a series of interviews, it was collected a heterogeneous set of several million relations of positive and negative knowledge that a group of one thousand of people had regarding a set of about twenty-two thousand Portuguese proverbs
Data preparation	it was necessary to transform the data table extracted from the database to a table (transaction, item) with a set of all transactions where each transaction contains a subset of items.	the data had many faults and inconsistencies that were corrected using the knowledge gathered during the interviews. A data table was prepared using a query from the clean database.
Modelling	firstly, the input dataset is transformed into a graph-based structure and then the maximum-weighted clique problem is solved using a meta-heuristic approach in order to find the most frequent itemsets.	the problem was tackled by reduction of the dataset and rule covering algorithms. A new two-phase algorithm was presented. First, the problem is transformed by generating a matrix with the disjoint constraint. Second, the minimal subset of attributes is chosen using a well-known Set Covering Problem.
Evaluation	to validate the Similis algorithm a real dataset of frozen-food and other datasets from the Frequent Itemset Mining (FIM) Implementations Repository were used. Several measures were calculated to compare results.	using Leave-One-Out cross-validation, several measures like area under ROC curve and k statistics were calculated. The dataset reduction results in a dozen of attributes and a hundred of rules.
Deployment	the algorithm was made available to modellers	the new algorithm was made available to modellers
Handicaps	customer disinterest in the implementation	the main difficulties were in data preparation, because of faults in database.
Reference	Cavique (2007)	Cavique et al. (2011)

In the same table, for each application, the CRISP-DM handicaps, like customer disinterest in the implementation, no business understanding, or very long data preparation due to inconsistencies in data.

The following cases are described in more detail to serve as case study of CRISP-DM application.

3.1. Business UNDERSTANDING:

In these cases, top EDA (Electricidade dos Açores) management was the client, being the users the IT specialists that designed and developed the system with the authors. Those last ones were profoundly involved in the system development and were also responsible for all the communication with the client's project.

Following CRISP-DM process model, we first collected data over the company and main business. This was an easy phase, because the EDA collaborators collected all the data and answered all the questions. The following phases were trickier and so much more interesting for case study proposes.

Both decisions engaged in mentioned the projects needed learning from data, as it is defined in knowledge management literature (see for instance). For learning we used two main approaches: a business Intelligence project based on OLAP technologies (MS. SQL Server) and a data mining project which used statistics and machine learning models. The Cross Industry Standard Process for Data Mining Process Model (CRISP-DM) was applied as a way to define methodological phases and to integrate business intelligence in a data mining framework.

This work reports on the methodological knowledge generated by several projects which intended to support decisions in the Electric Company of Azores Islands (Portugal), EDA - Electricidade dos Açores. These are real life applications of Business Intelligence and Data Mining technologies. All the projects meant to extract organizational knowledge from data records to support different kinds of decisions.

One strategic decision would be whether EDA communications, among the islands, should be moving to Voice over IP (VoIP) from present telephone lines and, if the feedback is positive, how to do it. This decision is not properly structured and must be based on technical and non-technical criteria's. For the technical criteria, a Decision Support System (DSS) was developed based on

data of an external telecommunications company, and MS SQL technologies. This project generated several others related to fraud detection in using telecommunications within EDA installations in all the nine islands. The results were published, in detail in a previous publication (Mendes, *et al*. 2009).

Another project intended to analyse the relation between the climatic factors and the consumption of electric power. This is, certainly, a different type of decision, more frequent and operational. But is not completely structured, since the climate influences direct and indirectly the consumption of electric power. With this project, we plan to develop models and knowledge to support consumption forecasts. These are critical decisions for a power producing energy, since this type of energy cannot be efficiently stocked, and so production must always be in phase with consumption.

Most papers about this subject focus on improving the prediction of electricity demand and on how to obtain forecasts as soon as possible, for better resemblance between production and consumption (see for instance: , and). All these papers mention the relevance of weather in electricity demand. In Smith (1989) the climate, is considered the major error factor in electricity demand forecasts. This problem can be even more complex when it comes to islands, subordinated to many weather variations, and where the investment in alternative renewable energies is higher. Decisions about how much energy to produce by flexible ways, like burning fuel, are especially relevant. In this context, any new knowledge is welcome.

The EDA Company (www.eda.pt) is responsible for the production, transportation and trade of electric power in all of the nine Azores islands. Other companies can also produce electric power, but they must sell it to EDA, because this the only one certified company to transport and resell electricity to consumers. Data of 2010 fiscal year indicate a turnover of 199 million euros and a total of 121.164 customers spread over the nine islands of the archipelago of Azores. EDA company has 687 employees, and it’s the head of a group, which include 6 other companies, approaching 870 permanent employees. EDA has a particular complex communication system, because of the dispersion of clients spread over a wide discontinuous area of 66 thousand square kilometres (see Figure 2).

