



Centro de Estudos de  
Economia Aplicada do Atlântico

## WORKING PAPER SERIES

**CEEApIA WP No. 09/2008**

### **Suporte à Decisão em Tecnologias de Comunicação: Um caso com utilização de metodologias OLAP e Data Mining**

**Paulo Jorge Álfaro  
Armando Mendes  
Aires Ferreira**

**July 2008**

# **Suporte à Decisão em Tecnologias de Comunicação: Um caso com utilização de metodologias OLAP e Data Mining**

**Paulo Jorge Álfaro**

Universidade dos Açores (DM)

**Armando Mendes**

Universidade dos Açores (DM)  
e CEEApIA

**Aires Ferreira**

Electricidade dos Açores, S.A.

Working Paper n.º 09/2008  
Julho de 2008

## RESUMO/ABSTRACT

### **Suporte à Decisão em Tecnologias de Comunicação: Um caso com utilização de metodologias OLAP e Data Mining**

Este artigo descreve um Sistema de Apoio à Decisão capaz de fornecer informação precisa e de qualidade sobre as comunicações na Electricidade Dos Açores (EDA). A decisão imediata a apoiar consistia em saber se as comunicações entre ilhas deveriam passar para tecnologias Voice over IP (VoIP), um serviço actualmente contratado a uma empresa de comunicações externa. Num projecto de business intelligence e usando tecnologias Microsoft SQL Server, o sistema lê e pré-processa ficheiros CSV de grande dimensão, recebidas da empresa de comunicações, combina esses dados com bases de dados existentes e apresenta os resultados sobre a forma de cubos multidimensionais. Posteriormente, este trabalho foi integrado num projecto de data mining, usando a metodologia CRISP-DM, tendo sido possível além de apoiar a decisão pretendida identificar situações ineficientes e mesmo de utilização fraudulenta de equipamentos de comunicação. Vários modelos foram construídos e disponibilizados a diferentes decisores para apoiar decisões estratégicas e de controlo.

**Palavras chave:** Decision Support Systems; OLAP; Data Mining; Business Intelligence; Comunicações telefónicas.

Paulo Jorge Álfaro  
Departamento de Matemática  
Universidade dos Açores  
Rua da Mãe de Deus, 58  
9501-801 Ponta Delgada

Armando B. Mendes  
Departamento de Matemática  
Universidade dos Açores  
Rua da Mãe de Deus, 58  
9501-801 Ponta Delgada

Aires Ferreira  
Electricidade dos Açores, S.A.  
Rua Dr. Francisco Pereira Ataíde  
9504-535 Ponta Delgada

# **Suporte à Decisão em Tecnologias de Comunicação: Um caso com utilização de metodologias OLAP e *Data Mining***

Paulo Jorge Alfaro  
Universidade dos Açores, Depto. de Matemática  
Rua da Mãe de Deus, 9501-801 Ponta Delgada, Portugal, pjalfaro@gmail.com

Armando B. Mendes  
CEEApIA e Universidade dos Açores  
Rua da Mãe de Deus, 9501-801 Ponta Delgada, Portugal, amendes@uac.pt

Aires Ferreira  
Electricidade dos Açores, S.A.  
Rua Dr. Francisco Pereira Ataíde, 9504-535 Ponta Delgada, Portugal, airesfer@eda.pt,

*Working Paper de Fevereiro de 2008*

## **Resumo:**

Este artigo descreve um Sistema de Apoio à Decisão capaz de fornecer informação precisa e de qualidade sobre as comunicações na Electricidade Dos Açores (EDA). A decisão imediata a apoiar consistia em saber se as comunicações entre ilhas deveriam passar para tecnologias *Voice over IP* (VoIP), um serviço actualmente contratado a uma empresa de comunicações externa. Num projecto de *business intelligence* e usando tecnologias *Microsoft SQL Server*, o sistema lê e pré-processa ficheiros CSV de grande dimensão, recebidas da empresa de comunicações, combina esses dados com bases de dados existentes e apresenta os resultados sobre a forma de cubos multidimensionais. Posteriormente, este trabalho foi integrado num projecto de *data mining*, usando a metodologia CRISP-DM, tendo sido possível além de apoiar a decisão pretendida identificar situações ineficientes e mesmo de utilização fraudulenta de equipamentos de comunicação. Vários modelos foram construídos e disponibilizados a diferentes decisores para apoiar decisões estratégicas e de controlo.

**Palavras chave:** *Decision Support Systems*; OLAP; *Data Mining*; *Business Intelligence*; Comunicações telefónicas.

## **1. Introdução e Definição do Problema**

Este artigo descreve um caso de aplicação de metodologias de apoio à decisão para a construção de um sistema desenhado para suportar decisões sobre o investimento em infraestruturas de comunicação na Electricidade Dos Açores (EDA), a companhia eléctrica do Arquipélago dos Açores. A decisão principal consistia em saber se as comunicações inter-ilhas da EDA deveriam passar a ser efectuadas usando *Voice over IP* (VoIP), sendo actualmente subcontractadas a uma empresa de comunicações externa. Esta é uma decisão complexa e estratégica envolvendo pontos de vista técnicos e não técnicos. Para o cálculo de descriores de impacto, medidas precisas e de qualidade para os critérios técnicos, desenvolveu-se um Sistema de Apoio à Decisão com base em tecnologias *MS SQL Server* e dados provenientes de várias origens.

