**INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA**

**Área Departamental de Engenharia de Electrónica e Telecomunicações e de Computadores**

# Análise de Eficiência Energética de Transportes Rodoviários

JOSÉ ANTÓNIO DIAS CORREIA DE ALMEIDA
(Bacharel em Engenharia Electrónica e Telecomunicações)

Projecto para obtenção do grau de Mestre em Engenharia Informática e de Computadores

Orientador:

Professor Adjunto Doutor João Carlos Amaro Ferreira

Júri:

Presidente: Professor Adjunto Mestre Vitor Jesus Sousa de Almeida

Vogais:

Professor Coordenador Doutor Helder Jorge Pinheiro Pita

Director I&D da TECMIC Pedro Alexandre Vasconcelos Marques

Setembro de 2013

*"Per ardua ad astra"*

## Nota Biográfica

Nascido em Agosto de 1968, José António Dias Correia de Almeida é filho único, casado e pai de três crianças.
Desde novo começou a interessar-se por tecnologias de informação tendo sido incentivado por uma amigo da família, Eng.º Canossa, a iniciar-se nas lides informáticas com um glorioso ZX Spectrum aos 15 anos.
Este interesse levou-o como trabalhador-estudante à frequência do curso de Bacharelato em Engenharia Electrónica e de Telecomunicações - Ramo de Sistemas Digitais, que concluiu em Julho de 1994.
Tendo desempenhado funções de Engenheiro de Sistemas Informáticos na Direcção de Manutenção e Engenharia da TAP Portugal até Março de 1997, reorientou a carreira profissional para o desempenho de funções mais estreitamente ligadas ao cerne da actividade de manutenção aeronáutica, desempenhando actualmente funções de gestão técnica de frota de aeronaves A330 e de sistemas de Comunicações e Entretenimento a Bordo de Aeronaves (IFEC).
Participa também na definição e preparação de entrada em serviço de aeronaves AIRBUS A350. As aeronaves de última geração têm arquitectura de sistemas aviónicos centrada em redes, são extremamente ricas em dados, suscitam a necessidade de utilização e o conhecimento de Sistemas de Extracção de Conhecimento que são motivadoras deste trabalho.

## Agradecimentos

Agradeço ao meu orientador Professor Doutor João Carlos Amaro Ferreira o apoio, disponibilidade, orientação avisada, paciência, confiança e incentivo dispensados ao longo deste projecto.

Agradeço à TECMIC, S.A. a disponibilidade dos dados e aos Engenheiros Fernando Pão-Mole e Pedro Marques a colaboração prestada.

Agradeço aos que me formaram na minha *alma mater*, ao Engenheiro Mário Araújo pelo encorajamento e aos colegas dos diferentes grupos com os quais percorri este caminho, Vasco Silva, Adelaide Alhinho, Ilesh Gamanbhai, José Luis Paulino, João Ferreira e Ricardo Fernandes pelo apoio mútuo e espírito de equipa.

## Dedicatória

Ao Afonso, ao Dinis, à Leonor e à...

*Paula,*
*“Stat rosa pristina nomine, nomina nuda tenemos”,*
*Umberto Eco*

**Abstract**

A operação de sistemas de transporte público rodoviário em ambiente urbano de forma eficiente, minimizando a energia despendida, é relevante pelo impacto no ambiente, satisfação no serviço prestado e contribui para a optimização de custos de operação. Foi estabelecida uma parceria de colaboração entre o Instituto Superior de Engenharia de Lisboa (ISEL) e a empresa TECMIC, S.A. que desenvolve soluções de gestão de frotas de veículos automóveis pesados, da qual surgiu o presente projecto de trabalho Mestrado. O âmbito deste projecto de trabalho é a aplicação de métodos de extracção de conhecimento à informação existente na base de dados de parâmetros das viaturas, recolhidos aquando do acontecimento de um conjuntos de factores que espoletam o registo, por forma a obter conhecimento de valor para a gestão da operação na optimização da utilização e dispêndio de energia associado.
Simultaneamente pretende-se identificar padrões de utilização por condutor, por veículo, por tempo ou data e outras dimensões que se venham a mostrar relevantes.

Palavras-chave: Extracção de Conhecimento; Transportes Públicos Rodoviários; Autocarros; Data Warehouse; Data Mining; Armazém de Dados; Extracção de Conhecimento de Dados; Padrões

Esta dissertação foi escrito em LaTeX de acordo com a ortografia anterior ao Acordo Ortográfico de 1990.

**Abstract**

Energetically efficient operation of bus based public transportation systems is relevant to the environmental impact, service satisfaction and contributes to operational costs optimization. In this scope, a cooperation partnership between Instituto Superior de Engenharia de Lisboa (ISEL) and TECMIC, S.A., a company that provides fleet management systems as been set, from which the present Master degree in Informatics and Computers project work arose. The project work scope is the aplication of knowledge discovery methods to the existing vehicule parameters database, which are collected in event-driven basis, as to extract knowledge of value to the management of the operation by optimizing the operation and associatted energy waste. Simultaneously identification of utilization patterns by driver, by time or date and any other relevant dimensions, is intended.

Keywords: Knowledge Discovery in Databases; Public Transportation; Buses; Data Warehouse; Data Mining; Knowledge; Patterns

This dissertation was written in LaTeX.

# Conteúdo

# Lista de Figuras

# Capítulo 1

# Introdução

Neste capítulo apresentam-se as motivações e contexto, problema, objectivos deste trabalho de modo a permitir a familiarização com o ambiente e enquadramento que assistirá ao maior detalhe prosseguido em capítulos subsequentes deste relatório, com cuja organização se concluirá.

## 1.1 Motivação

De acordo com a Agência Internacional de Energia [9, pág 108], o sector dos transportes é responsável pelo maior consumo de petróleo e assim se manterá de acordo com as previsões, aumentando mesmo a sua quota de consumo de 53% em 2009 para 60% em 2035, conforme se mostra na Figura 1.1.

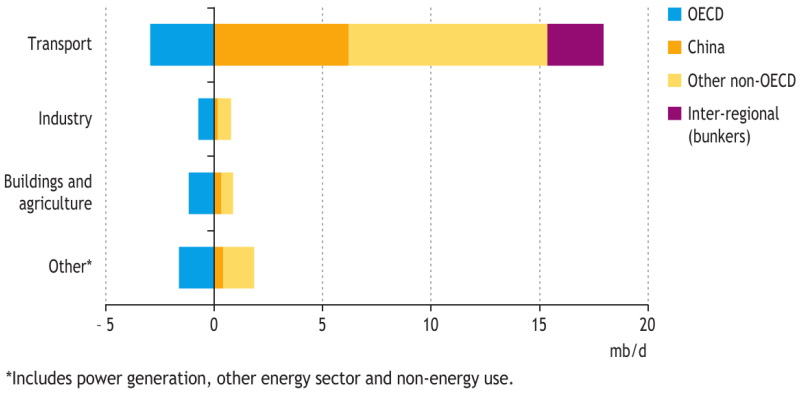



Figura 1.1: Evolução da procura primária de petróleo por sector e região, 2009-2035 [9].

## 1.2 Familiarização

Este aumento de consumo implica a procura de ganhos de eficiência na utilização desta fonte de energia e a mesma entidade emitiu em 2008 um conjunto de recomendações aos seus Estados membros para utilização eficiente, de forma a suscitar mudanças de política sectoriais.

No sector de transportes a previsão da evolução de consumo de combustível por tipo de transporte identifica o rodoviário como o sub-sector com a maior percentagem de consumo de combustível [9, pág 109], como documentado pela Figura 1.2.

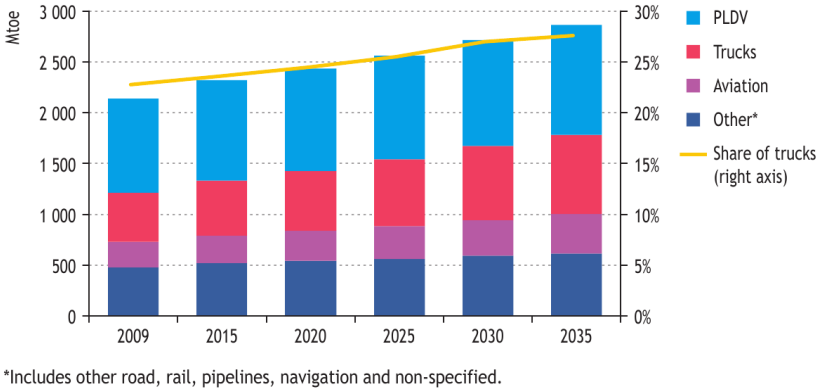



Figura 1.2: Consumo de petróleo por tipo de transporte, 2009-2035 [9].

Para o sector de transportes foram listadas quatro recomendações focadas no modo rodoviário para criação de políticas de:

- melhoria da eficiência energética de pneus;
- padrões de economia de combustível para veículos ligeiros;[1]
- padrões de economia de combustível para veículos pesados;
- condução ecologicamente responsável.

Esta Agência publicou em 2010 um relatório de acompanhamento da adopção destas medidas, onde indica que, por via da adopção de padrões de economia de combustível um Estado membro, obteve um aumento de eficiência de 5,4% de 2002 a 2009 [8, pág. 34] e afirma que é possível obter ganhos permanentes de eficiência entre 5% e 10% para todos os condutores [8, pág. 37].

Surgem assim naturalmente propostas de sistemas de acompanhamento da condução de veículos, sobretudo em operadores de frotas, para suprir a necessidade de motivação cíclica dos condutores e sua adesão às boas práticas de condução ecologicamente responsável, por forma a aproximar os ganhos permanentes de eficiência ao limite superior do intervalo identificado.

[1] PLDV - Passenger Light-Duty Vehicles

Surgiu a possibilidade de colaboração com a empresa TECMIC, S.A. que está presente no mercado de sistemas de gestão profissional de frotas, reaproveitando dados recolhidos pelo sistema XTraN [10] dos barramentos CAN [6][2] durante a operação dos veículos de um operador de transportes públicos rodoviários urbanos, para a análise da eficiência energética da frota gerida. Por estas razões aos dados assim recolhidos chamaremos de operacionais.

## 1.3 Objectivos

Da proposta de projecto de trabalho e discussão subsequente com os interlocutores foi estabelecido como objectivo analisar a eficiência energética por:

- Veículo;
- Condutor;
- Rota;
- Data;
- Hora do dia; e
- Meteorologia.

Para se facilitar a análise por dados de Meteorologia, escolheu-se recolher do sítio Weather Underground [7], um conjunto de variáveis tipicamente disponibilizadas nos relatórios METAR de previsão meteorológica, considerando os dados relativos a um ponto como aproximação suficiente das condições em que decorreu a operação dos veículos. A este conjunto de variáveis chamaremos de meteorológicas. Como é evidente estes dados são externos ao sistema de recolha de dados embarcado nos veículos da frota a estudar, existindo muitos outros cujo a consideração para análise seria interessante, por exemplo a intensidade de tráfego, a existência

[2] CAN bus

de engarrafamentos de trânsito por rota, que não foram disponibilizados nem requeridos como objectivos.

Pretende-se extrair informação dos dados disponibilizados de forma a possibilitar a análise da eficiência energética segundo os critérios propostos. Para tal será criado um protótipo de demonstração recorrendo ao paradigma de armazéns de dados[3].

Pretende-se também extrair conhecimento dos dados armazenados e consolidados, recorrendo a ferramentas e técnicas de extracção de conhecimento por prospecção de dados[4], para descrever os factores que mais influenciam a eficiência de operação de veículos, os grupos de eficiência que se encontram e estimar qual a melhoria que a alteração de determinados comportamentos poderá proporcionar. Neste âmbito utilizaremos o consumo médio de combustível como medida de eficiência de condução.

Elaborou-se um protótipo de solução construindo um armazém de dados para facilitar a análise pretendida, as estruturas e modelos de prospecção de dados capazes de satisfazer os objectivos supra-mencionados.

## 1.4 Abordagem ao problema

Na sequência dos conhecimentos adquiridos pela conclusão de diversas unidades curriculares, optou-se por uma abordagem OLAP [5] para a integração dos dados operacionais com os dados históricos de meteorologia, para posterior extracção de vista sobre o cubo multidimensional a ser submetida a técnicas de prospecção de dados[6] para a descoberta de padrões presentes nos dados disponíveis e subsequente extracção de conhecimento.

[3] Data Warehousing, ou DW
[4] Extracção de Conhecimento ou ECD; Data Mining ou DM
[5] on-line analytical processing
[6] Data Mining

## 1.5 Selecção de metodologia e plataforma

Segundo a abordagem de Fayyad, et al [2] o processo de descoberta de conhecimento em bases de dados divide-se em cinco etapas, conforme se mostra na Figura 1.3

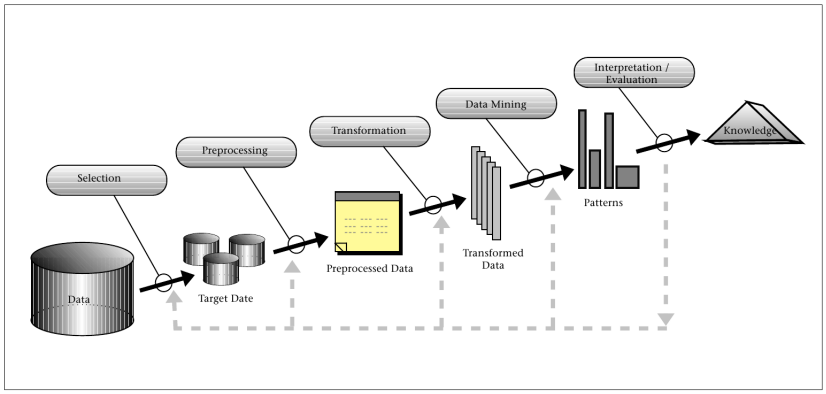



Figura 1.3: Passos constituintes do processo de descoberta de conhecimento [2].

Azevedo e Santos [3] compararam aquelas que são percepcionadas como os processos ou metodologias mais aplicadas no desenvolvimento de projectos de descoberta de conhecimento em bases de dados, seleccionando as metodologias CRISP-DM e SEMMA para comparação entre si e com as cinco etapas advogadas por Fayyad [1]. Conclui-se que existe equivalência entre os passos CRISP-DM e SEMMA como se mostra na tabela 1 da referência [3, pág. 4] citada:

| KDD | SEMMA | CRISP-DM |
|---|---|---|
| Pre KDD | ---------- | Business understanding |
| Selection | Sample | Data Understanding |
| Pre processing | Explore | |
| Transformation | Modify | Data preparation |
| Data mining | Model | Modeling |
| Interpretation/Evaluation | Assessment | Evaluation |
| Post KDD | ---------- | Deployment |

Figura 1.4: Comparação de metodologias KDD CRISP-DM e SEMMA [3].

Dado que diversos autores apresentam a metodologia CRISP-DM como sendo a mais comum em projectos de descoberta de conhecimento em bases de dados, sendo também normalmente indicada como mais prescritiva e função da relativa inexperiência do autor, optou-se por esta metodologia para o desenvolvimento deste projecto.

A metodologia SEMMA aparenta ser uma aproximação equivalente, mais ligeira, que um autor experiente em projectos de descoberta de conhecimento em bases de dados poderá eleger trocando a abordagem mais guiada do CRISP-DM por um grau de liberdade superior.

No âmbito deste relatório designaremos as fases CRISP-DM:

- Familiarização - como tradução de *Business understanding*;
- Interpretação de dados - como tradução de *Data understanding*;
- Modelação - como tradução de *Modeling*, e;
- Avaliação - como tradução de *Evaluation*.

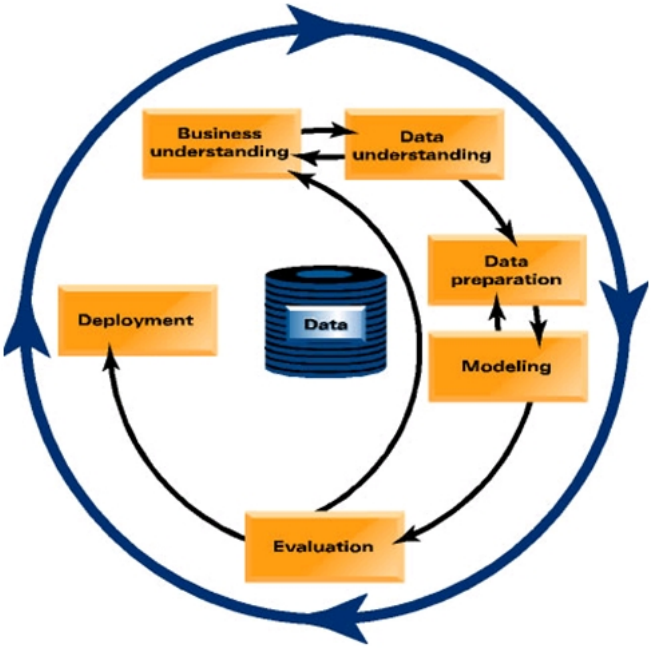



Figura 1.5: Visão da metodologia CRISP-DM.

