Business Intelligence no Suporte a Decisões sobre Comunicações: Descrição de um Caso

Armando B. Mendes
CEEAplA e Universidade dos Açores
amendes@uac.pt

Paulo Jorge Alfaro
Universidade dos Açores, Depto. de Matemática
pjalfaro@gmail.com

Aires Ferreira
Electricidade dos Açores, S.A.
airesfer@eda.pt

Resumo

O projecto descrito tem o objectivo de apoiar decisões de investimento em infraestruturas de comunicação na Electricidade Dos Açores (EDA), a empresa responsável pela geração, transporte e venda de corrente eléctrica na Região Autónoma dos Açores. 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 em regime de *outsourcing*. Foi estabelecido um projecto de *business intelligence*, usando tecnologias OLAP do *Microsoft SQL Server*, para ler e pré-processar os ficheiros CSV de grande dimensão, combinar esses dados com bases de dados existentes e apresentar os resultados sobre a forma de cubos multidimensionais. Posteriormente, implementaram-se igualmente algoritmos de *data mining*, integrando na metodologia CRISP-DM as duas técnicas utilizadas. Construindo vários modelos foi possível, além de apoiar a decisão pretendida, identificar situações ineficientes e mesmo fraudulentas. Os modelos construídos foram ainda disponibilizados aos decisores estratégicos e de controlo, assim como toda a estrutura de reutilização, manutenção e realimentação que suporta o OLAP e os modelos de *data mining*.

Palavras-chave: *Decision Support Systems; OLAP; Data Mining; Business Intelligence*; Telecomunicações.

Title: Business intelligence in decision-support on communications: Describing a case

Abstract

This project addresses decisions of investment on communication infrastructures in Electricidade dos Açores (EDA), the local Electric Company in the Azores Islands. The main decision was that EDA communications should be moving to Voice over IP (VoIP) from present telephone lines, outsourced to an external communications company. At the beginning, a business intelligence project was set, with the objective of getting data from the communications company and analyzing it in order to offer useful information to decision makers. The system uses Microsoft SQL server technologies to establish an OnLine Analytical Processing (OLAP) application. It

translates big CSV flat files in a ROLAP infrastructure and presents the results as multidimensional data cubes. Latter some data mining models were implemented and both techniques were incorporated in the CRISP-DM process model. Different models identified several inefficient procedures and even fraud situations as long as supporting the investment decision. These models as long as all the technology developed for gathering data, maintain an manage the OLAP cubes and data mining models were made available to control and strategic decision makers.

Keywords: Decision Support Systems, OLAP; Data Mining; Business Intelligence; Business Communications.

1 O Contexto e a 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 investimentos 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 subcontratadas 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.

Para a descrição do contexto onde o problema surge é importante compreender que a EDA S.A. (www.eda.pt) é a companhia responsável pela produção, transporte e venda de energia eléctrica na Região Autónoma dos Açores (RAA). Dados do ano fiscal de 2007 indicam cerca de 112.000 clientes dispersos pelas nove ilhas habitadas do arquipélago e 870 trabalhadores permanentes. A EDA possui um sistema de comunicações complexo devido às muitas localizações numa área dispersa com 66 milhares de quilómetros quadrados. Para a tomada de decisão é relevante o facto de a empresa possuir 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.

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 a uma companhia externa. Os dados identificados como necessários incluem 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.

Num levantamento de publicações apresentado por [Eom e Kim, 2006] foram identificadas várias aplicações de conceitos de SAD na indústria e na gestão de operações, numa grande variedade de problemas. Entre elas, Eom e Kim incluem 4 aplicações no desenho de redes de comunicações. Nestes artigos, como é exemplo o trabalho muito completo apresentado por Cortes e colaboradores [Cortes *et al.*, 2001], o objectivo principal é o desenho da rede de

telecomunicações, descrevendo-se principalmente algoritmos de optimização em rede. Outros trabalhos na indústria de telecomunicações, como o apresentado Tcha e Choi [Tcha e Choi, 1999] são essencialmente problemas de alocação de recursos. Nenhum destes é semelhante ao caso apresentado neste estudo, uma vez que se pretende decidir qual a melhor tecnologia para utilizar numa rede com desenho conhecido. Mesmo na revisão de bibliografia recente apresentada por Even Kobbacy e colaboradores [Kobbacy *et al.*, 2007] não foi possível identificar nenhuma aplicação semelhante à descrita neste trabalho.