A EDA S.A. ([www.eda.pt](http://www.eda.pt)) é a companhia responsável pela produção, transporte e venda de energia eléctrica. Outras organizações podem produzir energia eléctrica mas têm de a vender à EDA que se encarrega da transportar e revender ao consumidor final. Dados do ano fiscal de 2006 indicam um valor de vendas total de 81 milhões de euros, 112.000 clientes dispersos pelas nove ilhas habitadas do arquipélago. A EDA tem 683 empregados, mas o Grupo EDA, que inclui 5 companhias nas áreas da energia, engenharia e serviços de consultadoria; emprega cerca de 870 trabalhadores permanentes. Apesar de a sua dimensão não ser excessiva a nível europeu, possui um sistema de comunicações complexo devido às muitas localizações numa área dispersa com 66 milhares de quilómetros quadrados.

No âmbito do projecto foram definidos como objectivos a redução de custos com as comunicações envolvendo terminais fixos pertencentes ao Grupo EDA. A situação anterior à implementação do projecto consistia na quase total ausência de informação sobre comunicações internas ao grupo, com utilização das infraestruturas pertencentes à companhia externa. Os dados identificados como necessários incluíam padrões de funcionamento, número de chamadas, duração e frequência de uso em horas de pico. Pretende-se ainda verificar se existem tendências crescentes ou decrescentes nas medidas anteriores.

Para a tomada de decisão tem-se em conta que a empresa possui conhecimento em telecomunicações e especificamente em tecnologias VoIP e *IP Telephony* e redes

cobrindo todas as ilhas com serviços IP. O Grupo EDA possui ainda cerca de 700 equipamentos telefónicos fixos com acesso a chamadas externas e internas.

Para enfrentar o problema definido, foi sugerido e aceite a utilização de um projecto de *data mining* com uma componente forte em tecnologias OLAP para exploração de dados e cálculo de medidas de descritores de impacto considerados necessários à tomada de decisão. Os objectivos definidos para este projecto incluíam a necessidade de compreender os custos envolvidos nas comunicações telefónicas e as durações das chamadas, explorados para vários níveis de agregação incluindo vários períodos temporais, destinos e origens, entre outros. Tendo em conta os objectivos definidos foi igualmente decidida a utilização da metodologia processual CRISP-DM.

O *Cross Industry Standard Process for Data Mining* (CRISP-DM) é considerado muito útil neste tipo de projectos caracterizados por, no essencial, aplicarem metodologias de *data mining*. A versão 2.0 está por esta altura em discussão e pode ser participada por todos os interessados, ver em [www.crisp-dm.org](http://www.crisp-dm.org) para mais informações. Na Figura 1 representa-se as seis fases do modelo processual de forma esquemática. Para uma descrição completa desta metodologia ver Chapman, *et al.* (2000).