Estando escolhida a metodologia para o desenvolvimento do protótipo deste projecto, optou-se pelo pragmatismo de recorrer à solução de arquitectura proporcionada pela plataforma de *Bussiness Intelligence* da Microsoft sobre SQL Server 2008R2 [12] com SSIS[7], a explorar em Visual Studio 2008, instalada numa máquina virtual[8] correndo Windows 7 Professional.

A razão da escolha pela abordagem OLAP e plataforma seleccionada apoia-se nas vantagens que os:

- processos de Extracção, Limpeza e Carregamento de dados, iniciais na construção de um armazém de dados, apresentam para as ferramentas de prospecção de dados como garantia de disponibilidade e de correcção de dados que frequentemente os requerem assim preparados [11, pág. 1];

- utilizadores obtêm das facilidades de *slice & dice* típicas de um armazém de dados OLAP na familiarização com os dados, acessória ao processo iterativo e interactivo de descoberta de conhecimento durante a prospecção de dados;

- mecanismos e ferramentas integradas de uma plataforma OLAM[9] permitem na prospecção de dados, como é o caso da plataforma de BI escolhida em que a prospecção pode ser feita com recurso ao Microsoft Excel, ferramenta muito popular entre utilizadores empresariais, facilitando assim a utilização por peritos da área, mas leigos na utilização deste tipo de técnicas, para assim explorar os cubos multi-dimensionais e melhorar os modelos de prospecção construídos no âmbito do protótipo elaborado, e;

- o SQL Server é o SGBD em uso pela empresa que facilitou os dados operacionais.

[7] SQL Server Integration Services
[8] VirtualBox da ORACLE
[9] On-line Analytical Mining

Não obstante são conhecidas outras plataformas de exploração de dados para extracção de conhecimento como por exemplo o R, SAS, SPSS, WEKA ou Rapid Miner para nomear alguns dos mais vulgares segundo um dos mais abrangentes inquéritos [41] sobre esta temática. Acresce que uma vez produzida a vista de extracção de dados sobre o cubo multidimensional do armazém de dados, é sempre possível submeter os mesmos a qualquer uma das ferramentas que se deseje utilizar, bastando replicar a definição dos modelos elaborados no âmbito do presente trabalho.

A escolha de desenvolver o protótipo numa máquina virtual fundamenta-se nas vantagens de segregação, portabilidade e controlo proporcionadas pela virtualização e pelo pragmatismo de reutilizar um ambiente já familiar e com o qual já se tinha desenvolvido actividade académica neste campo, aquando da frequência da unidade curricular de Sistemas de Informação para Apoio à Decisão.

Igualmente na senda dos conhecimentos anteriormente adquiridos, optou-se por armazém de dados seguindo um esquema em estrela, e abordagem preconizada por Kimball [13] em detrimento do esquema em floco de neve e da abordagem proposta por Inmon [15]. A razão fundamental é que o esquema em estrela adoptado facilita o cálculo de pré-agregados o que por sua vez proporciona facilidade de exploração e melhora o desempenho nas operações de interrogação do armazém de dados.

## 1.6 Organização do relatório

Este documento está organizado em quatro capítulos cujo conteúdo é:

- Capítulo 1: Introdução - Apresentando-se as motivações, a familiarização e contexto, elencam-se objectivos do trabalho realizado, detalhando a abordagem ao problema, a escolha de metodologia e plataforma, terminando com a organização do documento;

- Capítulo 2: Interpretação de dados - Enquadrando-se este projecto nos trabalhos prévios, descrevendo-se o trabalho desenvolvido na caracterização e elaboração de um protótipo de armazém de dados e apresentando exemplos de análises tipo OLAP que este possibilita;

- Capítulo 3: Modelação e resultados - Apresentando as estruturas, os modelos de dados, os algoritmos de prospecção utilizados e principais regras de conhecimento extraídas,e;

- Capítulo 4: Conclusões - Confronta-se o trabalho realizado com os objectivos enumerados, elencando-se as ideias a reter e perspectivando aspectos de desenvolvimento futuro.

Note-se que a ausência do capítulo ”tradicional” descrevendo o estado da arte é compensada pela colocação de parte dessa informação nas subsecções de apresentação de conceitos dos capítulos enumerados, que enquadram o trabalho desenvolvido nos princípios teóricos apresentados pelos respectivos autores, ”Gigantes aos ombros dos quais” nos apoiamos.

Note-se também que a utilização de uma plataforma que disponibiliza diversos algoritmos como ”ferramentas” e que cuja a descrição se encontra disponível em diversas fontes de informação entre as quais se destaca os textos de referência [14], [25] [21], [22], [24] e os diversos artigos enumerados na bibliografia.

Na pesquisa e consulta da bibliografia e elaboração deste relatório empregou-se cerca de 30 % do esforço despendido e seria difícil elaborar uma descrição do estado da arte que acrescentasse valor ao patente nas obras de referência citadas no anterior parágrafo.

# Capítulo 2

# Interpretação de dados

Neste capítulo apresenta-se o enquadramento no trabalho prévio, proporcionamos a familiarização com o negócio, caracterizam-se os dados de base, detalha-se a implementação do armazém de dados em que se empregou cerca de 40 % do esforço e com que se conclui este capítulo.

## 2.1 Trabalho prévio

A detecção de padrões é uma área de pesquisa importante nos campos de prospecção de dados e de descoberta de conhecimento em bases de dados, pois tal como declarado por [24, pág. 3], ”estamos soterrados em dados”, neste caso pelo enorme volume gerado pelo sistema de recolha de dados operacionais gerados pelo CAN bus da frota de veículos a analisar e registados pelo sistema de gestão de frotas XTraN combinados com os dados históricos de observação meteorológica. Segundo [33] ”However, the task of learning standard behaviours from raw data of real human drivers has not yet been tackled and will be an area of future research”.

O estilo de condução, visto como a ”atitude, orientação e modo de pensar durante a condução no dia-a-dia” é ”habitualmente baseada em questionários” [34, 35]. Trabalhos mais recentes utilizam simuladores de condução virtual para recolher dados realísticos de condutores humanos para modelar o seu comportamento [36], ou para classificar o estilo de

condução usando um método objectivo para ordenação de condutores [37].

No contexto da eficiência energética, muito foi já alcançado no que ao nível de desempenho de motores e veículos, obtendo-se melhorias consideráveis e poupança de energia. A qualidade de condução conducente à eficiente utilização de combustível e os métodos de promover continuamente a sua melhoria, tem sido foco de atenção restrita, em parte pela dificuldade de avaliar o desempenho de condutores.

Os condutores controlam a velocidade, aceleração, travagem e posição do veículo na estrada, num ambiente caracterizado por condições de tráfego, itinerário, carga, condições atmosféricas, entre outros parâmetros. Ao controlar o veículo, o condutor actua directamente não apenas na velocidade e posicionamento do veículo, mas também em parâmetros como a aceleração, travagem, regime de rotação do motor e velocidade engrenada [38, 40]. A maneira como o condutor actua e controla este parâmetros em relação ao ambiente determina o seu estilo de condução. Estilos de condução diferentes resultam em diferentes consumos instantâneos e médios de combustível e de forma mais genérica determinam a qualidade da condução. Contudo o ambiente também exerce influência no consumo de combustível interagindo e condicionando as decisões dos condutores.

## 2.2 Caracterização dos dados de base

Os dados operacionais foram disponibilizados na forma de um ficheiro de *backup* de uma base de dados SQL Server, contendo cinco tabelas extraídas de uma base de dados do sistema XTraNPassenger [10] da TECMIC, conforme se apresenta na Figura 2.1.

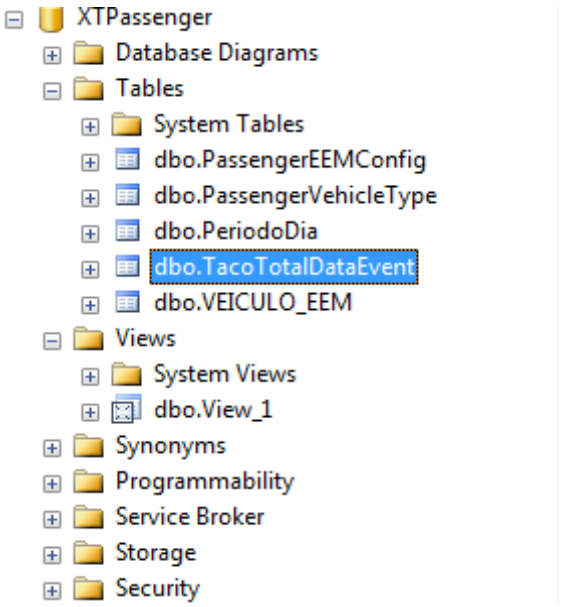



Figura 2.1: Extracto de ficheiro inicial fonte de dados.

Note-se que se houver a oportunidade de aceder directamente à base de dados, haverá a preocupação de minimizar o acoplamento e impacto do processo de extracção, transformação e carregamento de dados no armazém de dados, criando uma outra base de dados intitulada "`DataStaging_TECMIC`" que nada mais será que uma réplica local, actualizada incrementalmente, de todas as tabelas fonte do modelo multidimensional estabelecido.

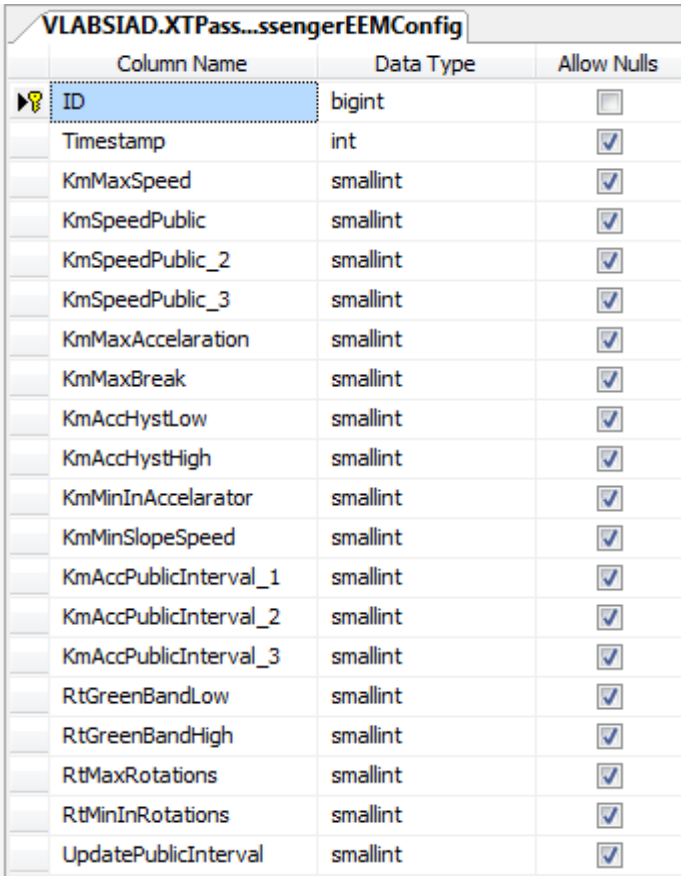
### 2.2.1 Análise preliminar

Avaliou-se então os dados de cada tabela de SQL conforme se descreve de seguida.

#### 2.2.1.1 Tabela PassengerEEMConfig

Trata-se uma tabela de parametrização/configuração do equipamento embarcado nos veículos da frota a estudar e que segundo os especialistas

de negócio *não deverá ser tomada em conta para o trabalho em curso*. Apresenta-se portanto apenas a lista de colunas na Figura 2.2.



VLABSIAD.XTPass...ssengerEEMConfig

| Column Name | Data Type | Allow Nulls |
|---|---|---|
| ID | bigint | ☐ |
| Timestamp | int | ☑ |
| KmMaxSpeed | smallint | ☑ |
| KmSpeedPublic | smallint | ☑ |
| KmSpeedPublic_2 | smallint | ☑ |
| KmSpeedPublic_3 | smallint | ☑ |
| KmMaxAccelaration | smallint | ☑ |
| KmMaxBreak | smallint | ☑ |
| KmAccHystLow | smallint | ☑ |
| KmAccHystHigh | smallint | ☑ |
| KmMinInAccelarator | smallint | ☑ |
| KmMinSlopeSpeed | smallint | ☑ |
| KmAccPublicInterval_1 | smallint | ☑ |
| KmAccPublicInterval_2 | smallint | ☑ |
| KmAccPublicInterval_3 | smallint | ☑ |
| RtGreenBandLow | smallint | ☑ |
| RtGreenBandHigh | smallint | ☑ |
| RtMaxRotations | smallint | ☑ |
| RtMinInRotations | smallint | ☑ |
| UpdatePublicInterval | smallint | ☑ |

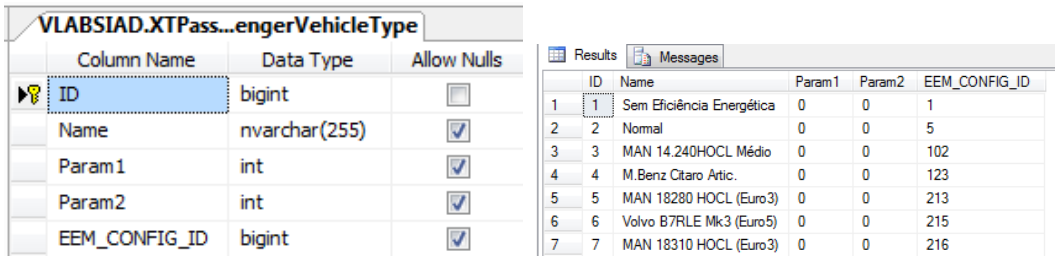
Figura 2.2: Tabela PassengerEEMConfig.

#### 2.2.1.2 Tabela PassengerVehicleType

Trata-se de uma tabela de descrição dos veículos nos quais se procedeu à recolha de dados, apresentando apenas sete registos e cinco colunas como se mostra na Figura 2.3.

#### 2.2.1.3 Tabela PeriodoDia

Trata-se de uma tabela com a descrição dos períodos do dia de acordo com as regras do negócio conforme se mostra na Figura 2.4



VLABSIAD.XTPass...engerVehicleType

| Column Name | Data Type | Allow Nulls |
|---|---|---|
| ID | bigint | ☐ |
| Name | nvarchar(255) | ☑ |
| Param1 | int | ☑ |
| Param2 | int | ☑ |
| EEM_CONFIG_ID | bigint | ☑ |

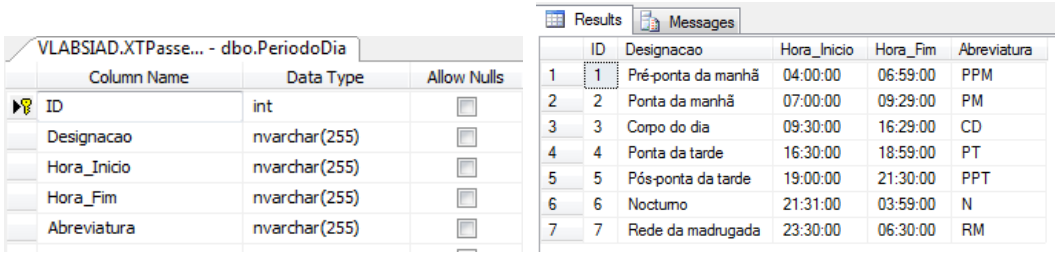
(a) Colunas

Results | Messages

| | ID | Name | Param1 | Param2 | EEM_CONFIG_ID |
|---|---|---|---|---|---|
| 1 | 1 | Sem Eficiência Energética | 0 | 0 | 1 |
| 2 | 2 | Normal | 0 | 0 | 5 |
| 3 | 3 | MAN 14.240HOCL Médio | 0 | 0 | 102 |
| 4 | 4 | M.Benz Citaro Artic. | 0 | 0 | 123 |
| 5 | 5 | MAN 18280 HOCL (Euro3) | 0 | 0 | 213 |
| 6 | 6 | Volvo B7RLE Mk3 (Euro5) | 0 | 0 | 215 |
| 7 | 7 | MAN 18310 HOCL (Euro3) | 0 | 0 | 216 |

(b) Registos

Figura 2.3: Tabela PassengerVehicleType.