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) revelou-se muito útil neste tipo de projectos caracterizados por, no essencial, aplicarem metodologias de *data mining* à resolução de problemas. Na Figura 1 representa-se esquematicamente as seis fases do modelo processual. Tal como em todas as restantes metodologias propostas, também esta apresenta imensos ciclos e retornos entre as fases enumeradas, a que alguns autores chamam a espiral de modelação e extracção de conhecimento [Lavrač *et al.*, 2004]. Para uma descrição completa desta metodologia ver [Chapman et *al.*, 2000].

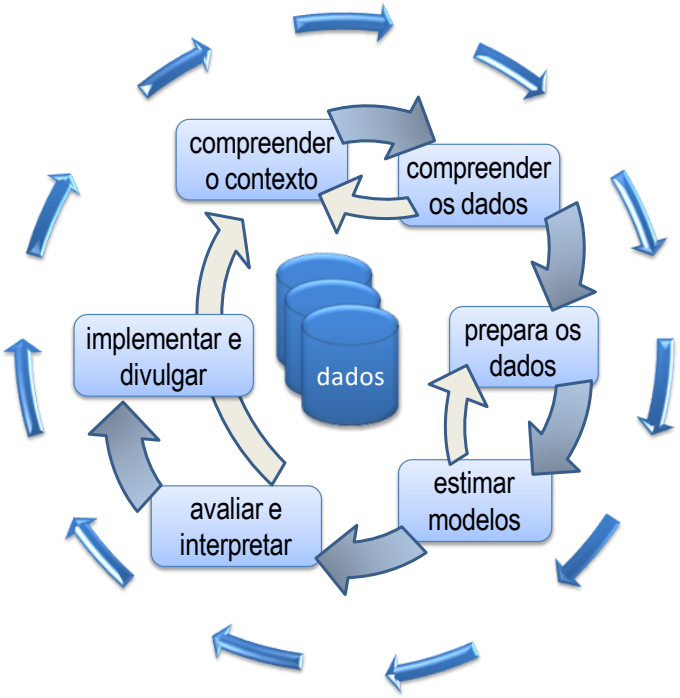

Figura 1 – Modelo Processual CRISP-DM v.1.0·
(reproduzido e adaptado com autorização de www.crisp-dm.org)

Seguindo a metodologia, começa-se por compreender o problema e o contexto onde surge, do ponto de vista do destinatário do projecto. Esta fase inclui o desenvolvimento de uma definição técnica do problema, objectivos a atingir e um plano de acção. A fase dois compreende a recolha de dados e, frequentemente, a integração de diversas fontes de dados, nem todas usando tecnologias de bases de dados ou *data warehouse*. Esta fase é igualmente responsável por uma avaliação prévia da qualidade dos dados, que pode incluir algumas análises simples.

A preparação e pré-processamento é uma fase essencial, uma vez que muitas vezes os dados não se apresentam no formato adequado. Deste modo, a preparação dos dados pode ser entendida como o conjunto de actividades que importam, transformam os dados e os ligam a dados existentes nas bases de dados transaccionais. As tarefas a empreender são além das habituais ETL - *Extraction, Transformation, Loading* ou extracção, transformação e carregamento, também tarefas de pré-processamento como a selecção de atributos, criação de novas dimensões e medidas por operações sobre as existentes e actividades de redução, transformação e limpeza de dados.

Na fase de modelação e análise de dados, são utilizadas diversas técnicas de aprendizagem ou algoritmos que permitem obter modelos alternativos. Os resultados são avaliados com métodos de validação ou teste. Durante esta fase, é muito frequente ter de voltar a efectuar actividades de pré-processamento, uma vez que podem ser identificadas variáveis com problemas de multicolinearidade ou observações atípicas ou influentes, por exemplo. Após a fase de modelação, obtêm-se modelos tecnicamente correctos. Mas serão correctos quando confrontados com o conhecimento de domínio? A fase 5 responde a esta questão, efectuando tarefas como a confrontação dos resultados com conhecimento prévio e a revisão dos passos efectuados para a construção do modelo. Esta última actividade serve para confirmar que nenhum aspecto foi esquecido e que cada decisão intermédia contribui para os objectivos do projecto. No final da fase 5, é necessário decidir quanto à implementação ou não dos modelos e conhecimento gerado.