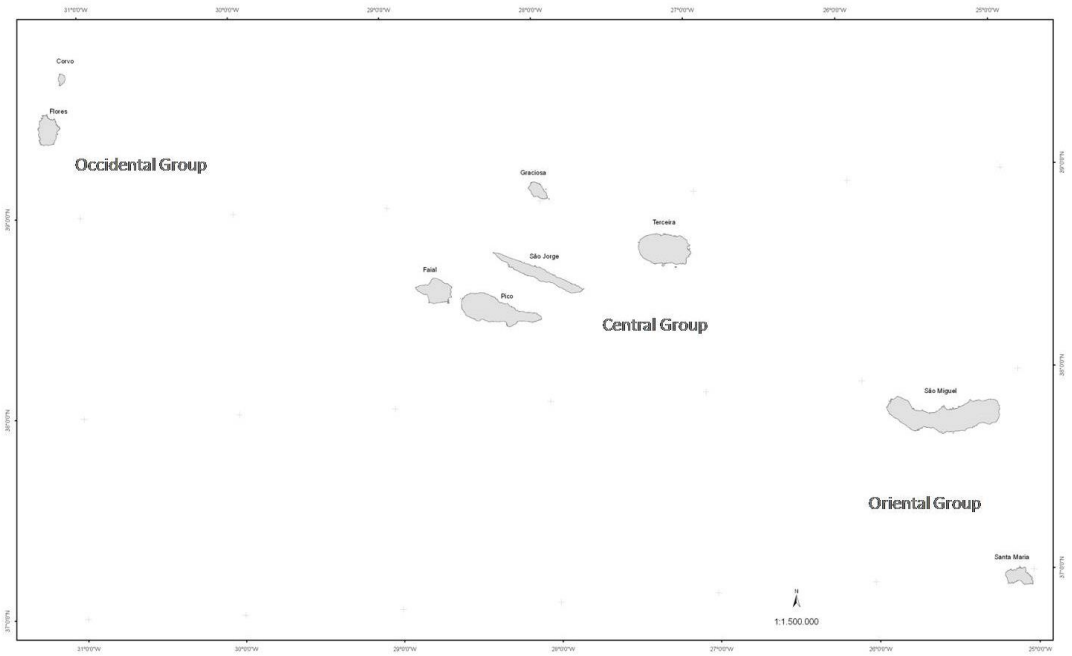

Figure 2: Azores archipelago showing the nine islands dispersion.

The EDA Company produces a mix of energy that is still largely dominated by thermoelectric power, although it also includes geothermic production (only in the biggest island), hydrologic and private production, mainly biogas. In recent years, the investment in renewable energy, such as geothermic, has been growing rapidly, for 43,5% of all energy consumed in São Miguel during 2010.

Most of the decisions were semi-structured. In Since Keen and Morton (1978) seminal work, has been shown that data analysis systems are very useful in the screening of these types of decisions. As the major part of the work meant to analyse data, in regular basis or relating a decision taken in one specific moment in time, we suggested an approach based on OLAP and data mining. This was discussed with EDA specialists and decision makers, and was accepted for both projects. In this way, MS. SQL Server software was selected as the adequate, and, more decisively, accessible for the EDA specialists to manipulate, as well as to improve the system in order to come within reach of the user's needs, in an iterative and interactive process initiated by projects like these. In fact, any data mining and BI software could be used in this context. We used an opportunistic criterion to select SQL Server software.

3.2. DATA UNDERSTANDING AND EXPLORATION: THE OLAP PHASE

In both projects, the contact with management to establish purposes and patronize the projects was easy. In both of them, the acknowledged purposes were to identify operational rules related to reducing costs and, in the case of the communications project, to also support the decision about moving from external telephone service operators to Voice over IP (VoIP), handled internally.

Also in both projects, a data warehouse was built from scratch, because it there weren't any data or information regarding both problems. The required data was, in both cases, external, for the communications decision it includes working patterns, number of calls, length, frequency and use in peak hours. In the forecast project: electricity consumption \ production values by the hour and mainly climate data, such as temperature, humidity, rain quantity, visibility, wind direction and intensity, and a record of climatic events like exceptional winds or rain. Both projects benefitted greatly from inside experience on communication technology, data and models.

In spite of the fact that nowadays we live submerged in big data waves, it is still very tricky to capture, evaluate quality and explore data. These are the main purposes of the CRISP-DM phase of data understanding. This phase was done before the data warehouse construction, using small parts of data and easily accessible and simple software, like statistical packages and R-software for table and graphic data exploration and the discussion of the results with EDA professionals. In this phase, the initial purpose grew deeper and the knowledge generated from discussion was shared.

Some poor communication between information systems was also an obstacle in the case of communications data, as we needed pre-existing data, as the locations, phone numbers, and the identification of the user accountable for the phone terminal.

The outputs of this phase concern knowledge about data. Facts, like the three almost perfect direct linear relationships between costs and call duration, were almost completely explained by the 3 different time periods cost values.