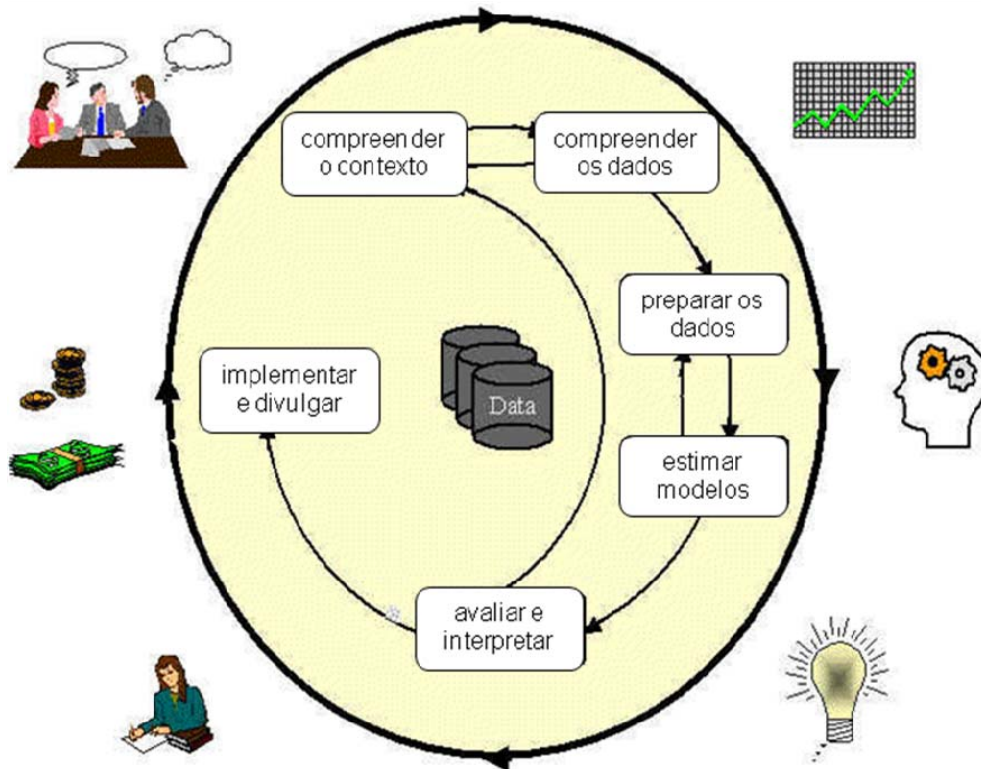


Figura 1 – Modelo de Processo CRISP-DM v.1.0 (adaptado de [www.crisp-dm.org](http://www.crisp-dm.org))

Seguindo a metodologia anterior começou-se por recolher os dados disponíveis e identificados como necessários à tomada de decisão. O conhecimento do contexto foi uma fase especialmente simples uma vez que a disponibilidade dos profissionais da EDA para recolher todos os dados e responder a todas as perguntas colocadas pelos analistas envolvidos no projecto foi total. As fases seguintes foram mais delicadas e demoradas e, logo, mais interessantes para análise de caso.

## 2. Exploração de Dados e Pré-Processamento: OLAP

Os dados da empresa de telecomunicações externa são recebidos em formato CSV (*comma separated values*), todos os meses, com cerca de 60 mil linhas correspondentes a chamadas individuais. Na Tabela 1 estão descritos os vários atributos incluídos na referida tabela de dados.

Tabela 1 – Descrição das tabelas de dados recebidas do operador externo.

Designação	Tipo de dados	Descrição
Data	MMDDAAAA	Número de mês, dia e ano
Hora	HHMMSS	Número de horas, minutos e segundos
Origem	Numérico (usado como nominal)	Identificação do equipamento que originou a chamada
Destino	Numérico (usado como nominal)	Número de telefone marcado
Tipo de serviço	Nominal	Chamada directa, Operador humano, Número especial (prefixo 808 ou 800)
Ilha	Nominal	Ilha de destino da chamada
Tipo de chamada	Nominal	Chamada de uma rede móvel, Chamada local, Outros tipos de chamadas
Período de custo	Nominal	Económico, Misto, Normal
Duração	Numérico	Duração da chamada em segundos
Custo	Numérico	Custo antes de impostos

Na sequência da metodologia CRISP-DM começou-se por entender os dados e o contexto em que são gerados. A exploração dos dados foi efectuada usando pequenas amostras de 2-3 meses e aplicações facilmente acessíveis como os pacotes

estatísticos SPSS *for Windows* e o R. Construíram-se tabelas e gráficos com estatísticas descritivas simples e os resultados foram discutidos com os profissionais da EDA. Um exemplo é apresentado na Figura 2. Nesta fase, os objectivos iniciais foram aprofundados e as estratégias de apoio à decisão foram desenhadas.

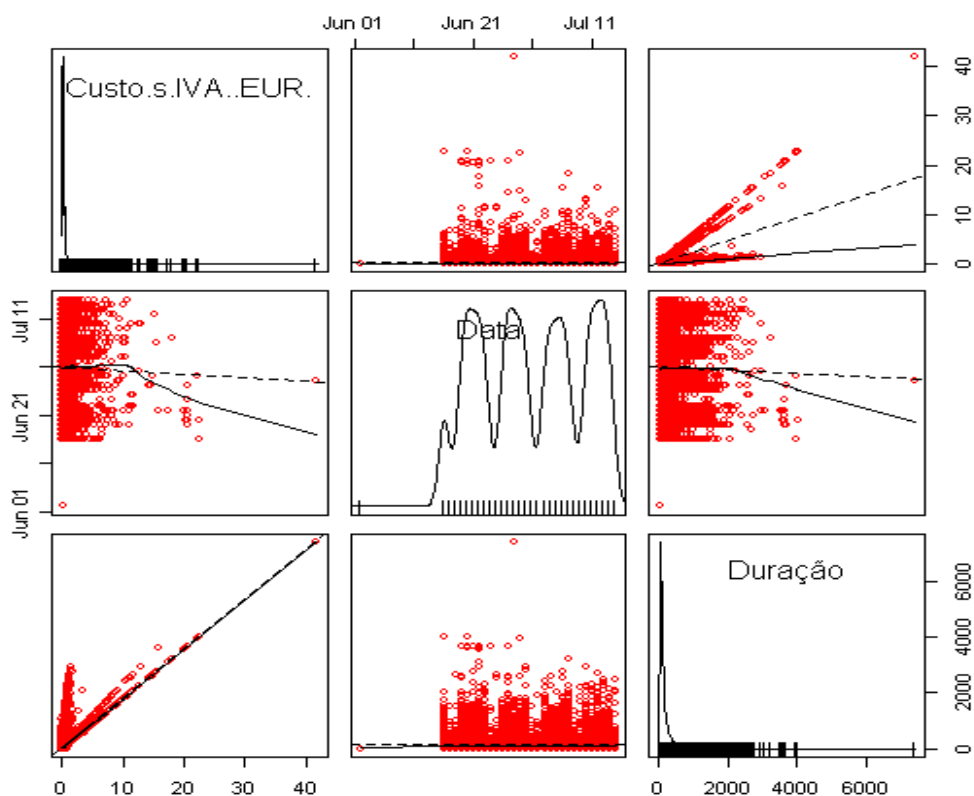


Figura 2 – Matriz de gráficos de dispersão para três meses de dados e 3 variáveis consideradas quantitativas (*software R*).