VLABSIAD.XTPasse... - dbo.PeriodoDia

| Column Name | Data Type | Allow Nulls |
|---|---|---|
| ID | int | ☐ |
| Designacao | nvarchar(255) | ☐ |
| Hora_Inicio | nvarchar(255) | ☐ |
| Hora_Fim | nvarchar(255) | ☐ |
| Abreviatura | nvarchar(255) | ☐ |

(a) Colunas

Results | Messages

| | ID | Designacao | Hora_Inicio | Hora_Fim | Abreviatura |
|---|---|---|---|---|---|
| 1 | 1 | Pré-ponta da manhã | 04:00:00 | 06:59:00 | PPM |
| 2 | 2 | Ponta da manhã | 07:00:00 | 09:29:00 | PM |
| 3 | 3 | Corpo do dia | 09:30:00 | 16:29:00 | CD |
| 4 | 4 | Ponta da tarde | 16:30:00 | 18:59:00 | PT |
| 5 | 5 | Pós-ponta da tarde | 19:00:00 | 21:30:00 | PPT |
| 6 | 6 | Nocturno | 21:31:00 | 03:59:00 | N |
| 7 | 7 | Rede da madrugada | 23:30:00 | 06:30:00 | RM |

(b) Registos

Figura 2.4: Tabela PeriodoDia.

#### 2.2.1.4 Tabela VEICULO_EEM

É a tabela que descreve o tipo e número dos veículos fonte dos dados, contem 869 registos (muitos incompletos, por exemplo NULL, Testes TECMIC, etc...) e quatro colunas como se mostra na Figura 2.5.

#### 2.2.1.5 Tabela TacoTotalDataEvent

Trata-se finalmente da tabela com os registos recolhidos pelo sistema embarcado a bordo dos veículos da frota a analisar, contendo 1 698 295 registos e 44 colunas conforme se mostra na Figura 2.6.

VLABSIAD.XTPass...dbo.VEICULO_EEM

| Column Name | Data Type | Allow Nulls |
|---|---|---|
| ID | bigint | ☐ |
| VEHICLETYPE_ID | bigint | ☑ |
| NR_VEICULO | int | ☑ |
| DADOS | nvarchar(255) | ☑ |

(a) Colunas

Results | Messages

| | ID | VEHICLETYPE_ID | NR_VEICULO | DADOS |
|---|---|---|---|---|
| 1 | 7 | 1 | 7 | Testes TECMIC |
| 2 | 8 | 2 | 8 | Testes TECMIC |
| 3 | 9 | 1 | 9 | Caixa de Testes TECMIC |
| 4 | 10 | 1 | 10 | Testes TECMIC |
| 5 | 11 | 1 | 11 | TECMIC - testes de bancada |
| 6 | 99 | 1 | 99 | Consola de testes |
| 7 | 100 | 1 | 100 | BUS |
| 8 | 101 | 1 | 101 | BUS |
| 9 | 102 | 1 | 102 | BUS |
| 10 | 103 | 1 | 103 | BUS |
| 11 | 104 | 1 | 104 | BUS |
| 12 | 105 | 1 | 105 | BUS |
| 13 | 106 | 1 | 106 | BUS |
| 14 | 107 | 1 | 107 | ELECTRICO HISTORICO |
| 15 | 108 | 1 | 108 | BUS |
| 16 | 109 | 1 | 109 | BUS |
| 17 | 110 | 1 | 110 | BUS |

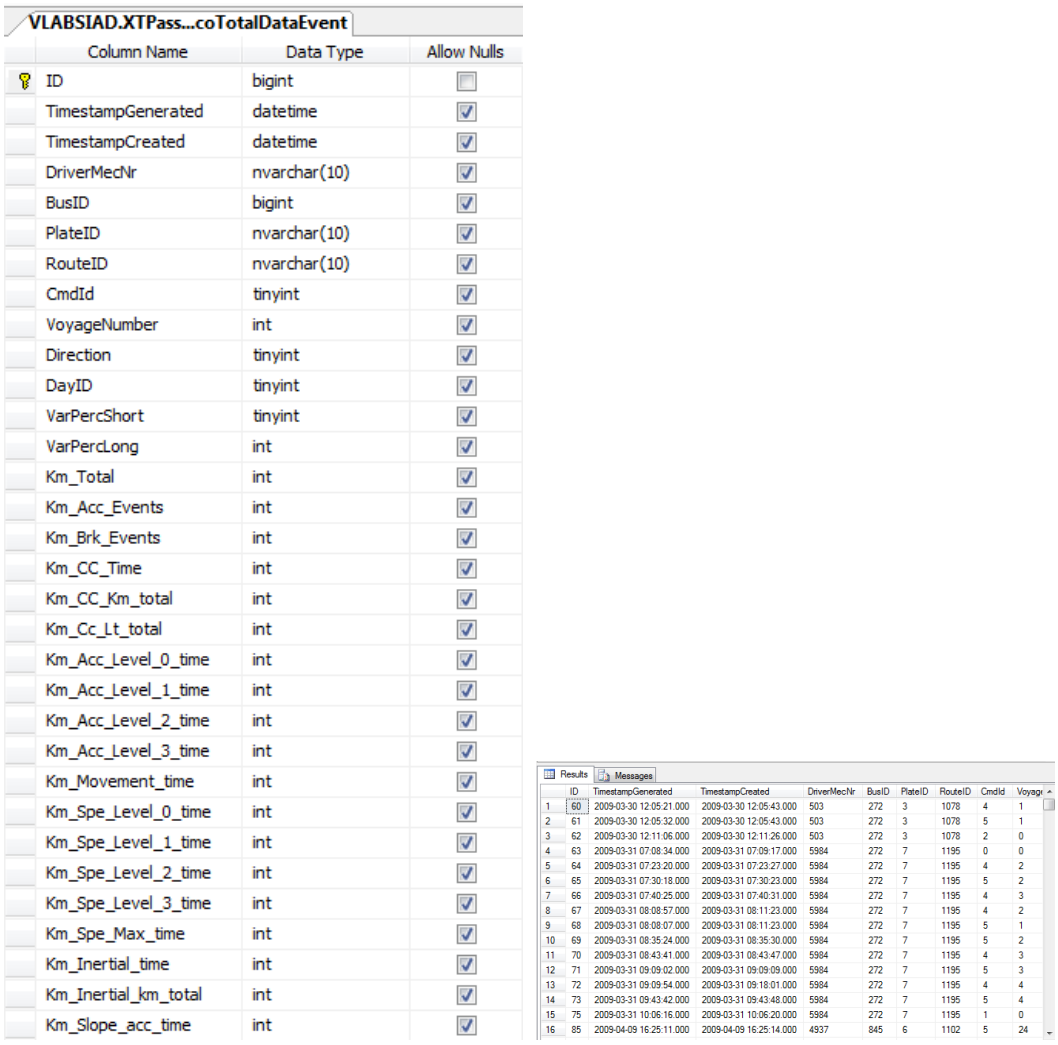
(b) Extracto de registos

Figura 2.5: Tabela VEICULO_EEM.

### 2.2.2 Descrição de dados

O passo seguinte foi a submissão desta tabela a uma tarefa SSIS de análise de perfil de dados[1], conforme se mostra na Figura 2.7 e no apêndice D.

[1] Data Profiling Task



VLABSIAD.XTPass...coTotalDataEvent

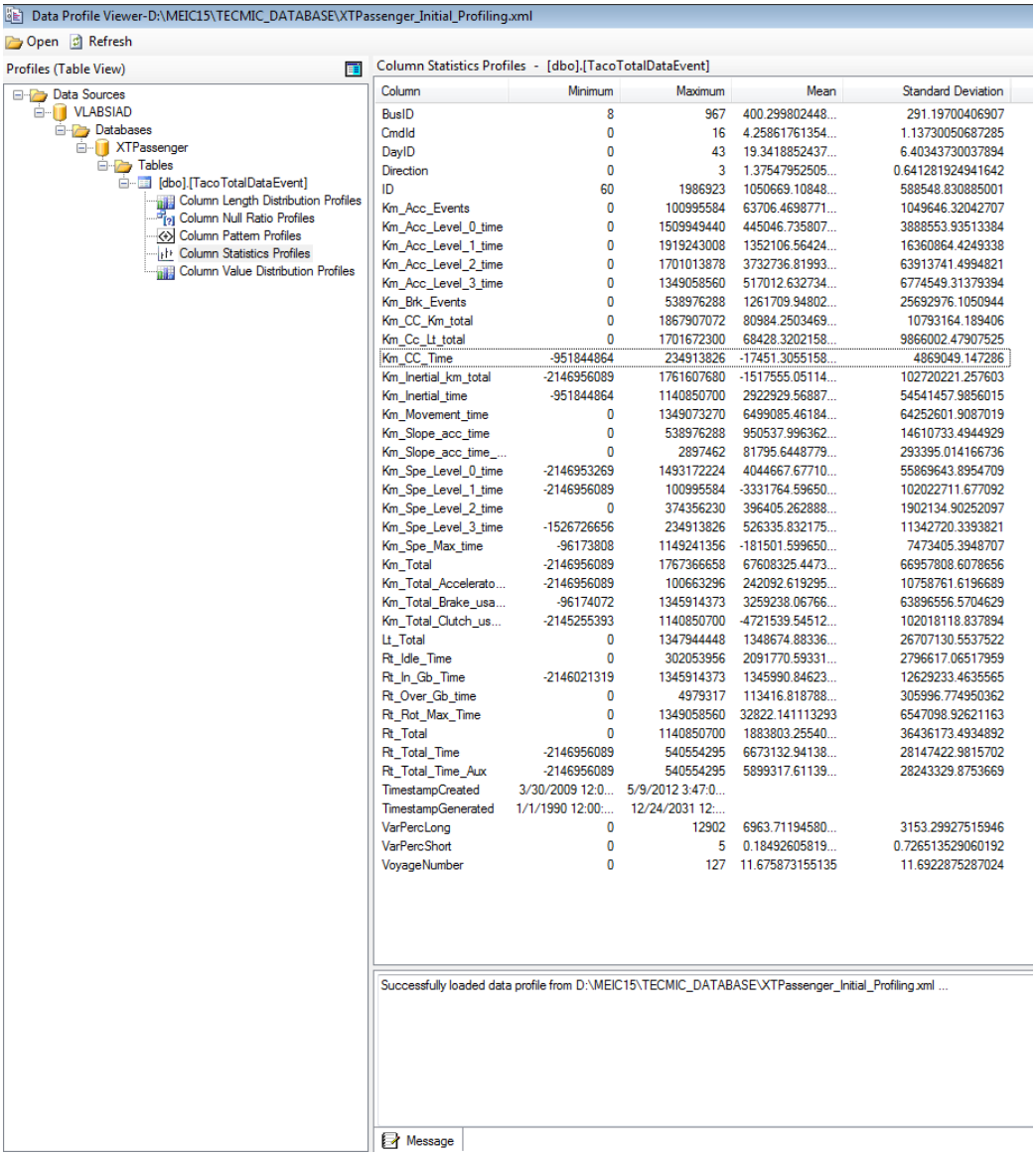
| Column Name | Data Type | Allow Nulls |
|---|---|---|
| ID | bigint | ☐ |
| TimestampGenerated | datetime | ☑ |
| TimestampCreated | datetime | ☑ |
| DriverMecNr | nvarchar(10) | ☑ |
| BusID | bigint | ☑ |
| PlateID | nvarchar(10) | ☑ |
| RouteID | nvarchar(10) | ☑ |
| CmdId | tinyint | ☑ |
| VoyageNumber | int | ☑ |
| Direction | tinyint | ☑ |
| DayID | tinyint | ☑ |
| VarPercShort | tinyint | ☑ |
| VarPercLong | int | ☑ |
| Km_Total | int | ☑ |
| Km_Acc_Events | int | ☑ |
| Km_Brk_Events | int | ☑ |
| Km_CC_Time | int | ☑ |
| Km_CC_Km_total | int | ☑ |
| Km_Cc_Lt_total | int | ☑ |
| Km_Acc_Level_0_time | int | ☑ |
| Km_Acc_Level_1_time | int | ☑ |
| Km_Acc_Level_2_time | int | ☑ |
| Km_Acc_Level_3_time | int | ☑ |
| Km_Movement_time | int | ☑ |
| Km_Spe_Level_0_time | int | ☑ |
| Km_Spe_Level_1_time | int | ☑ |
| Km_Spe_Level_2_time | int | ☑ |
| Km_Spe_Level_3_time | int | ☑ |
| Km_Spe_Max_time | int | ☑ |
| Km_Inertial_time | int | ☑ |
| Km_Inertial_km_total | int | ☑ |
| Km_Slope_acc_time | int | ☑ |

Results | Messages

| | ID | TimestampGenerated | TimestampCreated | DriverMecNr | BusID | PlateID | RouteID | CmdId | Voyage |
|---|---|---|---|---|---|---|---|---|---|
| 1 | 60 | 2009-03-30 12:05:21.000 | 2009-03-30 12:05:43.000 | 503 | 272 | 3 | 1078 | 4 | 1 |
| 2 | 61 | 2009-03-30 12:05:32.000 | 2009-03-30 12:05:43.000 | 503 | 272 | 3 | 1078 | 5 | 1 |
| 3 | 62 | 2009-03-30 12:11:06.000 | 2009-03-30 12:11:26.000 | 503 | 272 | 3 | 1078 | 2 | 0 |
| 4 | 63 | 2009-03-31 07:08:34.000 | 2009-03-31 07:09:17.000 | 5984 | 272 | 7 | 1195 | 0 | 0 |
| 5 | 64 | 2009-03-31 07:23:20.000 | 2009-03-31 07:23:27.000 | 5984 | 272 | 7 | 1195 | 4 | 2 |
| 6 | 65 | 2009-03-31 07:30:18.000 | 2009-03-31 07:30:23.000 | 5984 | 272 | 7 | 1195 | 5 | 2 |
| 7 | 66 | 2009-03-31 07:40:25.000 | 2009-03-31 07:40:31.000 | 5984 | 272 | 7 | 1195 | 4 | 3 |
| 8 | 67 | 2009-03-31 08:08:57.000 | 2009-03-31 08:11:23.000 | 5984 | 272 | 7 | 1195 | 4 | 2 |
| 9 | 68 | 2009-03-31 08:08:07.000 | 2009-03-31 08:11:23.000 | 5984 | 272 | 7 | 1195 | 5 | 1 |
| 10 | 69 | 2009-03-31 08:35:24.000 | 2009-03-31 08:35:30.000 | 5984 | 272 | 7 | 1195 | 5 | 2 |
| 11 | 70 | 2009-03-31 08:43:41.000 | 2009-03-31 08:43:47.000 | 5984 | 272 | 7 | 1195 | 4 | 3 |
| 12 | 71 | 2009-03-31 09:09:02.000 | 2009-03-31 09:09:09.000 | 5984 | 272 | 7 | 1195 | 5 | 3 |
| 13 | 72 | 2009-03-31 09:09:54.000 | 2009-03-31 09:18:01.000 | 5984 | 272 | 7 | 1195 | 4 | 4 |
| 14 | 73 | 2009-03-31 09:43:42.000 | 2009-03-31 09:43:48.000 | 5984 | 272 | 7 | 1195 | 5 | 4 |
| 15 | 75 | 2009-03-31 10:06:16.000 | 2009-03-31 10:06:20.000 | 5984 | 272 | 7 | 1195 | 1 | 0 |
| 16 | 85 | 2009-04-09 16:25:11.000 | 2009-04-09 16:25:14.000 | 4937 | 845 | 6 | 1102 | 5 | 24 |

(a) Colunas (b) Extracto de registos

Figura 2.6: Tabela TacoTotalDataEvent.