Obvious seasonalities on the number of calls, much higher during weekdays with strong reductions in weekends, were as well simple to explain, as only maintenance staff work on holidays and weekends. Similar seasonalities were easily identified in electric power consumption with much lower values during nights and week-ends. By graphing call duration in a histogram, a distribution similar to an exponential distribution was recognized, with almost all calls very short, tough some can be very long. Some long calls were specially noticed by EDA professionals. By graphing two years of daily costs in communications some other seasonalities were recognized as lower activity during the summer due to vacations and to a period during 2005, when EDA main building was transferred to a new location.

To make data exploration and dicing easy to any user, an OLAP application was implemented in both projects. This was a time consuming phase and comprehend pre-processing of the data, data exploration, data reduction and visualization. The Software tools used were based on the Microsoft SQL Server technologies, already known and used by the EDA systems professionals. The main components included the Data Base Engine, Analysis Services and Integration Services from Business Intelligence (BI) Studio. This became a major measure in both cases. But in the case of the strategic decision about communications network, the OLAP project was considered more relevant and the application was extensively used by professionals. In the case of electric power forecast, the OLAP software was used mainly for data exploration.

In spite of all this tools, some actually helpful, the construction and management of data cubes was a long and hard phase. Process flows from Integration Services were identified as one of the most important tools for data preparation, such as the generation of new fields, tables' relational integration and populate fields. One example, for the transformation and preparation of the foreign keys in relational tables for new months, and establishing several relations with existing ones, is shown in the Figure 3. The nodes in the process flow correspond to SQL coding and some other parameters. This is not uncommon in this kind of projects. Many other authors reported similar problems in supporting decisions in data rich ().

This phase is also very important for data quality evaluation. In both projects many of the problems described in the book by Chen (2001) were actually identified. That was the case especially for electric power data and data collected from internet. Many missing values or non-conform values were identified and coded. Another important problem identified in both projects was keys mismatch in related tables. This problem, often consequence of data fusion activities, is tackled in) and). For instance in matching electric power consumption with climatic data, a process flow was programmed to extract the records that matched the year, month, day and hour.

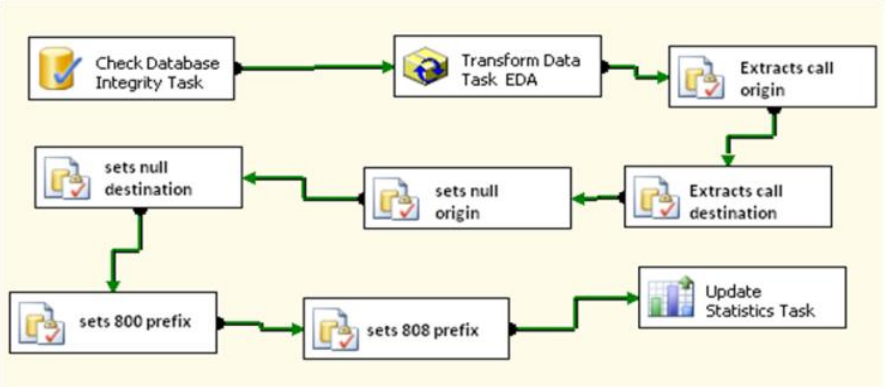

Figure 3: Process flow example for the communication network decision.

In 1993, E.F. Codd (sited in Larson, 2006), one of the fathers of relational database and On-Line Transaction Processing (OLTP) theory, proposed a different type of system that would be turned to the needs of data analysists. He called it an On-Line Analytical Processing System (OLAP). The criteria Codd originally developed for an OLAP system were not widely accepted, but the acronym and the name are widely used today for systems designed to quickly access, aggregate and explore large data tables.

For the data mart design in the communications project, 3 measures were defined: number of calls, simply the row counting of the data table, call duration and call cost. These are numeric quantities easily obtained from the external phone company, noticeably linked with the project purposes. The first one was only included in a later stage of the project, as it was considered relevant by the EDA professionals.

Dimensions are discrete fields used to define the aggregation degree of the measures. A very useful concept of MS. SQL 2005 is a hierarchy which is a way to organize dimension in various levels. For instance, in Figure 4 the time period is used in the following way: as the year dimension is above the trimester and this later one, above the month. Many other dimensions are used in the cube shown in Figure 4, as the company, telephone equipment, island of origin and destination, type of call, equipment user, time of the day, *etc.*. Other than new data, many are extracted from several OLTP data bases already available in the EDA group.

