BI and Data Warehouse Solutions for Energy Production Industry: Application of the CRISP-DM methodology

Armando B. MENDES[a,1]
[a] *Azores University and CEEAplA, Ponta Delgada, Azores, Portugal*

Abstract. This paper reports two projects for supporting decisions of the Company of Electricity in Azores Islands, Electricidade dos Açores. There were several decisions to support, such as whether communications between islands should moved from the present telephone lines to VoIP, and if better models to support forecast power consumption should be adopted. The solution established integrates OLAP cubes in a data mining project, based on CRISP-DM process model. Both for strategic and more operational decisions the objective was always to get accurate data, build a data warehouse and to get tools to analyze it in order to properly inform the decision makers. These DSS's translates big CSV flat files or acquire data in real time from operational Data Bases to update a data warehouse, including importing, evaluating data quality and populating relational tables. Multidimensional data cubes with numerous dimensions and measures were used for operational decisions and as exploration tools in the strategic ones. Data mining models for forecasting, clustering, decision trees and association rules identified several inefficient procedures and even fraud situations. Not only was possible to support the necessary decisions, but several models were also displayed so that control decision makers and strategists could support new problems.

Keywords. Decision Support Systems, Data Mining, OLAP, operations efficiency.

[1] Corresponding Author: Armando B. Mendes, Azores University, Mathematical Department, 9501-801 Ponta Delgada, Azores, Portugal. E-mail: amendes@uac.pt.

1. INTRODUCTION AND PROBLEM STATEMENT

This paper 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 wether 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 criterias. For the technical criterias, a Decision Support System (DSS) was developed based on data of an external telecomunications company, and MS SQL technologies. This project generated several others related to fraud detection in using telecomunications within EDA instalations in all the nine islands. The results were published, in detail in a previous publication [1].

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

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 consuption (see for instance: [2], [3] and [4]). All these papers mention the relevance of clima in electricity demand. In [2] the clima, 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 specially 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 2007 fiscal year indicate a turnover of 138 million euros and a total of 113.995 customers spread over the nine islands of the archipelago of Azores. EDA company has 646 employees, and it’s the head of a group, which include 5 other companies, approaching 870 permanent employees. EDA has a particular complex communication system, because of the dispersion of clients spread over a wide discontinous area of 66 thousand square kilometers.

The EDA company produces a mix of energy that is still largely dominated by termoelectric power, altough 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 41,5% of all energy consumed in São Miguel during 2008.

Most of the decisions were semi-structured. Since Keen and Scott Morton [6] seminal work, has been shown that data analysis systems are very useful in the screening of these type of decisions. As the major part of the work meant to analyze 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.

This paper describes OLAP and data mining solutions implemented to support different decisions. In fact, the solution was often found useful for other related tasks. Our primary research hypothesis is that a DSS based in OLAP and data mining, developed by using a process model, can be a fast way to support unstructured and semi-structured decisions. Because it is an interpretative case study, it illustrates contemporary practice and diminishes the gap between theory and practice. This kind of work was considered one of the major faults in DSS literature by Arnott and Pervan [7]. Eom and Kim [8], also corroborating this conclusion.

2. SOLUTION APPROACH

Both decisions engaged in mentioned the projects needed learning from data, as it is defined in knowledge management literature (see for instance [9]). 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. Version 2.0 is now in discussion, see www.crisp-dm.org for more information.

Several authors have been suggesting process models for knowledge extraction from data bases (*e.g.* [10]; [11]; [12]). 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 been 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 ([13]; [14]).

In Figure 1 are depicted the six phases of the process model version 1 used in this project (see [15]). This process model can be compared with the OR decision methodology [16] 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. In this process model, there are numerous feedback loops connecting the phases. This adaptive development process is not new to DSS literature. In 1980, Keen describes continuous actions cycles that involved significant user participation [17]. As each evolutive cycle is completed, the system gets closer to its established state like an evolutive spiral [14]. The CRISP-DM project 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 Arnott and Pervan noted, data warehouse and data mining development are dominated by IT departments. As IT professionals have little experience with decision support, some basic concepts of DSS have recently being rediscovered like evolutionary development [7].

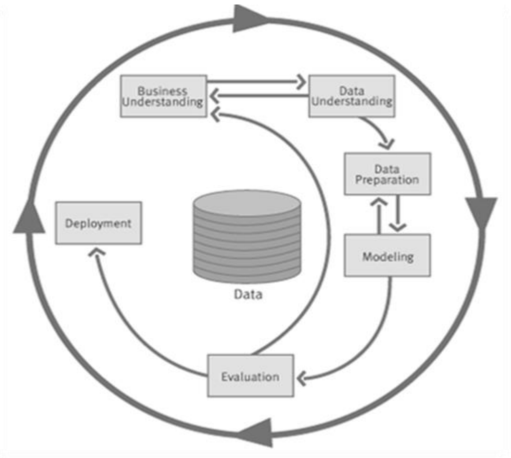


Figure 1: CRISP-DM v. 1.0 process model (reproduced with authorization from www.crisp-dm.org).