Factos como a existência de 3 relações lineares quase perfeitas entre o custo e a duração da chamada foram explicados de forma simples pelos 3 períodos de custo. No mesmo processo algumas questões interessantes se levantam, como os raros casos que violam a regra anterior. Usando um histograma das durações é possível identificar uma distribuição semelhante a uma exponencial, com a quase totalidade das chamadas de muito pequena duração, mas com algumas particularmente demoradas. Estas chamadas muito longas foram consideradas especialmente interessantes pelos profissionais da EDA.

Foram igualmente identificadas sazonalidades óbvias no número de chamadas, apresentando valores elevados durante os dias úteis da semana e reduzindo-se

fortemente durante feriados e fins-de-semana, quando apenas pessoal de manutenção se mantém activo. Sazonalidades semelhantes foram igualmente identificadas durante as 24 horas do dia. Usando gráficos com 2 anos de custos diários totais, foi possível identificar igualmente sazonalidades anuais, correspondendo a reduções de actividade durante o Verão. Identificou-se igualmente um período durante o ano de 2005, com reduzida actividade, devido a transferência de instalações da EDA, como se pode observar na Figura 3.

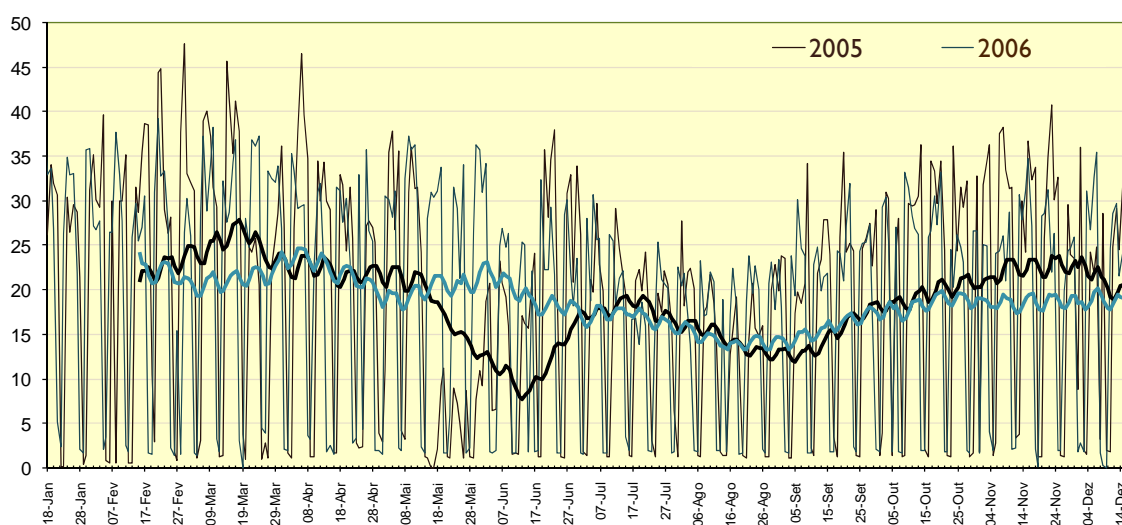


Figura 3 – Custos agregados para o dia, sobrepondo dados do ano 2005 e 2006 e linhas de médias móveis para 30 dias.

Tendo em conta que grande parte da decisão deverá ser apoiada com dados agregados e médios, foi considerado indispensável implementar um sistema OLAP (*OnLine Analytical Processing System*) para facilitar a exploração de dados futuros e gerar dados agregados para apoio a decisões.

Esta foi uma fase demorada, compreendendo o processamento das tabelas CSV recebidas mensalmente, a colocação destes dados num formato adequado à sua utilização posterior, a verificação da qualidade, e finalmente a fusão com bases dados existentes na organização. As ferramentas de *software* utilizadas foram as fornecidas pelo *Microsoft SQL Server* já conhecidas e utilizadas pelos profissionais em análise de sistemas da EDA. Os componentes mais usados foram o *Data Base Engine*, *Analysis Services* e *Integration Services from Business Intelligence (BI) Studio*.