| Column | Minimum | Maximum | Mean | Standard Deviation |
|---|---|---|---|---|
| BusID | 8 | 967 | 400.299802448... | 291.19700406907 |
| CmdId | 0 | 16 | 4.25861761354... | 1.13730050687285 |
| DayID | 0 | 43 | 19.3418852437... | 6.40343730037894 |
| Direction | 0 | 3 | 1.37547952505... | 0.641281924941642 |
| ID | 60 | 1986923 | 1050669.10848... | 588548.830885001 |
| Km_Acc_Events | 0 | 100995584 | 63706.4698771... | 1049646.32042707 |
| Km_Acc_Level_0_time | 0 | 1509949440 | 445046.735807... | 3888553.93513384 |
| Km_Acc_Level_1_time | 0 | 1919243008 | 1352106.56424... | 16360864.4249338 |
| Km_Acc_Level_2_time | 0 | 1701013878 | 3732736.81993... | 63913741.4994821 |
| Km_Acc_Level_3_time | 0 | 1349058560 | 517012.632734... | 6774549.31379394 |
| Km_Brk_Events | 0 | 538976288 | 1261709.94802... | 25692976.1050944 |
| Km_CC_Km_total | 0 | 1867907072 | 80984.2503469... | 10793164.189406 |
| Km_Cc_Lt_total | 0 | 1701672300 | 68428.3202158... | 9866002.47907525 |
| Km_CC_Time | -951844864 | 234913826 | -17451.3055158... | 4869049.147286 |
| Km_Inertial_km_total | -2146956089 | 1761607680 | -1517555.05114... | 102720221.257603 |
| Km_Inertial_time | -951844864 | 1140850700 | 2922929.56887... | 54541457.9856015 |
| Km_Movement_time | 0 | 1349073270 | 6499085.46184... | 64252601.9087019 |
| Km_Slope_acc_time | 0 | 538976288 | 950537.996362... | 14610733.4944929 |
| Km_Slope_acc_time_... | 0 | 2897462 | 81795.6448779... | 293395.014166736 |
| Km_Spe_Level_0_time | -2146953269 | 1493172224 | 4044667.67710... | 55869643.8954709 |
| Km_Spe_Level_1_time | -2146956089 | 100995584 | -3331764.59650... | 102022711.677092 |
| Km_Spe_Level_2_time | 0 | 374356230 | 396405.262888... | 1902134.90252097 |
| Km_Spe_Level_3_time | -1526726656 | 234913826 | 526335.832175... | 11342720.3393821 |
| Km_Spe_Max_time | -96173808 | 1149241356 | -181501.599650... | 7473405.3948707 |
| Km_Total | -2146956089 | 1767366658 | 67608325.4473... | 66957808.6078656 |
| Km_Total_Accelerato... | -2146956089 | 100663296 | 242092.619295... | 10758761.6196689 |
| Km_Total_Brake_usa... | -96174072 | 1345914373 | 3259238.06766... | 63896556.5704629 |
| Km_Total_Clutch_us... | -2145255393 | 1140850700 | -4721539.54512... | 102018118.837894 |
| Lt_Total | 0 | 1347944448 | 1348674.88336... | 26707130.5537522 |
| Rt_Idle_Time | 0 | 302053956 | 2091770.59331... | 2796617.06517959 |
| Rt_In_Gb_Time | -2146021319 | 1345914373 | 1345990.84623... | 12629233.4635565 |
| Rt_Over_Gb_time | 0 | 4979317 | 113416.818788... | 305996.774950362 |
| Rt_Rot_Max_Time | 0 | 1349058560 | 32822.141113293 | 6547098.92621163 |
| Rt_Total | 0 | 1140850700 | 1883803.25540... | 36436173.4934892 |
| Rt_Total_Time | -2146956089 | 540554295 | 6673132.94138... | 28147422.9815702 |
| Rt_Total_Time_Aux | -2146956089 | 540554295 | 5899317.61139... | 28243329.8753669 |
| TimestampCreated | 3/30/2009 12:0... | 5/9/2012 3:47:0... | | |
| TimestampGenerated | 1/1/1990 12:00:... | 12/24/2031 12:... | | |
| VarPercLong | 0 | 12902 | 6963.71194580... | 3153.29927515946 |
| VarPercShort | 0 | 5 | 0.18492605819... | 0.726513529060192 |
| VoyageNumber | 0 | 127 | 11.675873155135 | 11.6922875287024 |

Figura 2.7: Resultado da tarefa de análise ao perfil dos dados.

Desta análise e diálogo com os peritos de negócio analisou-se a semântica de cada coluna e o respectivo método de registo. Apresentamos a síntese das respostas desses interlocutores (transcritas caso a caso após travessão):

1. estão isentas de dados omissos, i.e. valores "NULL";

2. "TimestampGenerated" e "TimestampCreated" - Generated é relativo ao espoletar do evento na caixa negra e Created relativo ao momento em que é registado na BD;

3. "PlateID" - é o número da viagem que o autocarro faz, sempre no âmbito de uma carreira, tipicamente um autocarro que se encontra em serviço está a executar "a carreira 27 Chapa 3";

4. "RouteID" - é a identificação da carreira;

5. "CmdId" - é a identificação do tipo de evento que origina o registo:

   "START_SERVICE" = 0x00;

   "STOP_SERVICE_WITH_DEPOT" = 0x01;

   "STOP_SERVICE_WITHOUT_DEPOT" = 0x02;

   "DRIVER_SWAP_OUT" = 0x03;

   "DRIVER_SWAP_IN" = 0x06;

   "START_VOYAGE" = 0x04;

   "STOP_VOYAGE" = 0x05;

   "TIMED" = 0x09;

   IGNITION_ON" = 0x10;

   IGNITION_OFF" = 0x11.

   Acresce que uma viagem é sempre feita no âmbito de um serviço (quando inicia um serviço, inicia uma carreira, quando inicia uma viagem inicia uma chapa dessa carreira);

6. "VoyageNumber" - Trata-se de um nº identificador da viagem;

7. "Direction", apenas toma os valores 0, 1, 2 e 3. Qual o significado de cada um (0=ida, 1= volta, 2=circular, 3=desconhecido) - Certo;

8. "DayID", qual a relação com as colunas TimestampGenerated e TimestampCreated - DayID é uma chave para uma tabela em que

se detalha o tipo de horário e percurso a cumprir no dia de TimestampGenerated - não é para ser considerado;

9. "VarPercShort", com 6 valores diferentes e "VarPercLong" com 589 valores diferentes. Qual o significado, - este tb não é para ser considerado (trata-se de variantes de percurso);

10. "Km_Total", total de Km percorridos até ao evento ser gerado - totalizador (atenção trata-se de um totalizador – valor acumulado – para se perceber entre eventos, deve-se fazer um evento menos o anterior) de 0,5 kms da viatura na altura do evento;

11. "Km_Acc_Events", será Km percorridos com o actuação no acelerador - totalizador de acelerações bruscas da viatura, na altura do evento;

12. "Km_Brk_Events", idem para actuação de travão - totalizador de travagens bruscas na altura do evento;

13. "Km_CC_Time", Km ou tempo (segundos) em que o Cruise Control esteve activo - totalizador (em segundos) da viatura da utilização do CC na altura do evento;

14. " Km_CC_Km_total", Total de Km percorridos com Cruise Control activo - totalizador de kms da viatura em CC;

15. "Km_Cc_Lt_total", litros de combustível consumidos com Cruise Control activo - totalizador de litros consumidos em CC;

16. "Km_Acc_Level_0_time", "Km_Acc_Level_1_time", "Km_Acc_Level_2_time" e "Km_Acc_Level_3_time", Km ou segundos percorridos em cada um de 4 intervalos de ângulos de actuação do acelerador (Level_0 < Level_1 < ... < Level_3) – tempo total com o acelerador acima de nível x;

17. "Km_Movement_time" - Totalizador de Tempo de Condução (velocidade não nula);

18. `"Km_Spe_Level_0_time"`, `"Km_Spe_Level_1_time"`, `"Km_Spe_Level_2_time"`, `"Km_Spe_Level_3_time"` e `"Km_Spe_Max_time"` – Totalizador (em segundos) de Velocidade acima de nível x;

19. `"Km_Inertial_time"`, segundos com marcha em inércia - Totalizador de segundos em marcha de inércia;

20. `"Km_Inertial_km_total"`, Km com marcha em inércia - Totalizador de Kms em marcha de Inércia;

21. `"Km_Slope_acc_time"`, segundos de autuação do acelerador em slope (descidas ou subidas) - Tempo total de segundos com aceleração em declive;

22. `"Km_Total_Brake_usage"`, `"Km_Total_Clutch_usage"` e `"Km_Total_Accelerator_usage"` - Totalizador de actuação de travão, embraiagem e acelerador;

23. `"Lt_Total"` - total de litros de combustível consumido;

24. `"Rt_Total"`, `"Rt_Total_Time"`, `"Rt_Idle_Time"`, `"Rt_In_Gb_Time"`, `"Rt_Over_Gb_time"`, `"Rt_Rot_Max_Time"`, qual o significado de Rt

    `"Rt_Total"` – Totalizador de Rotações do Motor;

    `"Rt_Total_Time"` – Totalizador do Funcionamento do motor (segundos);

    `"Rt_In_Gb_Time"` – Totalizador (segundos) de rotações em Banda Económica;

    `"Rt_Over_Gb_time"` – Totalizador em segundos de rotações acima da Banda Económica;

    `"Rt_Rot_Max_Time"` – Totalizador em segundos em excesso de rotações;

25. `"Rt_Total_Time_Aux" e "Rt_Total_Time"` – o primeiro não é para considerar;

26. `"Km_Slope_acc_time_back"` – tb não é para considerar;

Ainda com a colaboração dos peritos de negócio foi decidido elaborar uma vista sobre esta tabela que permitisse agrupar os diferentes registos relativos a um facto num único registo da vista, como se mostra na listagem B. Como se verifica pela consulta ao código listado, decidiu-se agrupar os registos por `"TimestampGenerated"`, por `"BusID"`, por `"RouteID"`, `"PlateID"`, por `"DriverMecNr"`, por `"Direction"`, por `"VarPercLong"` e por `"VoyageNumber"`, filtrando os dados anteriores ao ano de 2010 porque os peritos de negócio indicaram que estes não deviam ser considerados. Deste modo os 1 698 295 registos e 44 colunas de dados operacionais inicialmente disponibilizados foram consolidados em 397 261 registos e 42 colunas.

## 2.3 Armazém de dados

Desenvolveu-se o *script* SQL para a criação do armazém de dados e perante os que estão disponíveis apostou-se nas dimensões data e hora do dia e em hipotéticas versões de dimensões local e veículo cuja listagem se apresenta no apêndice A. Este armazém de dados segue o desenho em estrela preconizado por Kimball [13] e resulta na organização apresentada na Figura 2.9.

Elaborou-se a solução de extracção, transformação e carregamento do armazém de dados [2] em solução Microsoft Visual Studio com SSIS, numa aproximação *top-down* conforme se apresenta na Figura 2.8.

[2] Extraction, Transformation and Loading, ou ETL

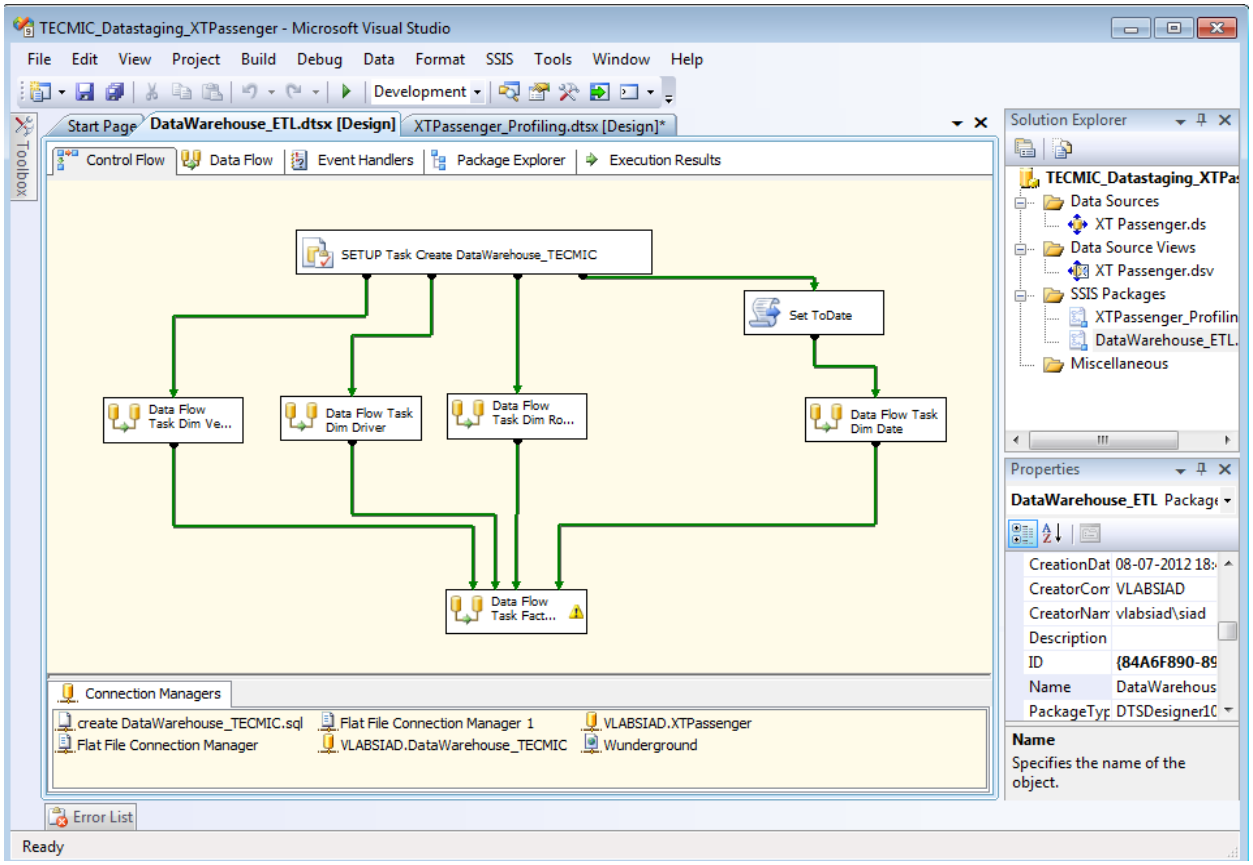



Figura 2.8: Solução Visual Studio SSIS para ETL.

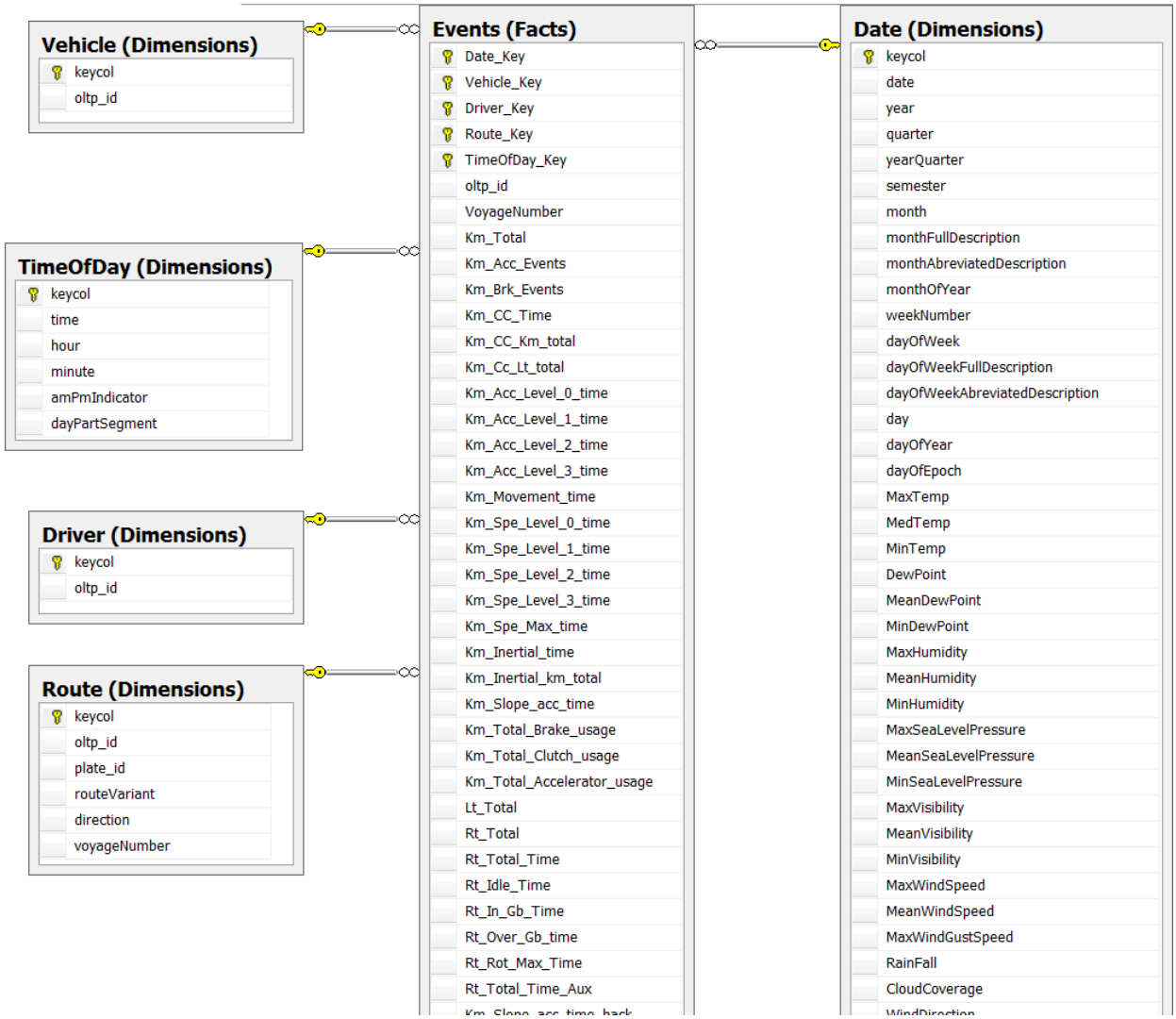



Figura 2.9: Esquema em estrela do armazém de dados.