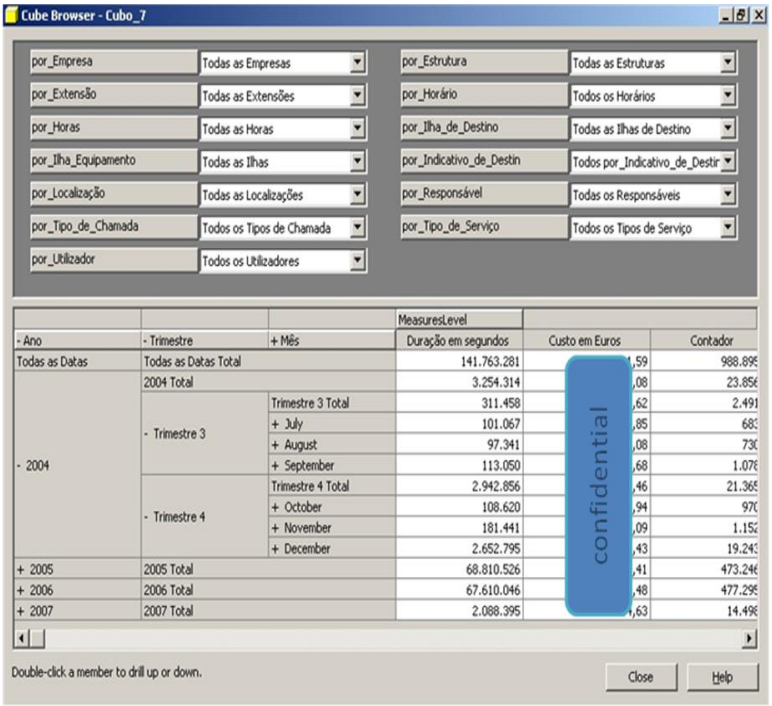

Figure 4: The final data cube for the communication network decision.

As you can see in Figure 4, the categorical fields can be used as aggregation dimensions (as the 3 dimensions of time period in the example above) or as filters as the other dimensions over the table. To interchange the dimensions used to filter and to aggregate data is only necessary click and drag between both areas.

Several data cubes were constructed in an evolutionary process, as the discussion about data mart design went on and new data were integrated in the data warehouse. The star scheme was selected as it is known to have less

performance problems in a ROLAP architecture (see Chen, 2001 and **Erro! A origem da referência não foi encontrada.**).

3.3. MODELING: THE DATA MINING PHASE

In fact, the OLAP project was much more than the data preparation and the exploration phase of CRISP-DM methodology. With the final data cube we answered many of the initial questions and actually generated knowledge and business intelligence, especially in the strategic decision.

In spite of this, it is also clear that the main data preparation necessary for an OLAP project is also required for the use of data mining algorithms. Many software houses recognize it by implementing both business intelligence technologies in the same framework. In the Microsoft case, the Business Intelligence Development Studio includes several tools for both OLAP analysis services and data mining. Both can use SQL Server Integration Services to extract the data, cleanse it, and put it in an easily accessible form.

For modelling, we employed the same data as the one used in the cube of both projects, as data source, in order to generate data tables for learning and testing. This data was used in a twofold validation scheme: circa of 2/3 of older data (130 thousand lines, years 2005 and 2006, in communications project and 16 thousand lines, years 2006 and 2007, in forecast project) for model estimation (or learning), and most recent data for validation (or testing).

The Business Intelligence Development Studio in MS SQL Server 2005 includes 7 mining algorithms, which perform the main tasks usually associated with data mining. These are: classification using a categorical target field, regression for a continuous target field, segmentation for defining clusters without a target field, association for rule induction, and sequence analysis for rules including a sequence of steeps.

In spite of the fact that data mining packages always have many algorithms for data model, is important to understand that they have different purposes and use different types of data. For instance, the forecast project had data which consists mostly on time series.

For the forecast purpose we need algorithms capable of identifying patterns in older data that can be extrapolated for future as well, as relations between the power consumption and other descriptive variables. For that propose, and also considering that the target variable has a continuous scale, the chosen algorithms were *Microsoft Decision Trees*, *Microsoft Linear Regression* and *Microsoft Neural Network*. A brief description of these algorithms can be found in Larson (2006).

In the communications project the object was less restrict in terms of algorithms that can be used. As the intention was to produce knowledge about the way telephone lines were used, almost any algorithm could be tried in this exploratory approach. Therefore, several algorithms were tested and four were regarded useful, considering their results and the project point: *Microsoft Naïve Bayes*, *Microsoft Decision Trees*, *Microsoft Clustering*, and *Microsoft Association*.

Other algorithms were regarded as not suitable for the defined data mining goals, inappropriate for the available data types, or we just couldn't find any interesting result. One example was the *Microsoft Time Series algorithm*. Since the available data is chronological series, this could be regarded as one of the major algorithms in the use of forecasting continuous variables. In spite of this, the unique autoregressive tree model used in this software does not allow the estimation of parameters we needed as seasonal factors (see Meek et al., 2002, for a complete description of the algorithm). For this reason we estimated regression models with dummy variables using statistical software for the calculation of the month seasonal factors.