Apesar da abundância de ferramentas, algumas realmente muito úteis, a integração de dados e a construção e gestão dos cubos de dados foi uma fase demorada e complexa, não apenas pela abundância de dados, mas também pela mudança de versão do *MS SQL server* 2000 para 2005. Foram criados fluxos de processamento (*process flows*), utilizando programação em SQL, para a preparação das tabelas de dados, como a geração de novos campos, tabelas e integração de dados (um exemplo é apresentado na Figura 4). A ferramenta de *Integration Services* foi considerada muito útil para esta tarefa e relativamente fácil de usar.

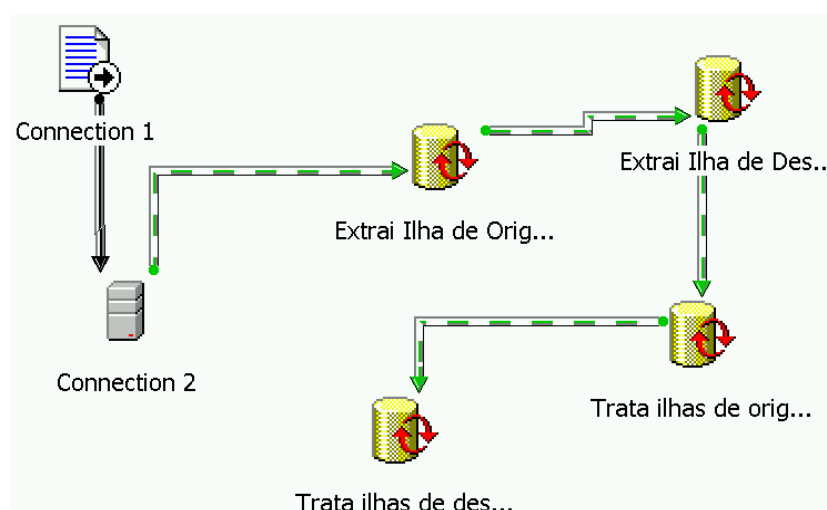


Figura 4 – Exemplo de fluxo de processamento usado para extrair a ilha de origem, de destino, corrigir erros verificando o domínio dos campos e liga-las a tabelas existentes.

Apesar de terem sido efectuados vários testes e verificados várias regras empíricas resultantes do conhecimento transmitido sobre o contexto, não foi possível identificar nenhum problema significativo de qualidade dos dados. De qualquer modo, as referidas regras foram implementadas em fluxos de processamento fáceis de usar com dados futuros. Estas incluem igualmente operações de limpeza dos dados como remoção de linhas apenas com valores nulos ou correcção de pontos decimais para vírgulas.

Em 1933, E.F. Codd (citado em Larson, 2006), um dos pais do modelo relacional para bases de dados e do *OnLine Transaction Processing (OLTP)*, propôs um novo tipo de sistema orientado para as necessidades dos analistas no apoio à decisão. A designação então proposta, *OLAP (OnLine Analytical Processing System)* mantém-se

hoje, ainda que os critérios propostos por Codd não tenham sido aceites pela comunidade em geral.

Com o *SQL Server 2005*, é introduzida a tecnologia UDM (*Unified Dimensional Model*), apresentada como de grande flexibilidade, uma vez que permite utilizar várias fontes de dados, em simultâneo se necessário, sem perder a facilidade de utilização de um sistema OLAP tradicional. Assim, seria possível usar e integrar origens de dados como *data marts*, *data warehouse* e bases de dados relacionais ou qualquer mistura destes (ver página 53 em Larson, 2006). Neste projecto, estas funcionalidades não foram todas utilizadas ou testadas. No entanto, foi possível construir um *data mart* que se actualiza periodicamente a partir de um sistema OLTP e das tabelas CSV recebidas da empresa de telecomunicações externa. Como grande parte destes dados não estavam num formato relacional, não foi possível utilizar a tecnologia UDM sem a construção de um *data mart* intermédio.

Para a construção do *data mart* foi escolhido um esquema em estrela e uma arquitectura relacional (ou ROLAP) uma vez que é apresentado como a combinação que permite menores problemas de desempenho (ver de Ville, 2001). Foram utilizadas 3 medidas, nomeadamente número de chamadas, resultante de uma contagem do número de linhas da tabela, duração e custo da chamada. Estas são quantidades numéricas, facilmente obtidas da tabela de dados fornecida pelo operador externo e fortemente relacionadas com os objectivos do projecto. A primeira delas foi apenas incluída numa fase posterior do projecto, uma vez que foi considerada relevante pelos profissionais da EDA, considerados decisores.

As dimensões são campos discretos, nominais ou ordinais, usados para definir níveis de agregação para as medidas. Um conceito muito útil do *MS. SQL Server 2005* é o de hierarquia de dimensões, o qual constitui uma forma de organização de dimensões por níveis. Por exemplo, na Figura 5, foi construída uma hierarquia relacionada com o período temporal, na sequência: ano > trimestre > mês. Muitas outras dimensões foram implementadas no cubo final, tais como a empresa do grupo EDA, o número da extensão do equipamento telefónico, a ilha, a localização do equipamento mais fina, o utilizador responsável pelo equipamento, tipo de chamada, tipo de serviço, etc. A maioria destas dimensões são directamente obtidas da tabela CSV, mas algumas outras são extraídas das bases de dados existentes, tal como toda a informação

relacionada com os equipamentos telefónicos e utilizadores ou responsáveis pelos mesmos.