Na fase 6, a divulgação e implementação pode ser tão simples como a escrita de um relatório, ou tão complexa como a criação de uma aplicação integrada no sistema de informação ou de *data warehouse* que permita apoiar decisões com base no conhecimento gerado. Em qualquer dos casos, inclui o registo e divulgação do conhecimento para que projectos futuros possam beneficiar da sua utilização.

Este artigo, tal como o projecto realizado, segue de perto a metodologia apresentada e descreve os problemas e soluções encontradas na sua aplicação, nomeadamente considera a integração de tecnologias OLAP na fase de pré-processamento e exploração dos dados.

2 Pré-Processamento e Exploração de Dados: OLAP

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 foi total. As fases seguintes foram mais delicadas e demoradas e, logo, mais interessantes para análise de caso.

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 dos 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. Dois exemplos são apresentados na Figura 2. Nesta fase, os objectivos iniciais foram aprofundados e as estratégias de apoio à decisão foram desenhadas.

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 indicados no primeiro gráfico da Figura 2 como económico, normal e misto. No mesmo processo algumas questões interessantes se levantam, como os raros casos que violam a regra anterior. Após novas explorações de dados foi possível explicar todos os casos de afastamento do custo esperado por telefonemas para números especiais com indicativos 707 ou 808 e telefonemas para redes móveis ou para destinos internacionais.

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.

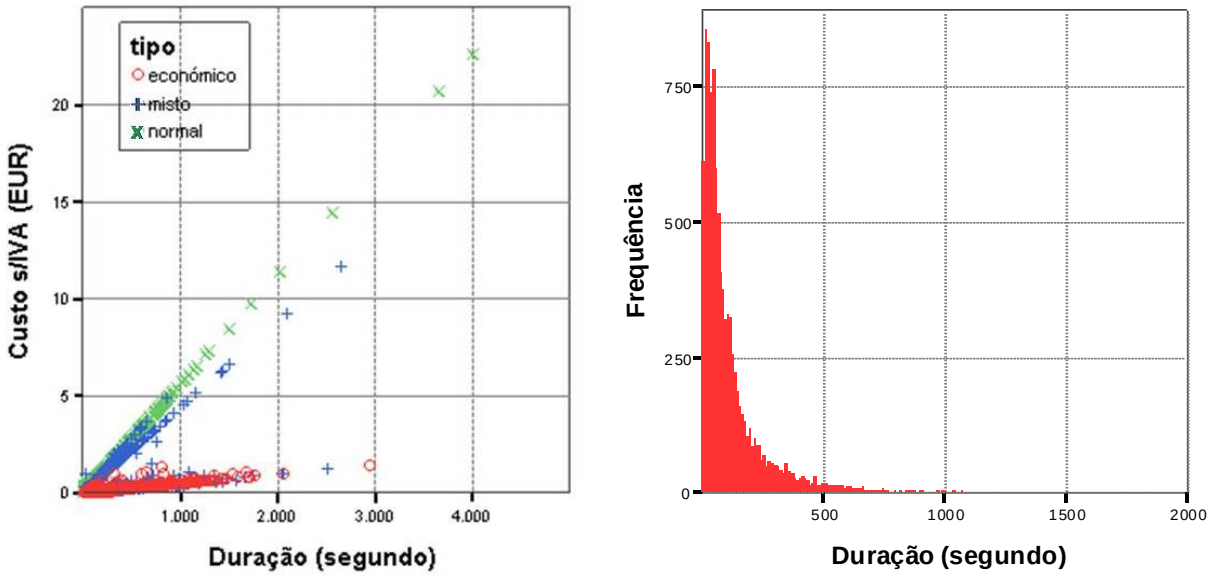

Figura 2 – Gráfico de dispersão evidenciando a relação entre o custo e a duração e histograma da duração em segundos, para 20% das chamadas escolhidas aleatoriamente de três meses de dados.

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 o 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.

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. As ferramentas de *software* utilizadas foram essencialmente as disponibilizadas 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. A ferramenta de *Integration Services* foi considerada muito útil e relativamente fácil de usar desde que se domine a linguagem SQL. 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.