One interesting feature of the software utilized was the dependency network which is a network where nodes represent attributes (or variables) and links represent causal relations or correlation dependencies between attributes. This is an interesting way to visualize algorithm results (). Other uncommon feature we would like to distinguish is the possibility of generating data tables on the fly by selecting a key variable for aggregation proposes. This is very handy when dealing with lots of data, since the models are usually generated from aggregated data. This means that it is not necessary to maintain several tables with data aggregated for different keys in order to generate different

models, but, as would be expected, the consequence is some delay in presenting algorithm results.

From these two projects and others in different contexts, we found that the data mining algorithms provided in the Business Intelligence Development Studio, the *Microsoft Naïve Bayes* was found one of the most useful. This was the case because usually many of the variables used in these projects are categorical. For instance, in applying this algorithm to the call data and only having in consideration the more expensive ones, we found that 80% of this were originated from the main island, where the headquarters are located, 80% of this calls had duration between 3 and 10 minutes; 51% were made directly and 42% by human operator (the remaining 7% were for special numbers). This last figure was considered too high by the professionals and other models were built to understand what was happening.

This algorithm can, also, produce a ranking of the best predictors of a dependent variable. For example, from climatic data we found that the best predictors of electric power consumption were, in order, humidity, dew point and temperature. The last variables in the list were wind velocity and climatic conditions. Note that the Naïve Bayes algorithm used in SQL Server 2005 does not consider combinations of variables (Larson, 2006), which is unexpected in data mining software (see any data mining text book as Witten and Frank (2005).

Microsoft Decision Trees is an algorithm that produces a tree structure defining logical rules for explaining a target variable, using several explaining categorical or categorized variables. In the MS implementation, it can be regarded as a generalization of Naïve Bayes algorithm or a form of Bayesian network (**Erro! A origem da referência não foi encontrada.**). Analysing many trees built for the call data and suing the cost variable as a target and other many variables as keys, used to define the aggregation levels, it became simple to recognize the obvious relation between call duration and cost. Excluding duration from the explanatory variables it was possible to conclude that when the destination of the call is the island of São Miguel (the biggest one with half of the total archipelago population) the majority of the calls were not direct calls, especially the most expensive ones.

The *Microsoft Clustering* algorithm builds clusters of entities based on proximity measures calculated from the training data set. Once the clusters are created, the algorithm describes each cluster, summarizing the values of each variable in each cluster. This algorithm has the originality of displaying results not only in tabular form, but also in a network scheme, where links and colour codes relate clusters. From the many clusters defined in this way, cluster 6 represented an especial interest as it was characterized by long calls; it had a strange distribution far from peak hours, and also abnormal destination numbers. This cluster represents a significant amount of suspicious calls.

Using this algorithm on the climatic data and day as key attribute, we found 10 clusters, some of that could be used to confirm the results obtained from other algorithms like the Naive Bayes. For instance, in one of the clusters high power consumption (bigger than 47 MW in 82% of the cases) corresponds to high temperatures (bigger than 19.7ºC in 98.3% of the cases).

The *Microsoft Association* is an *apriori* algorithm type association for the induction of association rules. It produces an ordered list of item sets, rules with precision values and it results on a dependency network. This algorithm was considered very interesting and was one of the most used in the communications study. For example, it was possible to conclude that there was a high support for calls by human operator with origin and destination in the same island, which seems suspicious as these calls may easily be made by a direct call using the company network.

All that models were validated using the main tools in MS SQL server, the lift chart and classification (or confusion) matrix. These are charts that compare the precision of the classification (or forecast for continuous variables) for the different models used. These charts can take a long time to be built and were useful only to compare models with each other in the worst case scenario. They confirmed that rules induced by decision trees and Naïve Bayes where the best ones for call cost forecast.

In both projects it was possible to find good validated models. In the communications study the quality measures calculated over the test data were coefficient of determination of 87%, root mean square of error of 5.9, a mean absolute deviation of 5.0 and the mean absolute percentage error of 19%. For

the climatic data identical procedure resulted in coefficient of determination of 94%, root mean square error of 3,52, a mean absolute deviation of 0,21 and the mean absolute percentage error of 2.92 %. We believe that these are fair good results which were corroborated with domain knowledge from EDA professionals.

3.4. RESULTS AND DECISIONS

The model results were discussed with EDA experts in order to consolidate the knowledge captured. For instance, for more exploratory study, where this phase is especially relevant, the peak hours are between 9 to 11 and between 14 to 16 hour each weekday and there is very low use at night and during lunch time, and also at weekends and holidays. There were no seasonalities between the weekdays, as they have almost the same high use. On the other hand, the month seasonal factors indicate less use during summer and around the New Year's Eve. The most common call destination goes to the tree bigger towns in the region, as it was expected, and the calls' length is usually lower than 3 minutes. The special numbers, like the call centre number, are of low usage.

This kind of exploratory information is enormously important for the particular decision to support. Especially the strong seasonalities identified mean that the equipment capacities must be planned for peak periods. From the trends estimated by the regression models, there is no evidence of increasing total duration of calls, or even in peak periods, as trend lines were always non-significant for the two years of data.