### 2.3.1 Declaração da granularidade de análise

Para satisfazer os objectivos enunciados em 1.3, a análise será feita por condutor, por veículo, por rota, por data e por hora. Os dados meteorológicos serão usados como característica de um dia, recuperados do registado por uma estação e considerados representativos de toda a área geográfica na qual ocorreu a operação dos veículos, pois não se conhe-

cem detalhes das rotas percorridas nem existem dados de localização geográfica.

### 2.3.2 Escolha das dimensões de análise

A escolha das dimensões de análise decorre da declaração de granularidade e da forma como os peritos de negócio descrevem os dados [13, pág. 31]. Assim foram elaboradas as dimensões Condutor, Veículo, Rota[3], Data e Hora, conforme se apresenta nas Figuras 2.10 a 2.12 e como mais se detalha no apêndice A para as dimensões data e hora.

[3]O fluxo de dados para as Dimensões Rota e Veículo é idêntico ao da dimensão Condutor, pelo que apenas se apresenta este.

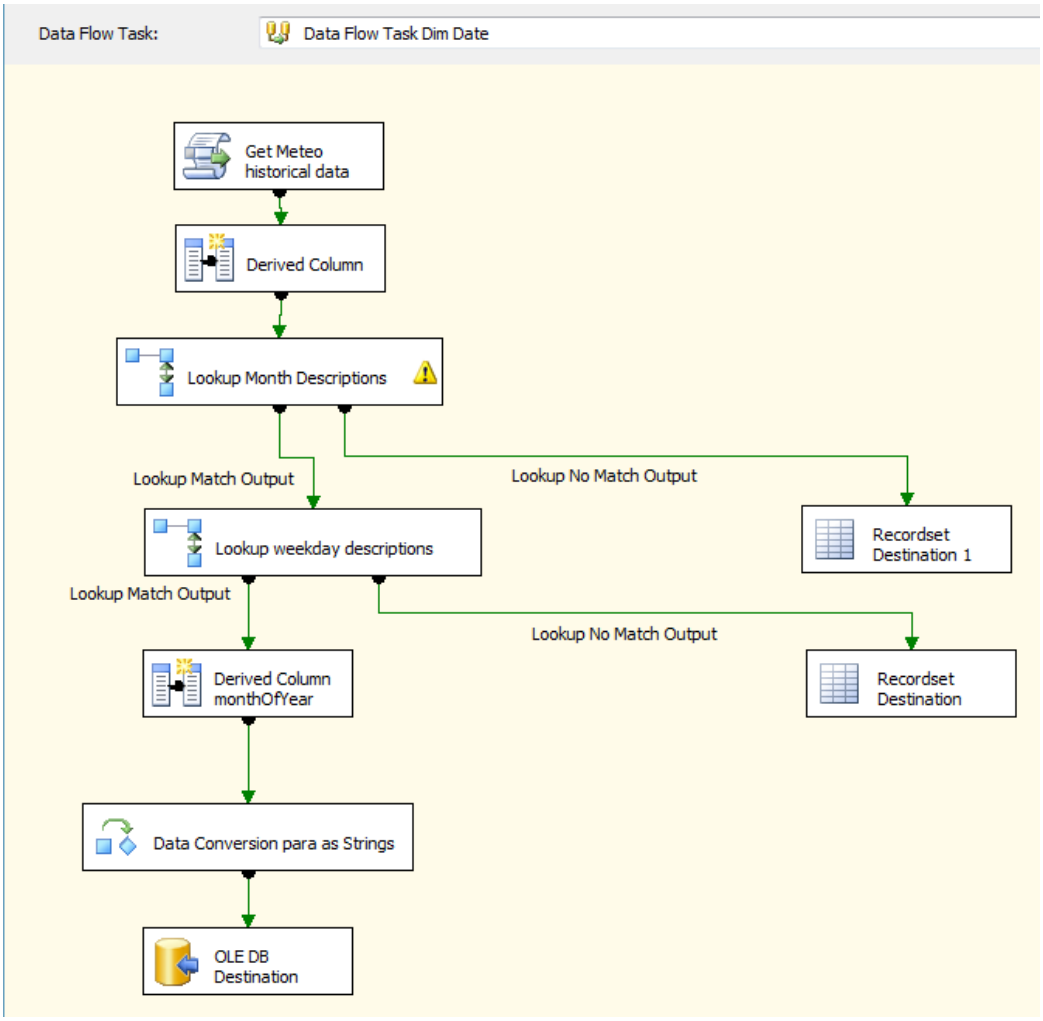



Figura 2.10: Fluxo de dados para Dimensão Data.

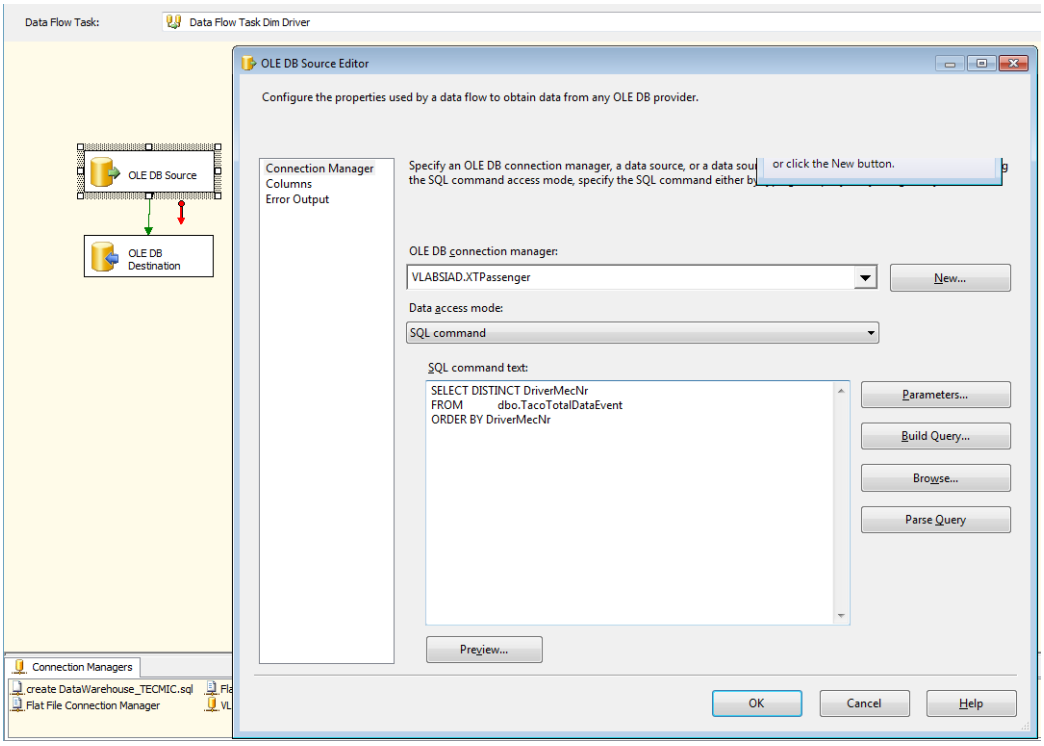



Figura 2.11: Fluxo de dados para a Dimensão Driver.

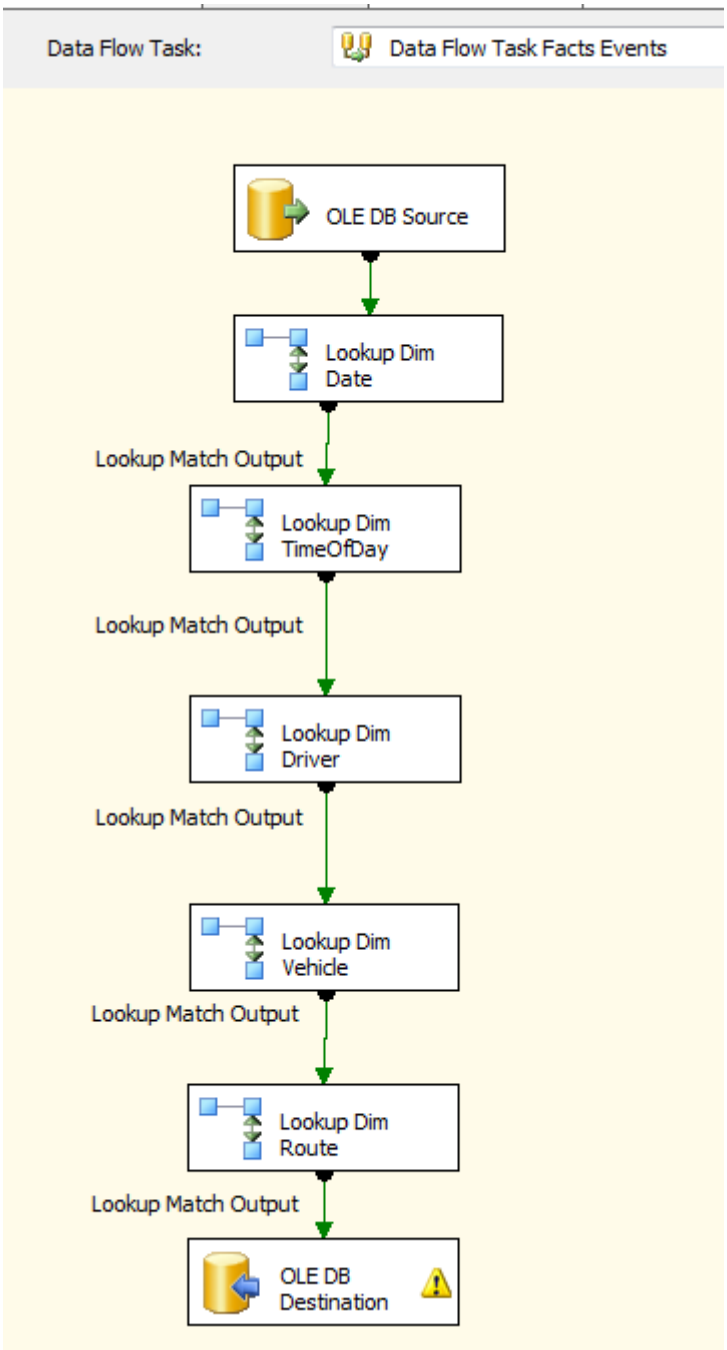



Figura 2.12: Fluxo de dados para Factos.

### 2.3.3 Descrição de factos

Este trabalho baseia-se em dados:

**Operacionais** obtidos por um sistema instalado em veículos que permite a recolha dos valores, disponibilizados pelos barramentos de controlo do veículo, mais um conjunto de sensores instalados nas viaturas. Estes barramentos permitem a troca de dados entre diversos micro-controladores e comunicam entre si usando a norma CAN bus [6];

**Meteorológicos** obtidos por interrogação do sítio internet Weather Underground, de forma a obter gratuitamente os dados históricos da observação meteorológica, assumindo um ponto diário como representativo das condições climatéricas da área geográfica em que os dados operacionais foram recolhidos, que se escolheu ser o do aeroporto internacional dessa área.

Procederemos pois à caracterização de cada um destes dois tipos de dados, (em 2.3.3.1 e 2.3.3.2) e depois já enquanto factos do armazém de dados do ponto de vista da análise OLAP e finalmente como atributos de estruturas de dados então já numa perspectiva de aplicação de técnicas de prospecção de dados para extracção de conhecimento (em 3.1.2.1 e 3.1.2.2).

#### 2.3.3.1 Caracterização de dados operacionais

Os dados operacionais são gerados por eventos e a cada evento são registados os valores de variáveis de tipo contador, logo com andamento monótono. Tal facto tornou necessário que no processo de ETL fosse necessário resolver a monotonia, optando-se por agrupar os registos do sistema OLTP por vista SQL de extracção das variações das variáveis que se consideraram pertinentes para a granularidade da analise desejada sobre os factos, transformando-as de seguida em rácios, o que levou a que estejam disponíveis 1 698 295 registos de partida para a construção do armazém de dados que depois de limpos deixam disponíveis 397

261 factos para submeter às análises OLAP e de prospecção de padrões. A semântica de cada coluna dos dados operacionais 2.2.2 está descrita nas respostas dos peritos de negócio.

#### 2.3.3.2 Caracterização de dados meteorológicos

Os dados históricos de observação meteorológica METAR acrescentam mais vinte variáveis às recolhidas durante a operação dos veículos, a saber:

1. MaxTemp - Temperatura máxima registada no dia em graus Celsius;
2. MedTemp - Temperatura média registada no dia em graus Celsius;
3. MinTemp - Temperatura mínima registada no dia em graus Celsius;
4. DewPoint - Temperatura em graus Celsius de condensação de humidade em orvalho;
5. MeanDewPoint - Temperatura média durante o dia em graus Celsius de condensação de humidade em orvalho;
6. MinDewPoint - Temperatura mínima durante o dia em graus Celsius de condensação de humidade em orvalho;
7. MaxHumidity - Percentagem máxima de humidade do ar;
8. MeanHumidity - Percentagem média de humidade do ar;
9. MinHumidity - Percentagem mínima de humidade do ar;
10. MaxSeaLevelPressure - Pressão máxima em hPa do ar ao nível médio da água do mar;
11. MeanSeaLevelPressure - Pressão média em hPa do ar ao nível médio da água do mar;
12. MinSeaLevelPressure - Pressão mínima em hPa do ar ao nível médio da água do mar;

13. MaxVisibility - Visibilidade máxima em km;
14. MeanVisibility - Visibilidade média em km;
15. MinVisibility - Visibilidade mínima em km;
16. MaxWindSpeed - Velocidade máxima do vento em km/h;
17. MeanWindSpeed - Velocidade média do vento em km/h;
18. MaxWindGustSpeed - Velocidade máxima da rajada de vento em km/h;
19. CloudCoverage - Parte do céu com nuvens em oktas;
20. Events - Observações de eventos como trovoada, chuva ou nevoeiro.

#### 2.3.3.3 Integração de dados operacionais com dados meteorológicos

Esta integração é feita no momento do processamento do ETL ao incluir como características de um dia os dados meteorológicos, conforme se descreve em 2.3.2, pois apenas os dados operacionais são considerados factos sobre os quais se pode analisar a eficiência energética enquanto que os dados meteorológicos são considerados possíveis factores de análise e influência da condução.

#### 2.3.3.4 Definição do cubo multidimensional

Estando o protótipo de armazém de dados elaborado, iniciou-se a elaboração do cubo multidimensional no SSAS tendo optado por uma estratégia de armazenamento de pré agregados de tipo MOLAP[4] pelas vantagens que apresenta na performance de resposta a interrogações típicas em OLAP.

[4]Multidimensional On-line Analytical Processing

Como no caso presente apenas há um tipo de dados e uma perspectiva inerente ao processo de negócio considerado, a matriz do barramento do armazém de dados apenas apresenta um *data mart* que utiliza todas as dimensões identificadas em 2.3.2, conforme se apresenta na Figura 2.13

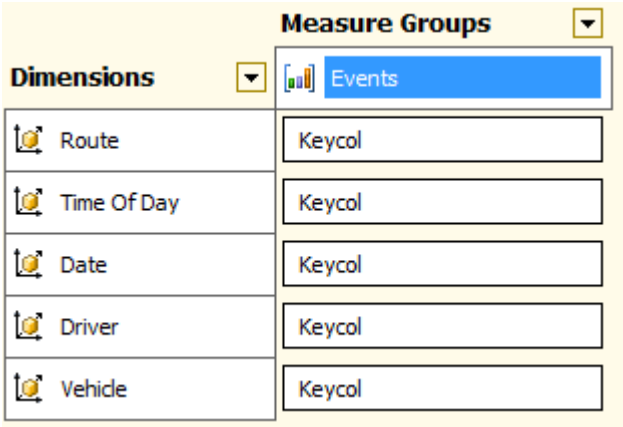



Figura 2.13: Matriz do barramento do Data Warehouse.

Como transparece da estrutura do armazém de dados apresentado em 2.3 das dimensões identificadas, apenas as de Data, de Hora e de Rota apresentam a complexidade suficiente para permitir interrogações do tipo *Roll-Up/Drill-Down* inerentes à hierarquização dos atributos delas contidos pelo que apenas para estas se elaboraram as relação de atributos e hierarquias não naturais em que se estabelece uma relação de 1 para n:

1. Hora

   AM PM, e;

   Parte do dia aproveitando a definição presente nos dados e descrita em 2.2.1.3.