Como se pode observar na Figura 5, os campos nominais podem ser utilizados tanto como dimensões de agregação, como são exemplos o ano, trimestre e mês apresentados na figura, ou como filtros, seleccionando um valor da lista pendente. Para alterar o papel desempenhado por cada dimensão basta clicar e arrastar as dimensões entre a área ao cimo e a área à esquerda das medidas.

Neste projecto, vários cubos de dados e interfaces foram construídas, num processo interactivo e evolucionário de apresentação de protótipos aos decisores e melhoria dos mesmos.

The screenshot shows the 'Cube Browser - Cubo\_7' window. At the top, there are two columns of filter settings, each with a label and a dropdown menu. Below the filters is a data table with columns for dimensions and measures. A large blue watermark with the word 'confidential' is placed over the table.

			MeasuresLevel			
- Ano	- Trimestre	+ Mês	Duração em segundos	Custo em Euros	Contador	
Todas as Datas	Todas as Datas Total		141.763.281	1,59	988.89€	
- 2004	2004 Total		3.254.314	,08	23.85€	
	- Trimestre 3	Trimestre 3 Total	311.458	,62	2.491	
		+ July	101.067	,85	68€	
		+ August	97.341	,08	73€	
	- Trimestre 4	+ September	113.050	,68	1.07€	
		Trimestre 4 Total		2.942.856	,46	21.36€
		+ October	108.620	,94	97€	
		+ November	181.441	,09	1.15€	
	+ December	2.652.795	,43	19.24€		
	+ 2005	2005 Total		68.810.526	,41	473.24€
+ 2006	2006 Total		67.610.046	,48	477.29€	
+ 2007	2007 Total		2.088.395	,63	14.49€	

Figura 5 – O aspecto final da interface do sistema OLAP.

### **3. Construção e Validação de Modelos: *Data Mining***

De facto, o projecto OLAP é muito mais do que a fase de exploração e pré-processamento da metodologia CRISP-DM. Com o cubo de dados final é possível responder a uma série de questões e fazer um diagnóstico da forma como os equipamentos telefónicos estão a ser utilizados na EDA.

Ainda assim, é igualmente claro que a preparação dos dados efectuada para a constituição do sistema OLAP é igualmente necessária para o uso de algoritmos de prospecção de dados (*data mining*). Muitos fabricantes de *software* reconhecem isto mesmo ao incluir ambas as tecnologias de apoio à decisão na mesma infra-estrutura informática. Este é o caso da Microsoft, já que o *Development Studio* inclui ferramentas tanto para *OLAP analysis services* como para *data mining*. Ambos podem usar *SQL Server Integration Services* para extrair, limpar, integrar e colocar os dados de uma forma acessível.

No projecto de prospecção de dados, utilizaram-se praticamente os mesmo dados que no projecto OLAP, na construção de tabelas de dados para aprendizagem (estimação) de modelos e para teste e validação. Na fase de teste usou-se sempre dois conjuntos de dados: 130 mil registos, correspondentes aos anos 2005 e 2006, para aprendizagem e 25 mil registos, alguns meses de 2007, para validação.

O *Business Intelligence Development Studio* do *MS SQL Server 2005* inclui 7 algoritmos de prospecção de dados, que cobrem as tarefas principais comumente utilizadas neste tipo de aplicações, tais como classificação para campos nominais, previsão para campos numéricos, segmentação para definir grupos em dados sem atributos de classificação, associação para indução de regras e análises sequenciais para a indução de regras correspondentes a uma sequência de etapas.

Foram ensaiados vários algoritmos e 4 foram identificados como sendo mais úteis: *Microsoft Naïve Bayes*, *Microsoft Decision Trees*, *Microsoft Clustering* e *Microsoft Association*. Os restantes não são aqui referidos, uma vez que foram considerados desadequados aos objectivos explicitados, inapropriados relativamente aos dados disponíveis ou simplesmente não conseguimos obter nenhum resultado interessante da sua utilização.

O *Microsoft Time Series* foi considerado um dos casos em que não foi possível obter resultados interessantes. Tendo em conta que os dados disponíveis são

essencialmente dados cronológicos, técnicas de previsão foram consideradas desde o princípio como essenciais ao estudo. No entanto, o pouco usual algoritmo de autoregressão em árvore implementado neste *software* não permite a estimação de parâmetros como os factores sazonais (ver Meek, *et al.*, 2002 para uma descrição completa do algoritmo). Por esta razão foram estimados modelos de regressão linear, usando uma aplicação estatística, incluindo variáveis binárias (mudas) para estimar os factores sazonais mensais e semanais, apresentados na Figura 6. Estes resultados foram validados pelo cálculo do coeficiente de determinação para os novos dados, obtendo-se 84%, a raiz do erro quadrado médio: 5,9, o erro absoluto médio: 5,0 e o valor absoluto médio percentual de 19%. Estes valores parecem-nos suficientemente bons, o que foi corroborado pelo conhecimento de domínio dos profissionais da EDA.