As it was recognized by other authors (see for instance Cortes et al., 2001) the key criteria for decisions relating a telecommunication investment is the cost of the different solutions. In this way, for a final decision, a cost analysis was also prepared using data collected from the previous analysis, comparing the costs for the existent communication system with two change scenarios. The three options defined in this phase are based in different technical solutions derived by the EDA experts.

Option A applies a minimum of investment using the existing lines and only buying the necessary equipment's. In spite of a reduced investment, reduces the annual operation costs in 15%, but there is no expected cost reduction for the new VoIP links between internal locations, due to low volume calls. This option maintains the two technologies presently used by the communication system until the end of equipment life: a PBx central for voice communication and VoIP, being this last one much more utilized than the one before as the connections between islands main stations would utilize this technology.

Option B consists on replacing progressively all equipment's resulting in a new telecommunications infrastructure based on VoIP Routers and Call Managers for voice and data communications. This option requires a big investment in new equipment's, 7 times option A, but, when finished, it will decrease the annual operation costs on 165% from the present values.

The current situation has no capital but high operational costs. As it was realized that was a complex decision with a multicriteria structure, all the factors considered for that decision are summarized in Table 2. Our conclusion, from cost analysis and business intelligence, is that both new solutions look attractive as the benefits compensate the costs in the long run. The decision aid group recommended the adoption of option B, as in a strategic view it will benefit the company, relating not only money, but also a "technological image" of the company and the simplification of operation activities. In spite of the fact that no numerical evaluation of criteria neither weight was calculated, as the decision seemed clear, the recommendation was adopted by decision makers executives and a project is now being implemented.

Table 2: Decision features of the two change alternatives.

actions	costs	benefits	critical factors
Option A - VoIP connections maintaining present WAN equipment's with minor changes			
Upgrade existing infrastructure; upgrade circuits bandwidth; provision of additional VoIP cards to existing equipment's; configure all network.	High ongoing and support costs due to obsolete and mixed technologies; High cost with service operators; High operation costs if traffic increases.	Reduced investment on infrastructure; Low reduction on telephone service costs.	Age of telephonic equipment and end of life (presently with 8 years); Cost of introduction of additional VoIP links exceed the actual costs.
Option B - VoIP connections replacing WAN equipment's			
Upgrade circuits bandwidth Provision of Routers, CallManager´s, Switching and terminal equipment; Change all WAN and LAN equipments supporting voice Configure all network.	High cost of initial investment; Change management.	Reduced ongoing cost and high reduction on support costs; Fixed telephone service; Renewed WAN infrastructures; New services introduced by IP telephony.	Long term benefits due to the reduction on ongoing costs and new services; Capability of existing service provider to implement and support the new infrastructures; Migration of existing services (Call Centre).

For the climatic data project there were no decisions to been taken, but there was a need for better models and for the understanding of climatic effects on electricity consumption. For the model constructing a simple to complex approach was adopted. Starting from simple regression models we ended by choosing a regression model tree as the final model. This model is a combination of regressions with a classification tree which divides the initial data in smaller sets. These smaller data sets are then used to obtain regression models. In Figure 5 we present the results for a terminal node of the model tree.

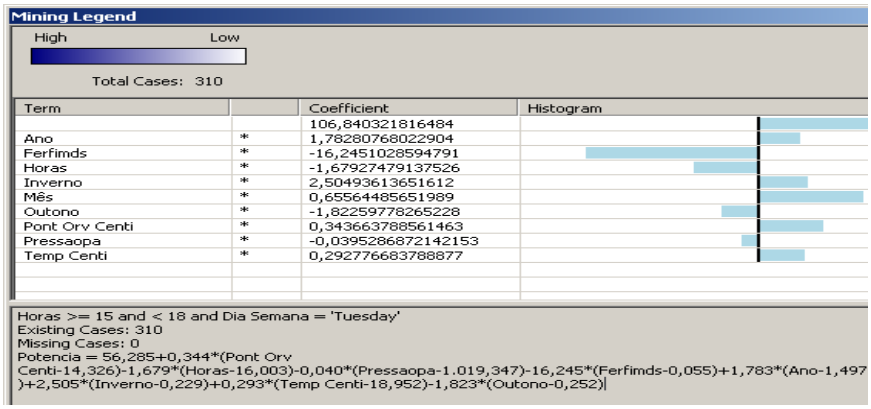

Figure 5: One terminal node of the model tree obtained with climatic data.

This model can be easily described. In every Thursday between 15 to 18 hours for one degree Celsius over the average temperature the power consumption is 0.293 MW higher, if we do not consider changes in all other variables. For the same expansion value of the dew point, the power is 0.344 MW. We can find, also, an improvement of the average annual power consumption of 1.783 MW and 0.656 MW monthly. In winter an improvement of 2.505*(1-0.229) or 1.93 MW is also expected and also a reduction in power consumption of 1.823*(1-0.252) or 1.36 MW is expected in autumn.