2. Data

   Numero de semana, mês, trimestre, semestre e ano;

Dia da semana, numero da semana, mês, trimestre no ano, semestre e ano;

Dia da semana, mês, trimestre no ano, semestre e ano.

3. Rota

   Número de viagem, Chapa e Rota, e;

   Direcção, variante de rota e Rota.

Com a colaboração dos peritos de negócio acordou-se na criação de uma perspectiva de análise *FuelEfficiency* conforme ilustrado pela Figura 2.14, que encapsulará todos os factos em bruto e que apenas disponibiliza transformações em rácios desses factos tendo-se decidido por:

1. Consumo médio de combustível em litro por 100 km;
2. Percentagem de tempo com rotação do motor em ralenti;
3. Percentagem de tempo com rotação do motor na banda económica;
4. Percentagem de tempo com rotação do motor na banda amarela;
5. Percentagem de tempo com rotação do motor na banda vermelha;
6. Quantidade de eventos de aceleração considerada excessiva por 100 km;
7. Quantidade de eventos de travagem considerada excessiva por 100 km;
8. Percentagem de distância percorrida aproveitando inércia;
9. Percentagem de tempo viajado aproveitado inércia;
10. Quantidade de actuações de travão por 100 km;
11. Quantidade de actuações de acelerador por 100 km,e;
12. Quantidade de actuações de embraiagem por 100 km.

| Cube Objects | Object Type | Perspective Name |
|---|---|---|
| Data Warehouse TECMIC | Name | Fuel Efficiency |
| Km Slope Acc Time | Measure | ☐ |
| Km Total Brake Usage | Measure | ☐ |
| Km Total Clutch Usage | Measure | ☐ |
| Km Total Accelerator Usage | Measure | ☐ |
| Lt Total | Measure | ☐ |
| Rt Total | Measure | ☐ |
| Rt Total Time | Measure | ☐ |
| Rt Idle Time | Measure | ☐ |
| Rt In Gb Time | Measure | ☐ |
| Rt Over Gb Time | Measure | ☐ |
| Rt Rot Max Time | Measure | ☐ |
| Rt Total Time Aux | Measure | ☐ |
| Km Slope Acc Time Back | Measure | ☐ |
| Events Count | Measure | ☐ |
| Dimensions | | |
| Route | CubeDimension | ☑ |
| Time Of Day | CubeDimension | ☑ |
| Date | CubeDimension | ☑ |
| Driver | CubeDimension | ☑ |
| Vehicle | CubeDimension | ☑ |
| Calculations | | |
| Average Fuel Comsuption In Litres per 100 Km | CalculatedMem... | ☑ |
| Time Percentage With Engine Rotation In Green Band | CalculatedMem... | ☑ |
| Time Percentage With Engine Rotation In Idle | CalculatedMem... | ☑ |
| Time Percentage With Engine Rotation In Yellow Band | CalculatedMem... | ☑ |
| Time Percentage With Engine Rotation In Red Band | CalculatedMem... | ☑ |
| Excessive Acceleration Events Per 100 Km | CalculatedMem... | ☑ |
| Excessive Braking Events Per 100 Km | CalculatedMem... | ☑ |
| Inertial Distance Travelled Percentage | CalculatedMem... | ☑ |
| Inertial Time Travelled Percentage | CalculatedMem... | ☑ |
| Brake Usage Per 100 Km | CalculatedMem... | ☑ |
| Accelerator Usage Per 100 Km | CalculatedMem... | ☑ |
| Clutch USage Per 100 Km | CalculatedMem... | ☑ |

Figura 2.14: Detalhe de encapsulamento por perspectiva.

### 2.3.4 Exemplos de interrogações OLAP

São exemplos de questões típicas de análise OLAP, por exemplo *Qual o veículo mais eficiente para uma determinada rota?*, *Qual o condutor mais eficiente numa rota ou de um veículo?* e *Para a operação de um determinado veículo numa determinada rota, qual o condutor com melhor eficiência?*.

Utilizando acções de rotação e partição da perspectiva visível do cubo multidimensional a resposta a este tipo de questões é trivial conforme se apresenta pelas Figuras 2.15 e 2.16, é preferível:

1. utilizar o veículo 886 na rota 1021 do que usá-lo para a rota 1024, e;
2. operar o veículo 192 com o condutor 3038 em vez do 3113.

| | Route Id | | | |
|---|---|---|---|---|
| | | ⊞ 1021 | ⊞ 1024 | ⊞ 1027 |
| Bus Id | Litres per 100 Km | Average Fuel Comsuption In Litres per 100 Km | Average Fuel Comsuption In Litres per 100 Km | Average Fue |
| 114 | | 203.276 | | |
| 839 | | 136.719 | | |
| 112 | | 61.610 | | |
| 126 | | 61.063 | | |
| 795 | | 59.838 | | |
| 865 | | 60.775 | | |
| 797 | | 59.809 | | |
| 650 | | 60.963 | | |
| 153 | | 61.841 | | |
| 950 | | 61.766 | | |
| 149 | | 60.911 | | |
| 187 | | 60.997 | | |
| 140 | | 60.414 | | |
| 792 | | 60.832 | | |
| 128 | | 60.422 | | |
| 188 | | 60.330 | | |
| 272 | | 58.874 | | |
| 150 | | 59.075 | | |
| 737 | | 59.308 | | |
| 141 | | 58.323 | | |
| 129 | | 46.016 | | |
| 886 | | 37.219 | 54.206 | |
| 894 | | 45.808 | | |
| 194 | | | 51.015 | |
| 154 | | | 51.718 | |
| 621 | | | 49.642 | |
| 151 | | | 51.575 | |
| 190 | | | 51.376 | |
| 192 | | | 48.735 | |
| 189 | | 29.983 | 48.247 | |
| 152 | | | 51.663 | |
| 690 | | | | |
| 144 | | | 49.973 | |
| 142 | | | 49.218 | |
| 143 | | | 48.161 | |
| 829 | | | 48.309 | |
| 195 | | | 48.461 | |
| 726 | | | | 45.204 |
| 193 | | | | |

Figura 2.15: Melhor rota para um veículo.

| Dimension | Hierarchy | Operator | Filter Expression |
|---|---|---|---|
| Vehicle | Bus Id | Equal | { 192 } |
| <Select dimension> | | | |

Drop Filter Fields Here

Drop Column Fields Here

| Bus Id | Driver Id | Average Fuel Comsuption In Litres per 100 Km |
|---|---|---|
| 192 | 3113 | 69.087 |
| | 2741 | 64.551 |
| | 5240 | 64.252 |
| | 5291 | 63.374 |
| | 5234 | 62.941 |
| | 3110 | 62.832 |
| | 2816 | 60.760 |
| | 5513 | 60.230 |
| | 4868 | 59.726 |
| | 5468 | 59.411 |
| | 2903 | 59.253 |
| | 3218 | 58.362 |
| | 2603 | 58.027 |
| | 2543 | 58.021 |
| | 5453 | 57.870 |
| | 3083 | 57.807 |
| | 3068 | 57.618 |
| | 2588 | 57.309 |
| | 2858 | 57.094 |
| | 2894 | 56.937 |
| | 5402 | 56.774 |
| | 3245 | 56.555 |
| | 4763 | 56.512 |
| | 5492 | 56.390 |
| | 2885 | 56.251 |
| | 2582 | 56.184 |
| | 5384 | 56.124 |
| | 5030 | 55.887 |
| | 5189 | 55.849 |
| | 4937 | 55.828 |
| | 2465 | 55.736 |
| | 5414 | 55.490 |
| | 3194 | 55.396 |
| | 5153 | 55.264 |
| | 2771 | 55.221 |
| | 5351 | 55.037 |
| | 3107 | 54.908 |
| | 5294 | 54.894 |
| | 5441 | 54.762 |
| | 3038 | 54.738 |

Figura 2.16: Melhor condutor para um veículo.

# Capítulo 3

# Modelação e Resultados

Neste capítulo transita-se do domínio típico das interrogações OLAP - quem, quando, quanto, como, onde? - para o das da aquisição de conhecimento por prospecção em bases de dados - porquê? - e que se mapeia para os passos Modelação e Avaliação da metodologia seleccionada em 1.5. Discute-se a preparação de dados, descrevem-se e aplicam-se aos atributos dos dados disponíveis conceitos de descrição estatística, contextualiza-se a questão da prospecção de dados e a familiarização com a ferramenta escolhida, sendo descritos os modelos de dados elaborados e apresentados os resultados obtidos e principais regras de conhecimento extraídas pelos algoritmos de prospecção de dados em que se empregou cerca de $1/3$ do esforço.

## 3.1 Preparação de Dados

Nesta secção apresentam-se alguns conceitos de descrição estatística, utilizados numa posterior descrição de atributos enquanto actividade preparatória da elaboração de estruturas de dados a submeter às técnicas de prospecção de dados aquando da elaboração de modelos dos dados.

### 3.1.1 Conceitos de descrição estatística

Seja um conjunto de dados numéricos organizado em matriz $X_{n\times d}$ em que $n$ é o número de casos ou linhas e $d$ é o número de atributos ou colunas da matriz e em que cada elemento dessa matriz $x^i_j$, contêm o valor do $i$-ésimo atributo do $j$-ésimo caso.

#### 3.1.1.1 Centralidade

Para se percepcionar características dos valores de um atributo é usual localizar a distribuição de valores à volta do valor médio desse atributo. A média do valor de um atributo $i$ é calculado pela Equação 3.1:

$$\overline{x}^i = \frac{1}{n}\sum_{k=1}^{n} x^i_k \tag{3.1}$$

Se quisermos conhecer o valor do atributo que divide os valores existentes em duas sequências com o mesmo número de elementos, necessitamos de determinar a mediana. Após ordenação crescente dos valores do $i$-ésimo atributo podemos determinar o valor da sua mediana pela aplicação da Equação 3.2:

$$mediana(x^i) = \begin{cases} \frac{1}{2}(x^i_k + x^i_{k+1}) & \text{com } n \text{ par } (n = 2k) \\ x^i_{k+1} & \text{se } n \text{ ímpar } (n = 2k-1) \end{cases} \tag{3.2}$$

Convém também definir:

$Moda$ de um atributo $x^i$ como sendo o valor mais frequente que esse atributo apresenta em todos os $n$ casos.

$Percentil$ e $Quartil$ de um atributo são pontos de divisão do conjunto de valores que esse atributo toma, semelhantes à mediana, mas que utilizam pontos de segmentação arbitrários, por exemplo: o $3^{o}Quartil$, ou $Q3$ de um atributo é o valor para o qual existem $75\%$ de valores inferiores. Essa é também a definição do $Percentil$ $75\%$, ou $P75$, desse atributo.

Quando confrontados com a necessidade de definir quais os valores

atípicos[1] para um atributo $x^i$, é usualmente necessário estimar a sua escala de valores. Para tal é habitual recorrer à amplitude interquartil, ao desvio médio absoluto[2] definido pela Equação 3.4 ou ao desvio mediano absoluto[3] definidos pelas Equações 3.3, 3.4 e 3.5:

$$IQR = Q3 - Q1 \tag{3.3}$$

$$AAD(x^i) = \frac{1}{n}\sum_{k=1}^{n}|x_k^i - \overline{x}^i| \tag{3.4}$$

$$MAD(x^i) = mediana(|x_1^i - \overline{x}^i|, \ldots, |x_n^i - \overline{x}^i|) \tag{3.5}$$

A medida da dispersão e distribuição de valores que um atributo $x^i$ toma são usualmente aferidas pelos seus $p$ momentos, definidos pela Equação 3.6.

$$momento_p(x^i) = \frac{1}{n-1}\sum_{s=1}^{n}(x_s^i - \overline{x}^i)^p \tag{3.6}$$

Tomando $p$ valores inteiros. Quando:

$p =$**1** , obtém-se $0$, o primeiro momento central;

$p =$**2** , obtém-se a $variância$, o segundo momento central;

$p =$**3** , obtém-se a $obliquidade$, o terceiro momento central;

$p =$**4** , obtém-se a $curtose$, o quarto momento central;

**etc** .

À raiz quadrada da $variância$ chama-se $desvio\ padrão$, representado por $\sigma$. É habitual *normalizar* os momentos[4] da distribuição dos valores de um atributo dividindo-os por $\sigma^p$.

[1] outliers
[2] AAD
[3] MAD
[4] Em diante ao referir variância, obliquidade e curtose, consideramos às suas versões normalizadas

#### 3.1.1.2 Dispersão

A variação e dispersão dos valores de um atributo $x^i$ são aferidos quanto ao intervalo e variância normalizada definidos pelas Equações 3.7 e 3.8.

$$intervalo(x^i) = max_{k=1...n}(x_k^i) - min_{k=1...n}(x_k^i) \tag{3.7}$$

$$variância(x^i) = \frac{momento_2(x^i)}{\sigma^2} = \frac{1}{\sigma^2(n-1)} \sum_{s=1}^{n} (x_s^i - \overline{x}^i)^2 \tag{3.8}$$

#### 3.1.1.3 Distribuição

A simetria da distribuição dos valores de um atributo $x^i$ em relação à média é a $obliquidade$ e define-se pela Equação 3.9.

$$obliquidade(x^i) = \frac{momento_3(x^i)}{\sigma^3} = \frac{1}{\sigma^3(n-1)} \sum_{s=1}^{n} (x_s^i - \overline{x}^i)^3 \tag{3.9}$$

A dispersão ou achatamento da distribuição dos valores de um atributo $x^i$ chama-se $curtose$ e define-se pela Equação 3.10.

$$curtose(x^i) = \frac{momento_4(x^i)}{\sigma^4} = \frac{1}{\sigma^4(n-1)} \sum_{s=1}^{n} (x_s^i - \overline{x}^i)^4 \tag{3.10}$$

Quando comparado com uma distribuição normal ou Gaussiana com média $0$ e $\sigma$ 1, a distribuição de valores de um atributo $x^i$ pode estar mais concentrada num dos lados da moda, ter um pico mais ou menos acentuado e ter uma dispersão maior ou menor, nomeadamente:

- obliquidade

  $= 0$: a distribuição é simétrica;

  $> 0$: a distribuição concentra-se à esquerda;

$< 0$: a distribuição concentra-se à direita.

- curtose

  $= 3$: a mesma dispersão da distribuição normal;

  $> 3$: uma dispersão mais concentrada do que a da distribuição normal;

  $< 3$: uma dispersão menos concentrada do que a da distribuição normal;

#### 3.1.1.4 Covariância

Até este momento apenas nos concentramos na definição das medidas estatísticas para descrição *per si* de cada atributo da matriz de dados, $X_{n\times d}$, mas convêm medir também se algum par de atributos tem uma variação relacionável, ou seja qual a forma como variam em conjunto. Isto pode ser avaliado recorrendo a Equação 3.11:

$$covariância(x^r, x^s) = \frac{1}{n-1}\sum_{k=1}^{n}(x_k^r - \overline{x}^r)(x_k^s - \overline{x}^s) \tag{3.11}$$

Quando a $covariância$ é:

- $= 0$: os atributos não variam de forma linear;
- $> 0$: os atributos variam directamente,e;
- $< 0$: os atributos variam inversamente.

Acresce que um par de atributos cuja escala de valores seja maior do que outro terá maior $covariância$ ainda que ambos os pares variem com à mesma proporção.

Usando esta Equação e a matriz de dados $X_{n\times d}$, obtêm-se uma matriz de $covariância$ $Cov_{d\times d}$, apresentado a diagonal a $variância$ dos $d$ atributos.

#### 3.1.1.5 Correlação

Para eliminar o efeito da escala na $covariância$, é habitual recorrer à $correlacão$ de acordo com a definição da Equação 3.12:

$$correlacão(x^r, x^s) = \frac{covariância(x^r, x^s)}{\sigma_{x^r}\sigma_{x^s}} \tag{3.12}$$

### 3.1.2 Análise exploratória de atributos

Em preparação da elaboração das estruturas e modelos de dados descritos em 3.2.6 e 3.2.7, tendo em atenção as definições estatísticas referidas em 3.1.1, optou-se por analisar os atributos disponíveis enquanto dados uni e multivariados, para familiarização com os mesmos, avaliar o ruído e indagar a existência de redundância. Optou-se por apresentar as medidas dos atributos em modo gráfico recorrendo a diagramas de caixa e bigodes[5], histogramas de valores e representações em projecções de duas dimensões de dois espaços tridimensionais, em que os eixos do plano horizontal são, no primeiro caso o condutor e rota, e no segundo o condutor e o veículo.