Figura 6 – Factores sazonais diários e mensais usando dados agregados para o dia incluindo os anos de 2005 e 2006.

Dos algoritmos de prospecção de dados utilizados, o *Microsoft Naïve Bayes* foi um dos mais úteis, apesar da sua simplicidade. Tal deve-se ao facto de existirem nos dados muitos atributos categóricos, especialmente adequados para a utilização deste algoritmo. Na explicação do custo da chamada, este algoritmo colocou a hora do dia

em primeiro lugar, seguido pela ilha de destino, ilha de origem e tipo de serviço. Tendo em conta apenas as chamadas de custo mais elevado, foi possível verificar que 80% são originárias da maior ilha, onde se situa a sede do grupo, com durações entre 3 e 10 minutos, 51% foram chamadas directas e 42% por operador humano (os restantes 5% são chamadas para números especiais). Este último valor foi considerado muito elevado pelos profissionais da EDA. Note-se que o algoritmo *MS. Naïve Bayes* usando no *SQL Server 2005* não considera a possibilidade de se combinarem atributos (Larson, 2006), o que é pouco comum em aplicações deste tipo (ver por exemplo: Witten and Frank, 2005).

O *Microsoft Decision Trees* é um algoritmo que constrói árvores correspondentes a modelos lógicos de classificação para um atributo categórico, usando como explicativos outros atributos categóricos ou divididos em classes. Na implementação da *Microsoft* pode ser interpretado como uma generalização do algoritmo *Naïve Bayes* ou uma forma simples de redes Bayesianas (de Ville, 2001). Foram construídas várias dezenas de árvores, utilizando vários atributos de classificação e vários atributos para agregação, tornando-se evidente a relação óbvia entre a duração da chamada e o custo. Excluindo a duração do conjunto de atributos explicativos do custo (agregação por hora) foi possível concluir que quando o destino da chamada é São Miguel (a maior ilha com metade da população do arquipélago) a maioria das chamadas não são efectuadas de forma directa, em especial as mais caras. Porque serão as chamadas usando operador humano tão comuns?

O *Microsoft Clustering* é um algoritmo para dividir os dados em grupos tendo em conta a semelhança entre entidades definida usando um conjunto de atributos. Após a definição dos grupos estes são caracterizados, resumindo-se os valores que melhor os distinguem. O algoritmo do *SQL Server 2005* apresenta a caracterização não apenas na forma tabular, mas também na forma de rede, onde os arcos e um código de cores tornam claras as relações entre os grupos. Dos muitos grupos formados deste tipo, o *cluster 6* surgiu como especialmente interessante uma vez que é caracterizado por chamadas longas, com uma distribuição estranha fora das horas de pico e, igualmente, destinos pouco usuais. Este grupo de chamadas foram consideradas suspeitas pelos profissionais da EDA.

O *Microsoft Association* é um algoritmo para indução de regras de associação. Deste algoritmo obtém-se uma lista ordenada de itens constituintes das regras, as regras

com valores de precisão e uma rede de dependência dos atributos. O algoritmo foi considerado muito útil e um dos mais discutidos nas reuniões. Por exemplo, foi possível concluir existir um forte suporte de chamadas longas com origem e destino na mesma ilha, o que parece suspeito uma vez que estas chamadas poderão ser efectuadas de forma simples por chamada directa.

Todos os modelos foram validados usando as ferramentas disponibilizadas pelo *MS SQL Server*, nomeadamente gráficos e matrizes de confusão ou classificação. Estas ferramentas comparam a precisão da classificação (ou previsão para atributos numéricos) para os diferentes modelos construídos. Os gráficos e tabelas podem demorar muito tempo a ser gerados e são úteis apenas para compara os modelos entre si e com o pior caso (classificação aleatória). Por este processo foi possível verificar que os modelos obtidos pelas árvores de regressão e *Naïve Bayes* são os que apresentam maior poder de previsão do custo da chamada (sem agregação).

Na fase de divulgação do CRISP-DM foi disponibilizado aos utilizadores um cubo OLAP e vários modelos de prospecção de dados. Foram igualmente organizadas reuniões de trabalho para transferência de conhecimento.

#### **4. Resultados e Conclusões:**

Neste artigo é descrito o desenvolvimento de um sistema de apoio à decisão, baseado em tecnologias de *business intelligence* e *data mining*. Este tipo de geradores de SAD's são bastante distintos dos utilizados anteriormente pelos autores (ver por exemplo: Mendes, *et al.*, 2006), mas são instrumentos muito potentes e úteis, em especial quando se pretende o acesso a grandes volumes de dados. Neste projecto, consideraram-se estas ferramentas muito adequadas à constituição de bases de dados para apoio à decisão, fusão de dados de várias origens, descrição de dados e identificação de ineficiências e fraudes.