In other branches of the tree we can see that temperature, dew point and humidity are the more important climatic factors and we can find an improvement in power consumption every time these factors present higher values. This was rationalized by specialists for the need to use refrigerator and humidity control systems when temperatures and humidity are higher. On the other side, there is a known physical phenomenon that explains higher losses in electric energy transportation in these conditions. There is also a possibility of other indirect factors may be influencing power consumption, like sun exposure and higher population in summer due to tourism.

This model is considered a good representation, not only because of the good quality statistics and graphs, like the one in Figure 6, but also because domain knowledge supports the fact that the very strong seasonalities identified make the division of data more adequate for regression models. This pattern in electrical power consumption is highly recognized in published work, being common the recommendation of modelling only particular periods of time like peak hours (see and).

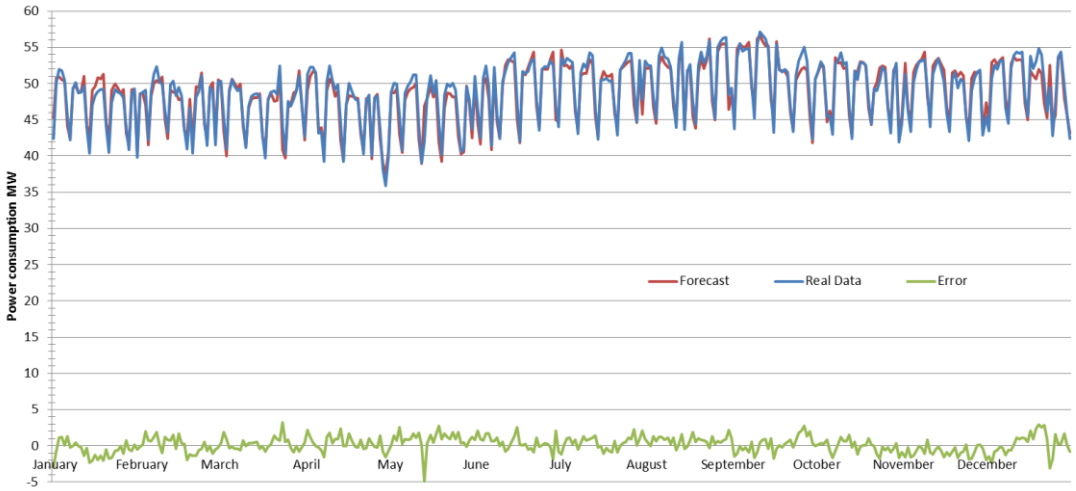

Figure 6: Mean daily power consumption for year 2007, both recorded and forecasted.

4. CONCLUSIONS

In this chapter we explain the need for a process model in quantitative modelling of management problems supported by database management systems, data warehouses and OLAP technologies. The CRISP-DM Process Model is presented and two applications are described.

In each application the decisions to support were completely different, in structure and frequency as well as in data requirements, the technologies used were found very useful and resourceful. The component of data mining seems to have as main purposes the easiness to use and automation. SQL Server 2005 Data Mining Add-ins was found especially interesting for easily exploring relatively small data sets. Some algorithms are black boxes and not very clear for the user. In spite of that, it can be actually relevant in management context for quantitative model, especially with big data tables and using database management systems.

In addition to the fact that was possible to support the right decisions, and producing reports or models for future use, this technology also allowed to collect actually good knowledge. Examples of that is all the knowledge about seasonalities in the climatic data, and more relevantly the relative importance of climatic factors.

But, the most fundamental example of information collected is all the relevant faults and inefficient procedures identified in the communications project. A

concrete example is the high number of long calls not related to business activities, for personal and shopping purposes. It was also possible to identify a miss configuration on automatic call distribution, resulting on additional external calls, which were more expensive, and terminal equipments not used but that had subscription costs. From these fault detection activities several terminal equipments have been eliminated and some ghost traffic reduced.
But the major and unexpected result was the high number of indirect calls, using human service operator, as a way around to the existing control system. Doing an indirect call, the link between the call origin and destination is much more difficult to establish. This fact led to new rules of operation, restricting calls by human operator. In the deployment phase applications (as the OLAP cube) were developed for use by several technicians and decision makers.
These successful projects are good examples of quantitative modelling in data mining supported by a comprehensive process model.

5. REFERENCES

Adriaans, P. and Zantinge, D. (1996). Data Mining. Addison-Wesley: Massachusetts, USA.

Arnott, D.; Pervan G. (2005). “A critical analysis of decision support systems research”. *Journal of Information Technology*. 20, pp. 67-87.

Cavique L, A. B. Mendes, M. Funk (2011) "Logical Analysis of Inconsistent Data (LAID) for a Paremiologic Study", proceedings in press, EPIA 2011.

Cavique, L. (2007), “A Scalable Algorithm for the Market Basket Analysis”, *Journal of Retailing and Consumer Services*, Special Issue on Data Mining in Retailing and Consumer Services, 14 (6), pp. 400-407.