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

Several authors, like Poon and Wagner [18], recognize as a major critical factor the executive and operational sponsorship in a DSS adoption. In this project, top EDA management was the client, being the users the IT specialists that designed and developed the system with the authors. Those last ones were profoundly envolved in the system development and were also responsible for all the comunication with the client's project.

This paper presents the development process of the DSS components. The introduction presents the problem and reviews the literature. This section presents the methodology followed to develop the system. The following sections will present the application of this methodology. At the end, the results and conclusions will be presented.

3. DATA UNDERSTANDING AND EXPLORATION: THE OLAP PHASE

In both projects, the contact with management to establish purposes and patronize the projects were 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.

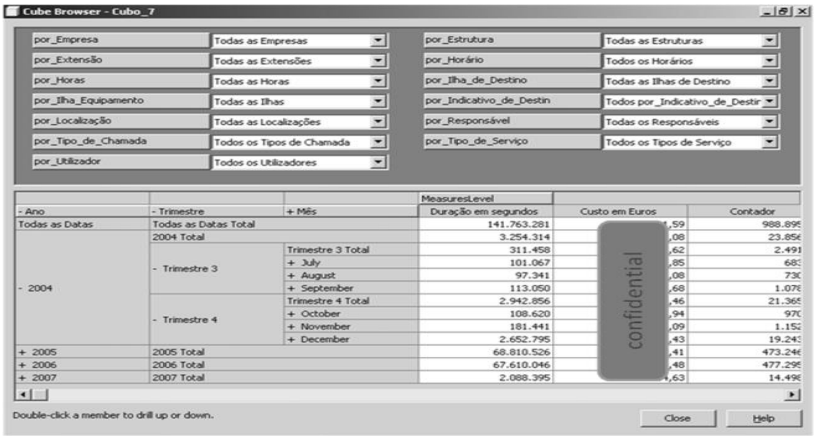
To make data exploration and dicing easy to any user, an OLAP application was implemented in both projects. This was a time consuming phase and comprehend pre-processing of the data, data exploration, data reduction and visualization. The Software tools used were based on the Microsoft SQL Server technologies, 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. 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 [20].

This phase is also very important for data quality evaluation. In both projects many of the problems described in the book by Chen [20] 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 [20] and [21]. 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.

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 2 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 2, 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.



Cube Browser - Cubo_7

por_Empresa	Todas as Empresas	por_Estrutura	Todas as Estruturas
por_Extensão	Todas as Extensões	por_Horário	Todos os Horários
por_Horas	Todas as Horas	por_Ilha_de_Destino	Todas as Ilhas de Destino
por_Ilha_Equipamento	Todas as Ilhas	por_Indicativo_de_Destin	Todos por_Indicativo_de_Destir
por_Localização	Todas as Localizações	por_Responsável	Todas os Responsáveis
por_Tipo_de_Chamada	Todos os Tipos de Chamada	por_Tipo_de_Serviço	Todos os Tipos de Serviço
por_Utilizador	Todos os Utilizadores		

			MeasuresLevel		
- Ano	- Trimestre	+ Mês	Duração em segundos	Custo em Euros	Contador
Todas as Datas	Todas as Datas Total		141.763.281	[redacted],59	988.895
- 2004	2004 Total		3.254.314	[redacted],08	23.856
	- Trimestre 3	Trimestre 3 Total	311.458	[redacted],62	2.491
		+ July	101.067	[redacted],85	683
		+ August	97.341	[redacted],08	730
		+ September	113.050	[redacted],68	1.078
	- Trimestre 4	Trimestre 4 Total	2.942.856	[redacted],46	21.365
		+ October	108.620	[redacted],94	970
		+ November	181.441	[redacted],09	1.152
		+ December	2.652.795	[redacted],43	19.243
+ 2005	2005 Total		68.810.526	[redacted],41	473.246
+ 2006	2006 Total		67.610.046	[redacted],48	477.295
+ 2007	2007 Total		2.088.395	[redacted],63	14.498

Double-click a member to drill up or down.

Figure 2: The final data cube for the communication network decision.

As you can see in Figure 2, 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 [20] and [22]).

4. 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 [19].

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 [23] 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 [19]. Other uncommon feature we would like to distinguish is the possibility of generating data tables on the fly by selecting a key attribute 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 attributes 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 attributes [19], which is unexpected in data mining software (see any data mining text book as [22]).