De todas as análises e modelos construídos foi possível resumir a seguinte informação sobre a utilização de linhas de telefone fixo na EDA:

- As horas de pico situam-se entre as 9 e as 11 e entre as 14 e as 16 nos dias de semana, com pouca utilização durante a noite e ao fim de semana.
- Não foram identificadas sazonalidades na série de custos diários para os cinco dias úteis da semana, com valores médios muito semelhantes.

- Não foi identificada qualquer tendência de crescimento ou decréscimo dos custos das chamadas diários, nem mesmo nos períodos de pico.
- Os factores sazonais mensais indicam menor uso durante os meses de verão e próximo do fim de ano.
- O destino das chamadas mais comum são as três maiores cidades da região, com duração abaixo dos 3 minutos e custo 3-4 cêntimos.
- Os números especiais como o *call center* da EDA são pouco utilizados.

A informação anterior é relevante para a tomada de decisão em consideração. Por exemplo a existência de sazonalidades anuais significa que os equipamentos a instalar terão de ser planeados para os períodos de utilização mais intensa. A ausência de uma tendência clara permite utilizar os valores médios do passado para planear a utilização do equipamento no futuro.

Para a tomada de decisão foi ainda efectuada uma análise de custos de duas alternativas de implementação de um novo sistema e para a situação actual. A opção 1, considera um investimento mínimo com utilização dos cabos actuais e adquirindo apenas, os equipamentos necessários à implementação de VoIP. Apesar do reduzido investimento, esta opção aumentaria os custos de operação anuais em 200%, não se verificando qualquer redução no custo nas chamadas VoIP internas, devido a utilização insuficiente.

A opção 2 considera um investimento elevado (7 vezes superior ao da opção 1) numa rede nova, mas mesmo assim continua a aumentar os custos de operação relativamente à situação actual em 165%. A situação actual apresenta assim ambos os custos de operação e de investimento mais reduzidos, do que qualquer uma das restantes opções.

A conclusão é que as duas opções de mudança em consideração surgem como pouco atractivas, pelo que apenas aspectos não considerados nesta avaliação, como uma utilização não espectável neste momento com aumento do volume de comunicações, ou a procura de uma imagem tecnológica da empresa pode justificar a mudança.

Além do suporte à decisão anterior, foram ainda identificadas ineficiências no sistema de comunicações da EDA ou mesmo fraudes, como o elevado número de chamadas

de longa duração não relacionadas com a actividade da empresa. Foi possível igualmente identificar problemas de configuração do sistema automático de encaminhamento de chamadas que resultava num aumento das chamadas externas, mais dispendiosas. Foi ainda possível identificar equipamentos telefónicos não utilizados, mas com custos de assinatura. Destas actividades de identificação de falhas vários equipamentos terminais foram eliminados e algum tráfego fantasma foi reduzido.

No entanto, o conhecimento com mais valor descoberto por este projecto, passou pelo elevado número de chamadas indirectas identificadas, com utilização do operador humano para contornar o actual sistema de controlo. Efectuando uma chamada indirecta, a ligação entre a origem e o destino da chamada é mais difícil de estabelecer. Este conhecimento levou à definição de novas regras de operação dos operadores humanos e ao ajuste do sistema de controlo de chamadas.

Este projecto, muito bem sucedido, poderá vir a ser continuado num futuro próximo, sendo agora mais orientado para actividades de controlo de falhas e optimização dos processos utilizados nas comunicações e actividades relacionadas.

### **Agradecimentos**

Os autores agradecem a atenção e colaboração prestada pela administração da EDA S.A., nas pessoas do Dr. Roberto de Sousa Amaral e da Dr<sup>a</sup> Maria José Martins Gil. Agradece-se igualmente a todos os profissionais da EDA envolvidos e responsáveis pelo sucesso deste estudo, com uma menção muito especial ao Eng. Edgar Ponceano.

## Referencias

- 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*. Boca Raton: CRC Press.
- Larson, B. (2006). *Delivering Business Intelligence with MS SQL Server 2005*. Emeryville: McGraw-Hill.
- Meek, C., Chickering, D. M. and Heckerman, D. (2002). Autoregressive tree models for time-series analysis. In *Proceedings of the 2<sup>a</sup> ed. of the Int. SIAM Conference on Data Mining*. Arlington: SIAM, pp 229-244.
- Mendes, A., Cardoso, M. and Oliveira, R. (2006). Supermarket site assessment and the importance of spatial analysis data. In Moutinho, L., Hutcheson, G. and Rita, P. (Eds.) *Advances in Doctoral Research in Management*. N.J.: World Scientific, pp 171-195.
- de Ville, B. (2001). *Microsoft Data Mining: Integrated business intelligence for e-commerce and knowledge management*. Boston: Digital Press.
- Witten, I. H. and Frank, E. (2005). *Data Mining: Practical machine learning tools and techniques*. The Morgan Kaufmann Series in Data Management Systems, San Francisco: Morgan Kauffman, 2<sup>nd</sup> edition.