Chapman, P.; Clinton, J.; Kerber, R.; Khabaza, T.; Reinartz, T.; Shearer, C.; Wirth, R. (2000). “CRISP-DM 1.0 - Step-by-step data mining guide”. SPSS Inc..

Chen Z. (2001). Intelligent Data Warehousing: From data preparation to data mining. CRC Press. Boca Raton.

Clifton, C.; Thuraisingham, B. (2001). “Emerging standards for data mining”. *Computer Standards & Interfaces* 23: 187-193.

Cortes, P.; Onieva, L.;Larrañeta, J.; Garcia, J.M. (2001). Decision support system for planning telecommunication networks: A case study applied to the Andalusian region. *J. Opl. Res. Soc.,* **52**, pp. 283-290

Engle, R.F.; Mustafa, C.; Rice, J. (1992). “Modelling peak electricity demand”. *Journal of Forecasting*, **11**: pp. 241-251.

Fayyad, M.; Piatetsky-Shapiro, G.; Smyth, P. (1996). “From data mining to knowledge discovery: An overview”. In Fayyad, U. M.; Piatetsky-Shapiro, G.; Smyth, P.; Uthurusamy, R. (Eds.) *Advances in Knowledge Discovery and Data Mining*. MIT Press: Menlo Park, USA, pp 1-34.

Hand, D.J.; Mannila, H.; Smyth, P. (2001).“Principles of Data Mining”. *Adaptive Computation and Machine Learning*, MIT Press: Cambridge, USA.

Keen, P.G.W. (1980). “Adaptive design for decision support system”. *Data Base* 12, pp. 15-25.

Keen, P.G.W.; Morton, M.S.S. (1978). “Decision Support Systems: An organizational perspective”. Addison-Wesley Series on Decision Support, Addison-Wesley: Reading, USA.

Klösgen, Willi e Żytkow, Jan M. (2002). 2 The knowledge discovery process. In Klösgen, Willi e Żytkow, Jan M. (Eds.) *Handbook of Data Mining and Knowledge Discovery*, 1ª ed. Oxford University Press: New York, USA, pp 10-21.

Larson B. (2006). Delivering Business Intelligence with MS SQL Server 2005. McGraw-Hill. Emeryville.

Lavrač, N.; Motoda, H.; Fawcett, T.; Holte, R.; Langley, P.; Adriaans, P. (2004). “Introduction: Lessons learned from data mining applications and collaborative problem solving”. *Machine Learning* 57: 13-34.

Liu, L.-M.; Harris, J.L. (1993). “Dynamic structural analysis and forecasting of residential electricity consumption”. *International Journal of Forecasting*, **9**: pp. 437-455.

Meek, C.; Chickering, D.M.; Heckerman, D. (2002). Autoregressive tree models for time-series analysis. In Proceedings of the 2ª ed. of the Int. SIAM Conference on Data Mining. SIAM: Arlington, pp. 229-244.

Mendes, A.B.; Ferreira, A.; Alfaro, P.J. (2009) "Supporting a Telecommunications Decision with OLAP and Data Mining: A case study in Azores". In Cruz-Cunha, Maria Manuela; Varajão, João Eduardo Quintela e Amaral, Luís Alfredo Martins Proceedings of the CENTERIS 2009. CENTERIS: Ofir, Portugal, pp 537-549.

Poon, P.; Wagner, C. (2001). "Critical Success Factors Revisited: Success and failure cases of information systems for senior executives". *Decis. Support. Syst* **23**, pp. 149-159.

Saporta, G. (2002). Data fusion and data grafting. *Computational Statistics & Data Analysis*, **38**: pp. 465-473.

Shearer, Colin (2000). The CRISP-DM model: The new blueprint for data mining. *Journal of Data Warehousing*, 5 (5), pp. 13-22.

Smith, D.G.C. (1989). "Combination of forecasts in electricity demand prediction". *Journal of Forecasting*. **8**: pp. 349-356.

Troutt, M.D.; Mumford, L.G.; Schultz, D.E. (1991). "Using spreadsheet simulation to generate a distribution of forecasts for electric power demand*". Journal of the Operational Research Society*, **42**: pp. 931-939.

de Ville B. (2001). Microsoft Data Mining: Integrated business intelligence for e-commerce and knowledge management. Digital Press: Boston.

White, D.J. (1975). "Decision Methodology". John Wiley & Sons: London, UK.

Wijnhoven, F. (2003). "Operational knowledge management: Identification of knowledge objects, operation methods, and goals and means for the support function". *J. Opl. Res. Soc.,* **54**, pp. 194-203.

Witten, Ian H. e Frank, Eibe (2005). Data Mining: Practical machine learning tools and techniques. *The Morgan Kaufmann Series in Data Management Systems*, Morgan Kaufmann Pubs: San Francisco, USA.