Microsoft Decision Trees is an algorithm that produces a tree structure defining logical rules for explaining a target attribute, using several explaining categorical or categorized attributes. In the MS implementation, it can be regarded as a generalization of Naïve Bayes algorithm or a form of Bayesian network [22]. Analysing many trees built for the call data and suing the cost attribute as a target and other many attributes 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 attributes 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 attribute 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 a 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 19.7°C 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 attributes) 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.

5. 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 [24]) 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 equipments. 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 equipments 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 equipments, 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. 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.

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 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 3, 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 [4] and [5]).

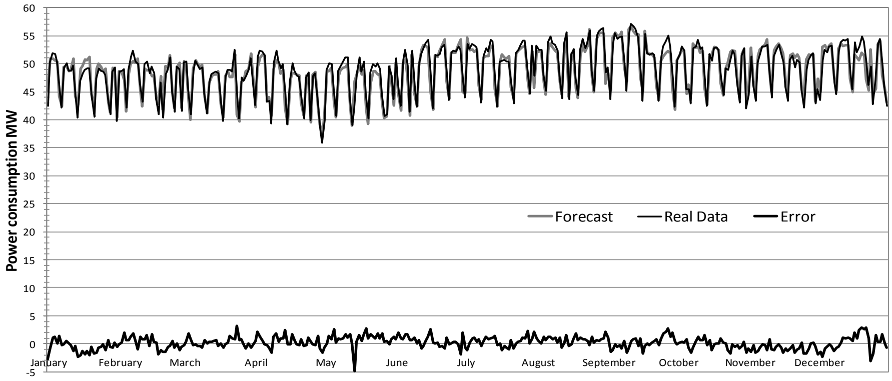


Figure 3: Mean daily power consumption for year 2007, both recorded and forecasted.

6. CONCLUSIONS

In this paper we described two main works which consist in the development of Decision Support Systems, based on business intelligence and data mining technologies. These are applications very different from the ones previously used by the authors for supporting other decisions (see for instance [25]), but it can also be very useful for supporting specific solutions. In spite of that, they are considered to be particularly adequate to data description and classification, with applications for the identification of fraud and inadequate procedures concerning communications.

Because 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, especially for ETL and data management. 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 compact and not very clear for the user. That is not a kind of software I would use in research activities, but it can be actually relevant in management context.

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 fact 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 we developed applications (as the OLAP cube) for use by several technicians and decision makers. We also organized workshop meetings for knowledge generation and transfer.

These successful projects are good examples of BI and Data Warehouse technologies as DSS generators.

7. ACKNOWLEDGMENTS

The authors wish to thank the Administration Office of the EDA, S.A. Dr. Roberto de Sousa Amaral and Dr^{a} Maria José Martins Gil all the support and consideration. We also would like to thank all the EDA professionals concerned with the success of this study, in special to Aires Ferreira and Edgar Ponceano. We also wish to thank Dr^{a} Sandra Mendes for the Azores Archipelago picture and Paulo Alfaro and José Ferreira for doing the actual work.

REFERENCES

[1] Mendes, A.B.; Ferreira, A.; Alfaro, P.J.; "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 (2009).

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

[3] 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.

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

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

[6] 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.

[7] Arnott, D.; Pervan G. (2005). "A critical analysis of decision support systems research". *J. Inform. Technol.* 20, pp. 67-87.

[8] Eom, S.; Kim, E. (2006). "A survey of decision support system applications (1995-2001)". *J. Opl. Res. Soc.* **57**, pp.1264-1278.

[9] 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.

[10] Klösgen, W.; Żytkow, J. M. (2002). "The knowledge discovery process". In Klösgen W and Żytkow J M (Eds.) Handbook of Data Mining and Knowledge Discovery, Oxford University Press: New York, USA, pp 10-21.

[11] Hand, D.J.; Mannila, H.; Smyth, P. (2001). "Principles of Data Mining". Adaptive Computation and Machine Learning, MIT Press: Cambridge, USA.

[12] 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.

[13] Clifton, C.; Thuraisingham, B. (2001). "Emerging standards for data mining". *Comput. Standards Interfaces* **23**: 187-193.

[14] 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.

[15] 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..

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

[17] Keen, P.G.W. (1980). "Adaptive design for decision support system". *Data Base* **12**, pp. 15-25.

[18] 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.

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

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

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

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

[23] 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.

[24] 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.

[25] Mendes, A.; Cardoso, M.; Oliveira, R. (2006). Supermarket site assessment and the importance of spatial analysis data. In Moutinho, L.; Hutcheson, G.; Rita, P. (Eds.) Advances in Doctoral Research in Management. World Scientific: N J , pp. 171-195.