#### 3.1.2.1 Dados univariados

Objectos ou casos descritos por dados univariados são descritos por um só atributo e o $i$-ésimo atributo da matriz $X_{n\times d}$ pode representar-se por $x^i = \{x_1, x_2, \ldots, x_n\}$ pode então ser analisado como como se explanou em 3.1.1 e apresentar os resultados da forma gráfica descrita em 3.1.2.

Estes quatro diagramas, produzidos com o código listado no apêndice F, apresentam para cada um dos atributos:

1. *de caixa e bigodes*, que descreve:

   $Q1$ - a linha inferior da caixa;

   $Q2$, ou mediana - a linha central da caixa;

[5] boxplot

$Q3$ - a linha superior da caixa;

$\times$ - a média;

bigode inferior - o limite de $Q1 - 1,5 \times IQR$, e;

bigode superior - o limite de $Q3 + 1,5 \times IQR$.

2. a distribuição dos valores do atributo;

3. a representação dos valores do atributo (no eixo vertical) *versus* as dimensões Condutor e Rota, e;

4. a representação dos valores do atributo (no eixo vertical) *versus* as dimensões Condutor e Bus.

As Figura 3.1 a 3.4 apresentam o conjunto de diagramas que constituem a descrição de um atributo enquanto dados univariados, remetendo a consulta dos mesmos gráficos para todos os restantes ao anexo H.

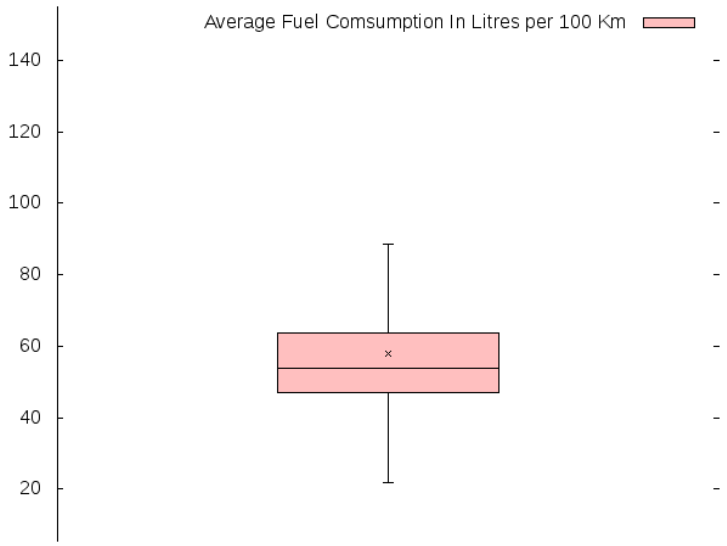



Figura 3.1: Diagrama de caixa do atributo Consumo médio de combustível.

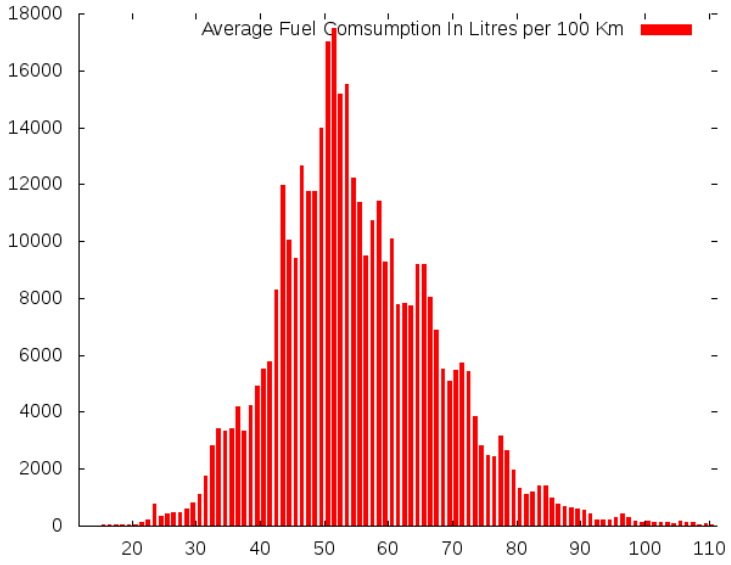



Figura 3.2: Histograma do atributo Consumo médio de combustível.

Resulta da análise das Figuras 3.1 e 3.2 que este atributo apresenta uma distribuição de valores interessante, quase simétrica, próxima de uma distribuição normal e com existência de pouco valores espúrios.

As Figuras 3.3 e 3.4, permitem a familiarização com distribuição de casos nos espaços definidos pelos pares de dimensões de análise mais relevantes, a saber {Condutor, Rota} e {Condutor, Veículo}, sendo trivial a adaptação à visualização noutros espaços, por exemplo {Rota, Veículo}.

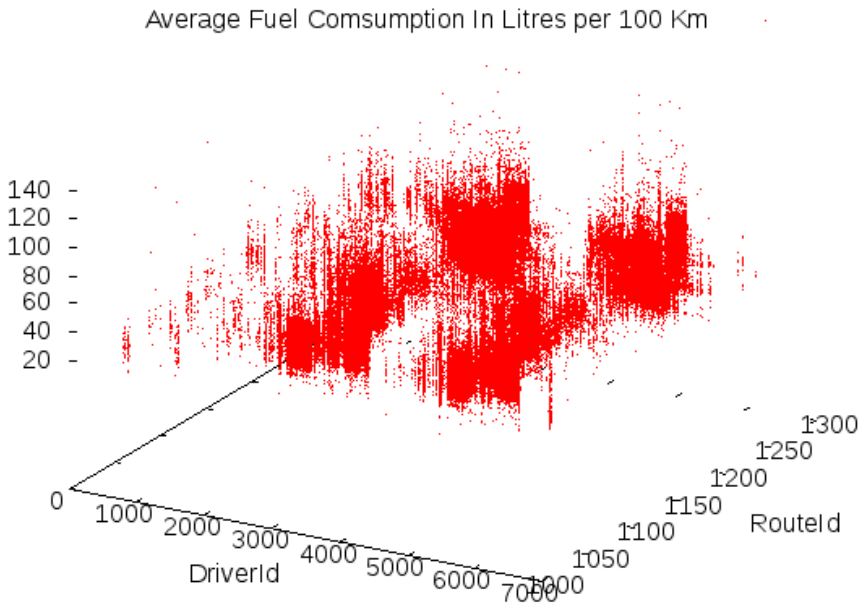



Figura 3.3: Descrição atributo Consumo médio de combustível versus Condutor e Rota.

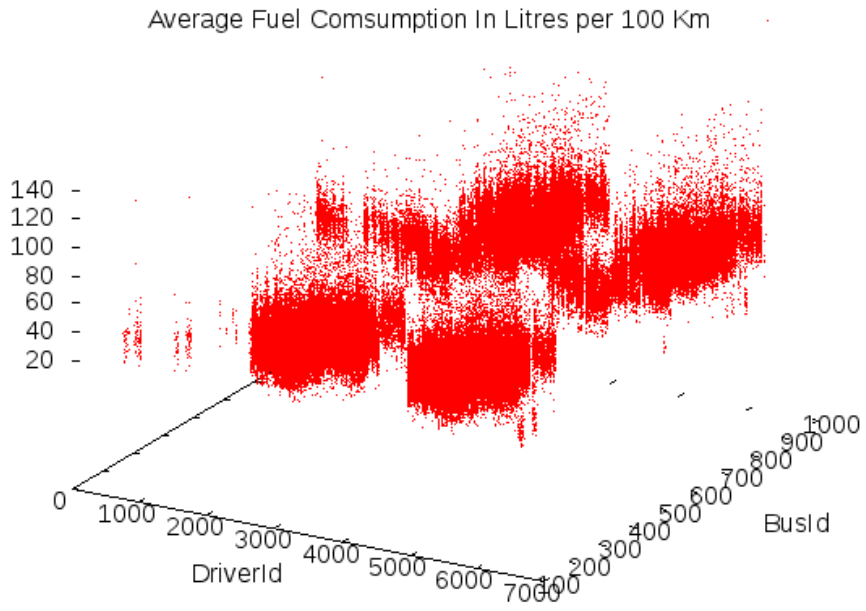



Figura 3.4: Descrição atributo Consumo médio de combustível versus Condutor e Veículo.

### 3.1.2.2 Dados multivariados

Pelas razões descritas em 3.1.1.5 analisou-se a $correlacão$ dos atributos apresentado-se a matriz respectiva na Figura 3.5.

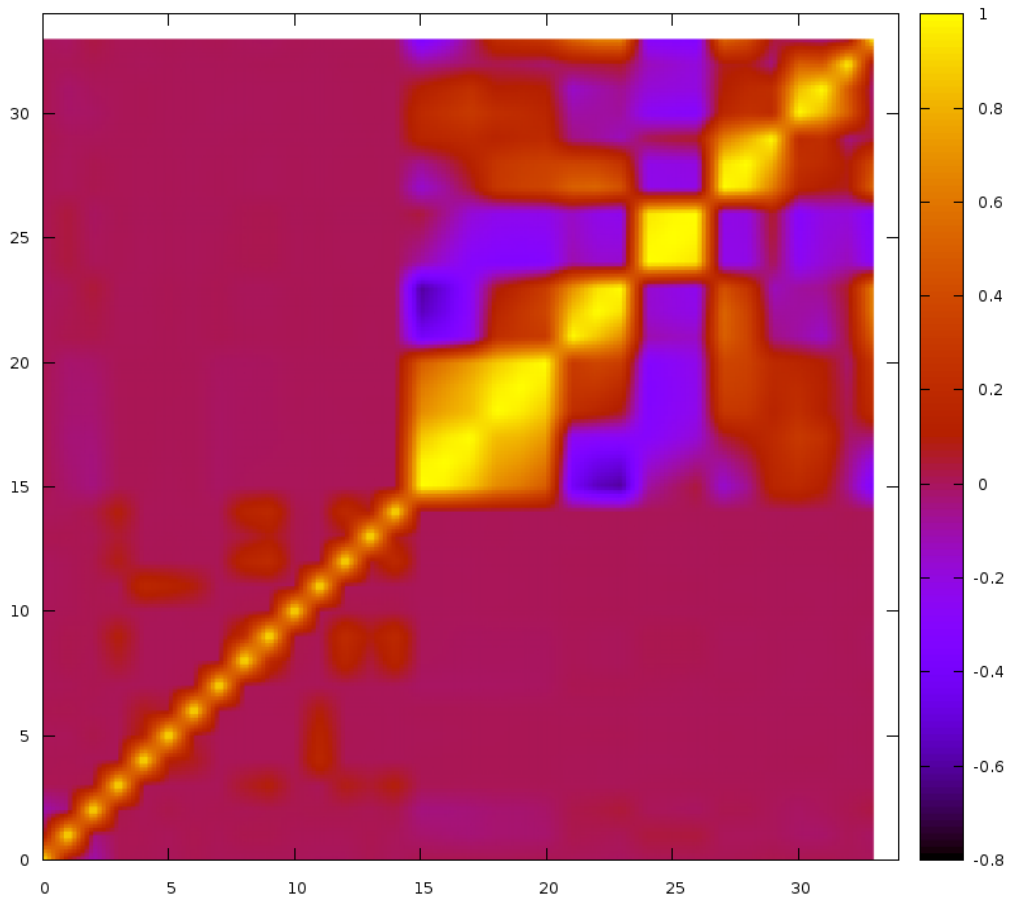



Figura 3.5: Matriz de correlação dos atributos numéricos de modelos.

Na Figura 3.5 os eixos vertical e horizontal representam os atributos:

1. Condutor;
2. Veículo;
3. Rota;
4. Consumo médio de combustível em litro por 100 km;

5. Percentagem de tempo com rotação do motor na banda económica;
6. Percentagem de tempo com rotação do motor em ralenti;
7. Percentagem de tempo com rotação do motor na banda amarela;
8. Percentagem de tempo com rotação do motor na banda vermelha;
9. Quantidade de eventos de aceleração considerada excessiva por 100 km;
10. Quantidade de eventos de travagem considerada excessiva por 100 km;
11. Percentagem de distância percorrida aproveitando inércia;
12. Percentagem de tempo viajado aproveitando inércia;
13. Quantidade de actuações de travão por 100 km;
14. Quantidade de actuações de embraiagem por 100 km;
15. Quantidade de actuações de acelerador por 100 km;
16. Temperatura máxima;
17. Temperatura média;
18. Temperatura mínima;
19. Ponto de orvalho;
20. Média do ponto de orvalho;
21. Mínimo do ponto de orvalho;
22. Humidade máxima;
23. Humidade média;
24. Humidade mínima;

25. Máxima da pressão atmosférica;
26. Média da pressão atmosférica;
27. Mínima da pressão atmosférica;
28. Visibilidade máxima;
29. Visibilidade média;
30. Visibilidade mínima;
31. Velocidade máxima do vento;
32. Velocidade média do vento;
33. Velocidade máxima da rajada de vento;
34. Parte do céu enublada.

A análise da matriz de correlação sugere que não há significativa correlação entre os atributos do grupo de dados operacionais (de 1 a 15) e o grupo dos dados meteorológicos (de 16 a 34), constatando-se que estes últimos apresentam um grau de correlação entre si. 6

## 3.2 Prospecção de dados para extracção de conhecimento

### 3.2.1 Conceitos

Apresentaremos nesta sub-secção um conjunto de conceitos em que se basearam as actividades de prospecção de dados.

A extracção de conhecimento em bases de dados, por máquinas, é um assunto que emerge na confluência das áreas de inteligência artificial enquanto abarcando a aprendizagem automática, estatística descritiva e

bases de dados, numa tentativa de por métodos automáticos ou semi-automáticos aprender novos factos até então soterrados, pelo que convirá assentar num significado para ”aprender”.

Segundo um dicionário de língua Portuguesa, aprender é adquirir conhecimento ou domínio (de assunto, matéria, etc.) através do estudo ou da prática, obtendo conhecimento. Porém a definição de conhecimento é um assunto que a epistemologia continua a estudar, mas para a qual não existe ainda uma definição absoluta. Numa definição clássica, conhecimento seria ”um acreditar verdadeiro e justificado” [16] e assim um individuo $I$ conheceria uma determinada proposição $P$ $sse$:

- $P$ é verdadeira;
- $I$ acredita em $P$, e;
- existe uma justificação para a crença de $I$ em $P$.

Em meados do Séc. XX, Gettier [18] provou que essa definição não era suficiente pelo que continua a tentativa de chegar a uma nova definição.

Ainda que ”o desempenho de um chinelo novo melhore após algum tempo sem que disso se possa dizer que aprendeu a forma do pé do proprietário” [24], não nos querendo substituir aos que continuam a estudar a questão, pragmaticamente diremos que ao ”aprender há uma melhoria de performance no desempenho futuro de alguma função na qual estamos empenhados”, com acumulação de experiência, [20], mesmo que desta forma possa não se distinguir aprendizagem de treino.

Esses ganhos de performance são normalmente potenciados pelas técnicas de prospecção de dados nas vertentes:

descritiva - em que autonomamente a máquina explora um conjunto de dados para produzir uma descrição sumária, encontrar ou grupos semelhantes, ou relações de associação entre os dados, e;

preditiva - em que se supervisiona a máquina, treinando-a a partir de um subconjunto de exemplos, procurando obter dela uma estimativa $\hat{f}$ de

uma hipotética função $f(X_{n\times(d-1)}) = Y$ que transforma um conjunto de atributos de entrada num atributo de saída, contínuo por regressão ou discreto por classificação e aferir da qualidade de $\hat{f}$ com um conjunto de teste.

Imagine-se que as descrições a produzir pelo processo de aprendizagem resultam num conjunto de preposições $P = p_1, p_2, \ldots, p_n$ que descrevem os dados $X_{n\times d}$, sendo que em cada preposição $p_i$ apenas pode haver um termo com referência a atributo ($p_i = t_1, t_2, \ldots, t_d$), que todos os atributos dos dados são discretos e que cada atributo $x^i$ assume um no máximo de uma quantidade limite de estados $l_{x^i}$.

Consideremos também que listamos todos os conjuntos possíveis de todas as regras e que vamos procurar dessa lista o conjunto de descrições que descreve os dados no máximo com tantas regras quantos os casos.