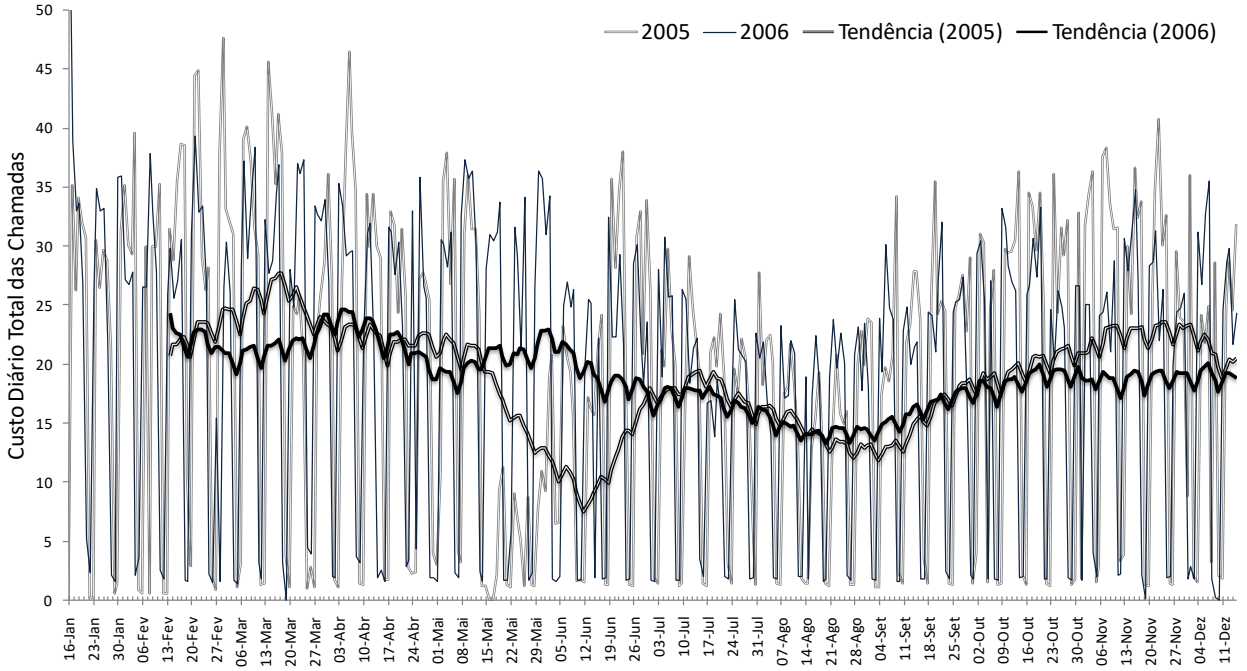

Figura 3 – Custos agregados para o dia, sobrepondo dados do ano 2005 e 2006 e linhas de médias móveis para os 30 dias anteriores.

No exemplo simples apresentado na Figura 4, começa-se por verificar a estrutura da base de dados principal no servidor em termos de consistência (passo *Check Database Integrity Task*). No passo seguinte, é feita uma ligação ao ficheiro de dados de comunicações (tabela CSV) e os dados pertinentes são copiados para a tabela relacional principal (*Transform Data*), como por exemplo os dois dígitos da hora do campo "hora". Em seguida, os valores dos campos "Origem" e "Destino" são comparados com os números dos equipamentos existentes na tabela de "Equipamentos". Esta comparação permite identificar correctamente a ilha de origem ou de destino das comunicações e deste modo preencher na tabela relacional o campo "ilha origem" ou "ilha destino". Nos dois passos seguintes, os valores "null" atribuídos quando a origem ou o destino não é identificável, são marcados como provenientes do exterior. Nos dois passos seguintes confirma-se se a chamada foi efectuada para um número especial (prefixos 800 e 808). Por fim a nova tabela é integrada com a tabela de factos e as tabelas das dimensões do armazém de dados. Note-se que o fluxo final implementado é bastante mais complexo apresentando várias dezenas de passos.

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.

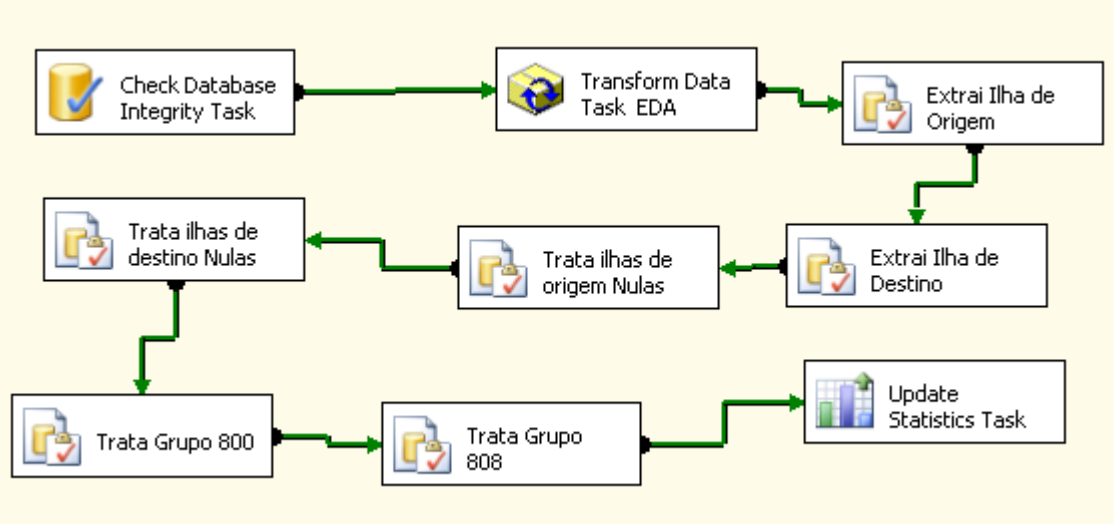

Figura 4 – Exemplo de fluxo de processamento usado para extrair a ilha de origem, de destino, corrigir erros verificando o domínio dos campos e liga-las a tabelas existentes.

Em 1993, 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 dado como *data marts*, *data warehouses* e bases de dados transaccionais 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 [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.

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.

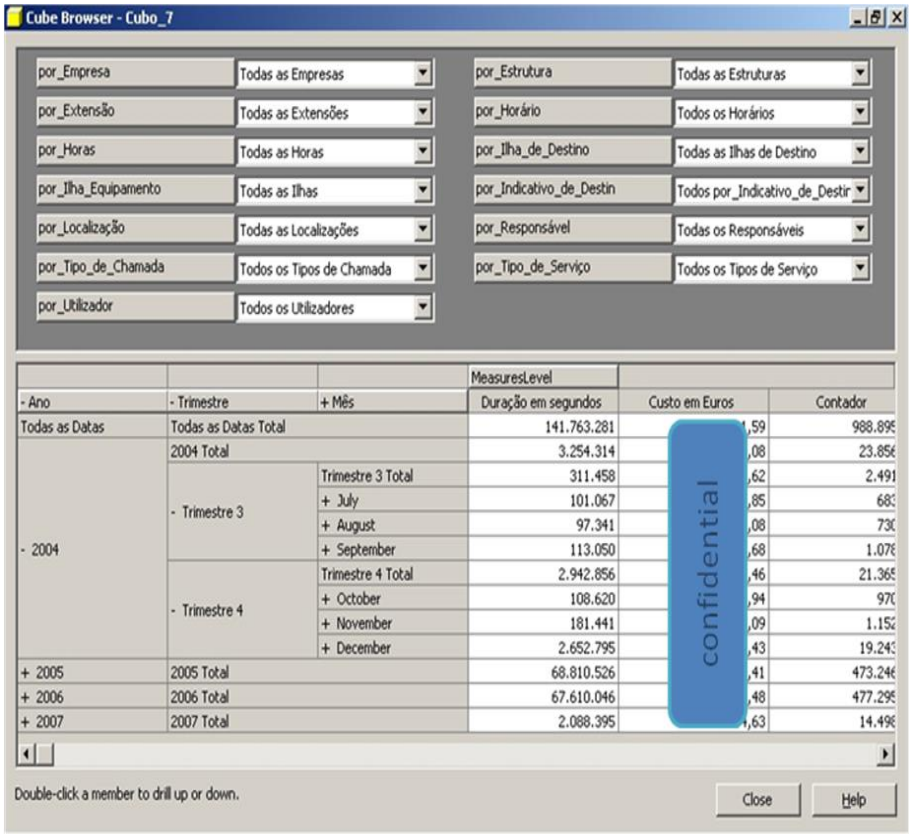

Figura 5 – O aspecto final da interface do sistema OLAP.

3 Construção e Validação de Modelos: *Data Mining*

De facto, o sistema OLAP permite 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.

Para construção dos modelos de prospecção de dados, utilizaram-se praticamente os mesmo dados já tratados para a construção do sistema OLAP. Na fase de teste usaram-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 comummente utilizadas neste tipo de aplicações, tais como classificação para campos nominais, previsão para campos numéricos, segmentação para definir grupos em dados sem atributos de supervisão, 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 foi possível 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 ou R^2 para os dados de 2007, obtendo-se 84%. Outras medidas de qualidade como a raiz do erro quadrado médio: 5,9, o erro absoluto médio: 5,0 e o erro absoluto médio percentual de 19%; confirmam a precisão das estimativas, o que foi ainda corroborado pelo conhecimento de domínio dos profissionais da EDA.

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* usado 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]) e resume o algoritmo a pouco mais do que uma análise descritiva univariada.

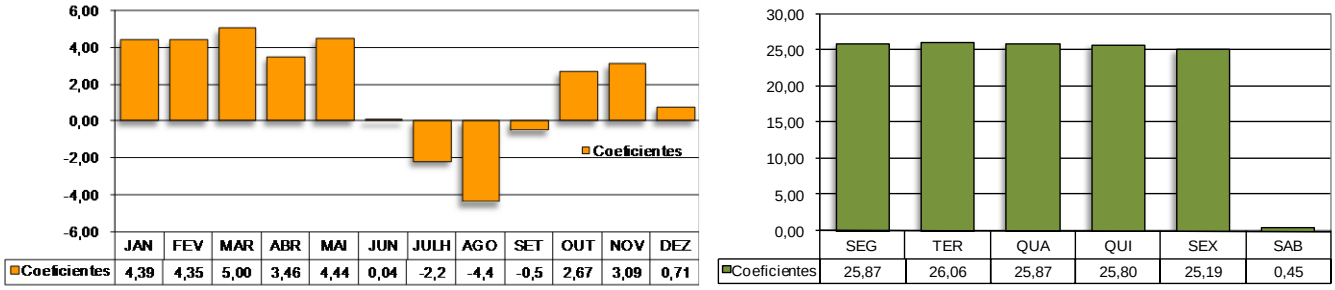

Figura 6 – Factores sazonais diários e mensais usando custos agregados para o dia incluindo os anos de 2005 e 2006.

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.

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. Os restantes *clusters* identificados foram considerados menos relevantes por apenas confirmarem o que já era conhecido.

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 que existe um forte suporte de chamadas longas com origem e destino na mesma ilha efectuadas via operador humano, 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 comparar os modelos entre si e com o pior caso de 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 *deployment* do CRISP-DM foram disponibilizados 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 são bastante distintos dos utilizados anteriormente pelos autores (ver por exemplo: [Mendes et *al.*, 2006]), mas são instrumentos de elevada capacidade e úteis para lidar com dados complexos ou pouco estruturados, 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 construção de modelos para identificação de ineficiências e fraudes.

De todas as análises e modelos construídos foi possível extrair o seguinte conhecimento 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 no tempo próximo do fim de ano.

- Os destinos das chamadas mais comuns são as três maiores cidades da região, com duração abaixo dos 3 minutos e custos de 30 a 40 cêntimos.
- Os números especiais e em especial 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 efectuada uma análise dos custos de duas alternativas de implementação de VoIP, comparando-se não só os custos com a manutenção e exploração da infra-estrutura actual (*hardware, software* e aluguer de circuitos), mas também os custos das chamadas telefónicas, de todas as centrais. A opção 1, considera um investimento mínimo mantendo as ligações actuais e adquirindo apenas placas de rede para as centrais telefónicas existentes e efectuando um estudo para a necessidade de *upgrade* de todos os circuitos da EDA. Com a aquisição das placas de rede para as centrais é possível passar a maior parte das chamadas para a rede WAN da EDA diminuindo os custos de exploração com o prestador de serviço telefónico. Na opção 2 considera-se uma renovação integral da infra-estrutura existente, com impacto e custos de migração em vários serviços prestados pela EDA, nomeadamente o centro de atendimento (*Call Center*).

Efectuando um balanço entre a eventual implementação dessas duas tecnologias verifica-se que na opção 1 existe um custo de implementação mais baixo, no entanto manter-se-ia sempre o custo de exploração de manutenção das centrais, bem como o custo de upgrade dos circuitos. Na opção 2, existirá um custo de implementação muito mais elevado aproximadamente 165% da opção 1. No entanto deixaria de existir custos com as ligações telefónicas internas e custo com a manutenção e exploração das centrais telefónicas. Em qualquer uma das opções, é necessário efectuar um *upgrade* dos circuitos, que nas condições actuais do mercado, terão sempre um custo de exploração (aluguer) superior aos custos actuais com as comunicações. A decisão teria de ser baseada numa estratégica de renovação dos equipamentos e serviços, não podendo ser explicada unicamente com uma redução de custos imediata. A administração decidiu-se pela opção 2, tendo-se decidido em Fevereiro de 2008 contratar em regime de *outsourcing* os serviços de WAN e de voz, prevendo-se a curto prazo a remodelação integral da infra-estrutura e a implementação de serviços de VoIP.

Este projecto foi considerado muito bem sucedido. De facto, ainda não é possível confirmar a esperada redução de custos pela substituição de equipamentos e utilização da tecnologia VoIP, por estas alterações ainda não estarem concluídas. Ainda assim, este trabalho permitiu não apenas suportar a decisão anterior de forma avaliada como muito satisfatória, mas também a identificação de ineficiências no sistema de comunicações da EDA ou mesmo fraudes. Através do conhecimento detalhado do tráfego e dos custos, os utilizadores passaram a ter conhecimento dos seus consumos, o que de forma indirecta contribui para uma maior contenção na utilização pessoal dos telefones. Foram identificados as excepções e os picos de consumo, apuradas as razões e definidas políticas que contribuíram para a redução de custos. Foram também implementados mecanismos de roteamento das chamadas internas para a rede móvel, que contribui para uma redução substancial dos custos. Um elemento bastante visível da

concretização dos objectivos, é a redução significativa dos custos com chamadas telefónicas que desde o início de 2007 teve uma redução superior a 60% (sessenta por cento).

No entanto, o conhecimento de maior 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. Por exemplo, tanto no reencaminhamento automático como na utilização do operador humano, passou-se a não permitir (com algumas excepções) ligações indirectas para chamadas de curta distância, nomeadamente entre ilhas ou dentro da mesma ilha.

Num futuro próximo, pretende-se desenvolver mais modelos de *data mining* orientados 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 Drª 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. Os autores agradecem ainda as sugestões de dois *referees* anónimos, que muito contribuíram para a clareza do texto apresentado.

Bibliografia

Chapman, P., Clinton, J., Kerber, R., Khabaza, T., Reinartz, T., Shearer, C. e 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.

Cortes P., Onieva L., Larrañeta, J., e Garcia, J.M. (2001). Decision support system for planning telecommunication networks: A case study applied to the Andalusian region. J. Operational Research Society,52, pp 283-290.

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

Eom S. e Kim E. (2006). A survey of decision support system applications (1995-2001). J. Operational Research Society, 57, pp 1264-1278.

Kobbacy, K.A.H., Vadera, S. e Rasmy, M.H. (2007). AI and OR in management of operations: History and trends. J. Operational Research Society, 58, pp 10-28.

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

Lavrač, N., Motoda, H., Fawcett, T., Holte, R., Langley, P., e Adriaans, P., (2004). Introduction: Lessons learned from data mining applications and collaborative problem solving. Machine Learning, 57 (1-2), 13-34.

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

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

Tcha, D.-W. e Choi, J.-S. (1999). Comparative economic analysis between direct and indirect wiring in the copper-based local loop. J. Operational Research Society, 50, pp 531-535.

Witten, I. H. e Frank, E. (2005). Data Mining: Practical machine learning tools and techniques. The Morgan Kaufmann Series in Data Management Systems, San Francisco: Morgan Kauffman, 2nd edition.