O pesquisa decorrerá sobre um espaço com um número muito grande mas finito de hipóteses dado pela Equação 3.13:

$$volume_{X_{n\times d}} = \left(\prod_{i=1}^{d} l_{x^i}\right)^n \tag{3.13}$$

Se o todos os atributos tiverem o mesmo limite de estados, $l$ o

$$volume_{X_{n\times d}} = \left(l^d\right)^n$$

Estamos pois perante um problema que rapidamente se pode tornar impraticável, devido em parte[6] à *maldição da dimensionalidade*. Por esta razão decidimos analisar as rotas apenas pela granularidade de Rota, descartando as variáveis de variantes de Rota, seguindo as recomendações dos especialistas de negócio e ainda porque mesmo se se considerassem somente as dimensões Condutor, Veículo e Rota com as suas variantes, os 397 261 casos que seriam descritos num espaço com um volume de $volume = 1487 \times 44 \times 5354 = 3.5 \times 10^8$ pelo e a densidade de casos andaria na ordem dos $11 \times 10^{-}3$ ou seja, tão espalhados que podem

[6] $l^d$

levar a dificuldades na detecção de padrões pela ”influência da presença de atributos irrelevantes” [20, pág. 235].

Convém notar que muitas técnicas de prospecção de dados não são baseadas em pesquisa, mas o conceito de um espaço $d - dimensional$ no qual estão $n$ pontos nas coordenadas dos valores dos atributos é uma imagem mental que geralmente auxilia o contexto em que são aplicadas e a razão de apresentar em 3.1.2.1 e no apêndice H, as Figuras do tipo de 3.3 e 3.4.

No campo de aprendizagem automática da inteligência artificial são por vezes aplicados princípios lógicos - em que existindo certeza e informação sobre os dados, existe um mecanismo de inferir qual a descrição em que se enquadram ou qual o valor da saída ”derivando dessa base de conhecimento por acção de um mecanismo de inferência” [21].

De uma forma geral [22] é possível agrupar os métodos usados em prospecção de dados na aplicação:

1. de mecanismos de pesquisa - em que se procura a melhor das soluções no espaço de todas as hipóteses;

2. de princípios estatísticos - em que por aplicação de princípios estatísticos se infere probabilisticamente;

3. do princípio da semelhança entre casos do mesmo conceito - em que por aplicação de métricas de distância se procura o resultado do(s) caso(s) mais próximos;

4. de estratégias de optimização - em que se optimiza o resultado a estimativa da função que produz o atributo de saída.

A Figura 3.6 apresenta um resumo possível [22] dos métodos mais comuns e clássicos de acordo com a natureza autónoma ou supervisionada da sua operação.

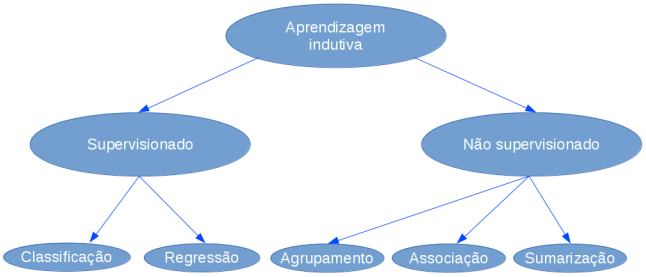



Figura 3.6: Taxonomia de métodos de prospecção de dados [22, pág. 15].

Sejam de que natureza forem, a descoberta de descrições em que se enquadram ou a estimativas do valor de uma saída perante novos casos, as actividades de prospecção de conhecimento em bases de dados constroem modelos sobre a informação contida nos dados disponibilizados para treino, modelos esses habitualmente depois validados com base em dados de teste[22, pág. 185-186].

#### 3.2.1.1 Teoria da informação

Segundo [29], "sempre que um atributo é numérico, contínuo e não binário" a selecção de atributos é realizada por interesse que é calculado tendo por base a entropia.

Shannon [19] mostrou que a incerteza sobre os valores que um atributo $x^i$ em que cada valor que toma tem uma probabilidade $p(x^i_j)$ correlaciona-se directamente com a quantidade de informação que se obtêm conhecelo. Essa quantidade de informação chamada de $Entropia\ de\ x^i$ ou $H(x^i)$ determina-se pela a Equação 3.14:

$$H\left(x^i\right) = -\sum_{j=1}^{n} p\left(x^i_j\right) \times \log_2\left(p(x^i_j)\right)\ \ [bit] \tag{3.14}$$

A entropia apresenta as seguintes propriedades:

1. $H\left(x^i\right) \in [0, \log_2(n)]$;

2. $H\left(x^i\right) = 0\ sse\ \exists j : p\left(x^i_j\right) = 1$ e;

3. $max\left(H(x^i)\right) = \log_2(n)\ sse\ p\left(x_j^i\right) = p\left(x_k^i\right), \forall j \neq k.$

Shannon também apresentou que dados dois atributos $x^i$ e $x^j$ a entropia conjunta $H\left(x^i, x^j\right)$:

$$H\left(x^i, x^j\right) \begin{cases} = & H\left(x^i\right) + H\left(x^j\right) & \text{se } x^i \text{ independente de } x^j \\ < & H\left(x^i\right) + H\left(x^j\right) & \text{se } x^i \text{ dependente de } x^j \end{cases} \tag{3.15}$$

Assim dados dois atributos $x^i$ e $x^j$ o ganho de informação sobre $x^i$ ao conhecer-se $x^j$ é dado pela entropia condicional, $H\left(x^i|x^j\right)$, conforme a Equação 3.16:

$$H\left(x^i|x^j\right) = -\sum_{k,l} p\left(x_k^i, x_l^j\right) \times \log_2\left(\frac{p\left(x_k^i, x_l^j\right)}{p\left(x_l^j\right)}\right)\ [bit] \tag{3.16}$$

A entropia condicional apresenta as seguintes propriedades:

1. $H\left(x^i|x^i\right) = 0$;
2. $H\left(x^i|x^j\right) = H\left(x^i\right) + H\left(x^j\right)\ sse\ x^i, x^j$ independentes e;
3. $H\left(x^i|x^j\right) = H\left(x^i, x^j\right) - H\left(x^j\right)$.

Ao conhecer um dos atributos há uma redução de incerteza, i.e. de informação que é dada pela informação mútua conforme a Equação 3.17:

$$I\left(x^i, x^j\right) = \sum_{k,l} p\left(x_k^i, x_l^j\right) \times \log_2 \frac{p\left(x_k^i, x_l^j\right)}{p\left(x_k^i\right), p\left(x_l^j\right)} \tag{3.17}$$

### 3.2.2 Modelação em Microsoft SQL Server SSAS

A plataforma em causa utiliza um conjunto de extensões para prospecção de dados, DMX, à linguagem estruturada de interrogação de bases de dados, SQL, sendo que ”DMX é a linguagem que transforma os dados que temos”, relacionais organizados em ”tabelas de registos e colunas nos requeridos pelos algoritmos de prospecção, casos e atributos” [25], recorrendo a dois objectos principais: estruturas e modelos de dados.

As estrutura de dados permitem a segmentação em subconjuntos de treino e teste, transformando e adaptando o tipo das colunas de dados aos atributos exigidos pelos algoritmos de prospecção, permitindo encapsular diversos modelos de exploração e introduzindo uma camada de abstracção e de limitação do acoplamento entre as fontes de dados e o processo de descoberta de conhecimento. Apresentaremos nas subsecções seguintes os passos e conceitos necessários à elaboração da modelação, remetendo para a literatura [25, 26] a descrição exaustiva das funcionalidades e implementação de algoritmos de análise disponibilizadas por esta ferramenta.

Não obstante prossegue-se nas próximas subsecções a explicitação das principais opções tomadas enquadrando-as no âmbito das temáticas clássicas de prospecção de dados.

### 3.2.3 Escolha de atributo alvo

Sempre que os métodos a empregar são supervisionados, é necessário estabelecer qual dos atributos disponíveis se pretende classificar ou estimar. A esse atributo chamaremos de atributo alvo e aos restantes atributos de entrada. No caso da análise realizada foi escolhida a variável ”Consumo médio em litros de combustível por 100 km” como atributo alvo sendo os restantes trinta e três atributos utilizados como entradas desta classe de algoritmos.

### 3.2.4 Tipificação dos atributos de entrada

Os atributos disponíveis nos dados são caracterizáveis quanto ao seu tipo em:

1. Qualitativos - Condutor, Veículo, Rota, Parte do dia;

2. Quantitativos

   Contínuos - os restantes enumerados em 3.1.2.2;

Discretos - os restantes das dimensões data e hora descritos em 2.3.3.4.

Esta análise é necessária pois alguns dos algoritmos requerem apenas dados categóricos pelo que é necessário aplicar uma técnica de os discretizar. Note-se que a discussão dual da transformação de valores categóricos em numéricos não é necessária pois a implementação Microsoft de todos os algoritmos de prospecção de dados aceita dados discretos.

### 3.2.5 Discretização

Sempre que os algoritmos de prospecção de dados o requerem os atributos quantitativos contínuos foram discretizados em até cinco intervalos cada um ou abrangendo áreas iguais do intervalo de valores ou por técnica de agrupamento de 1000 amostras aleatórias por maximização da expectativa conforme se descreve em detalhe em [27]. Para tal efeito parametrizaram-se com o tipo ”DISCRETIZED” disponibilizado pela ferramenta os atributos quantitativos contínuos.

#### 3.2.5.1 Selecção de proeminência

Conforme se apresentou em 3.2.1, por causa da *maldição da dimensionalidade*, é usual ser necessário agregar ou seleccionar dos atributos disponíveis aqueles que mais informação contêm sobre o atributo alvo ou para a descrição sumária dos dados.

Como se detalhou no capítulo 2 em especial na subsecção 2.3.3.4 acordou-se com a colaboração dos especialistas do negócio num conjunto de atributos a considerar, eliminado manualmente outros.

A agregação de atributos é uma possível forma de conter o problema da dimensionalidade e segundo [22] uma das técnicas mais utilizadas é a análise de componentes principais [31] . Porque no presente caso se pretende facilitar a interpretação dos resultados da aplicação das técnicas de prospecção, convêm preservar os valores dos atributos pelo que se preteriu esta abordagem.

Segundo [24, pág. 308] a selecção automática de atributos é útil e segundo [25] é mesmo indispensável em todas as ferramentas de prospecção de dados. No caso da implementação Microsoft existe para cada algoritmo a possibilidade de escolher e parametrizar o método pelo qual a selecção de atributos é feita, havendo para cada algoritmo as alternativas descritas em [29]. É referido na descrição citada que a selecção se aplica não só a atributos como também ao número de estados dos atributos cujo máximo é parametrizado sendo os estados menos interessantes agrupados e tratados como se em falta.

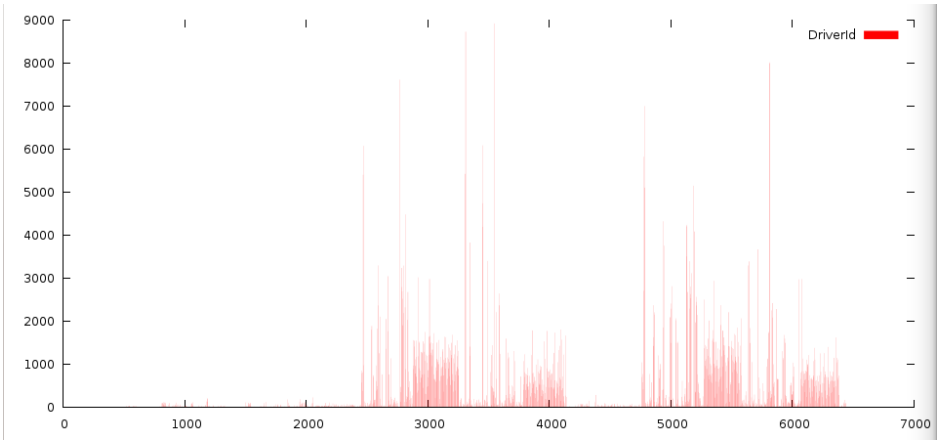
Note-se que devido ao esforço prévio de construção do armazém de dados, conforme transparece da análise uni e multivariada dos atributos, há acrescidas garantias de consistência, reduzida presença de ruído e redundância, sabendo-se que não existem valores omissos.

### 3.2.6 Estrutura de dados

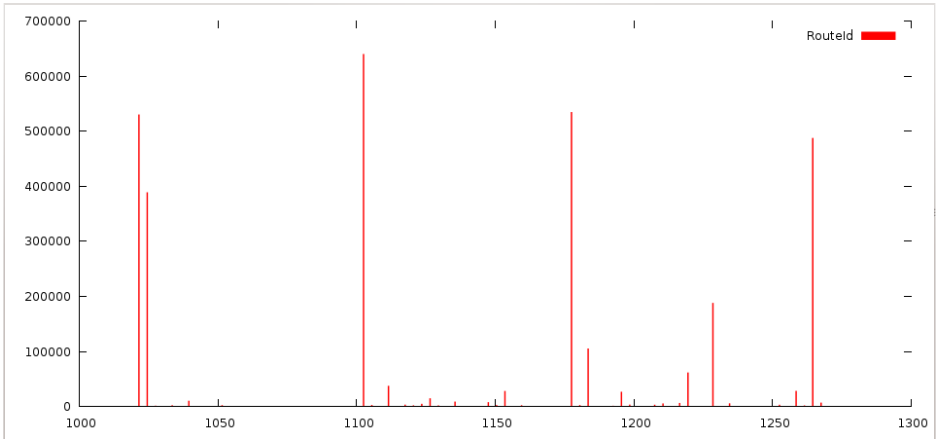
Conforme os objectivos, apoiados nos resultados da análise exploratória de atributos descrita em 3.1.2, escolhemos como principais perspectivas de análise do consumo de combustível segundo o condutor, a rota e o veículo, decidindo-se ainda numa primeira fase segmentar os atributos disponíveis nos conjuntos operacionais e meteorológicos.

Desta forma elaborou-se para cada perspectiva de análise uma estrutura de dados, todas ligadas à mesma fonte de dados, preservando trinta por cento dos casos para teste, discretizando todos os atributos contínuos pois a ”análise por intervalos é mais fácil que por valores e variâncias” [25, pág. 97].

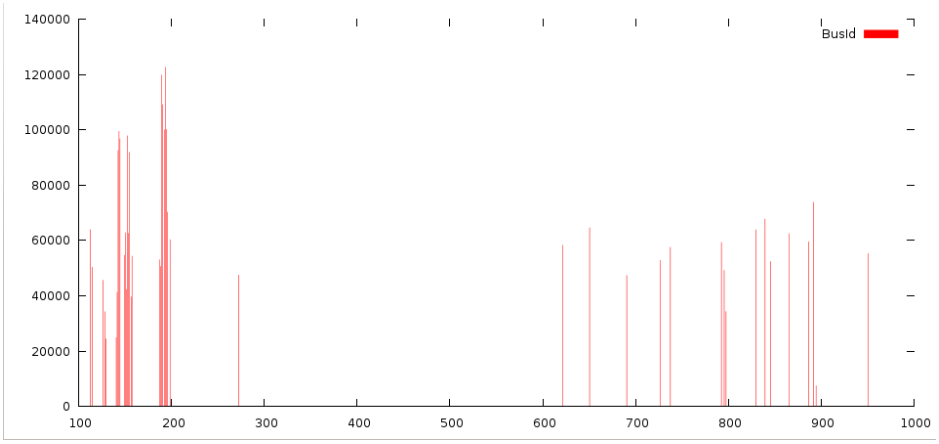
Note-se que ao analisar os dados disponíveis de acordo com estas três perspectivas, respeitam ao desempenho de 1487 Condutores, 73 Rotas e 44 Veículos, com a concentração de dados ilustrada pela Figura 3.7, pela qual se evidencia uma elevada concentração de casos num restrito número de veículos e rotas o que limitará os resultados segundo estas perspectivas.




(a) Por condutor.



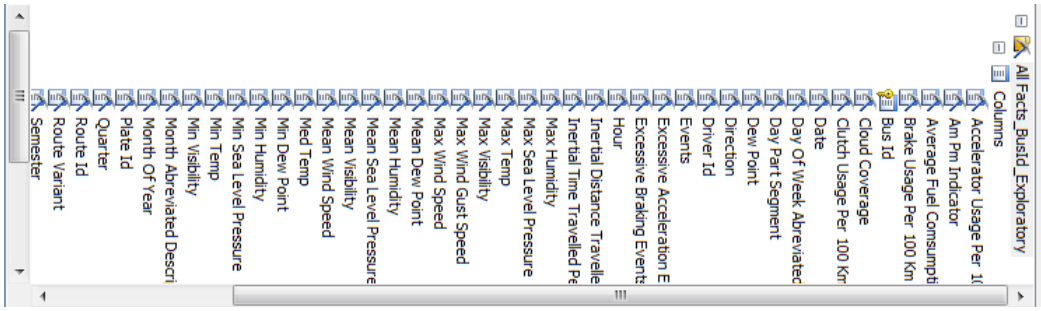

(b) Por rota.



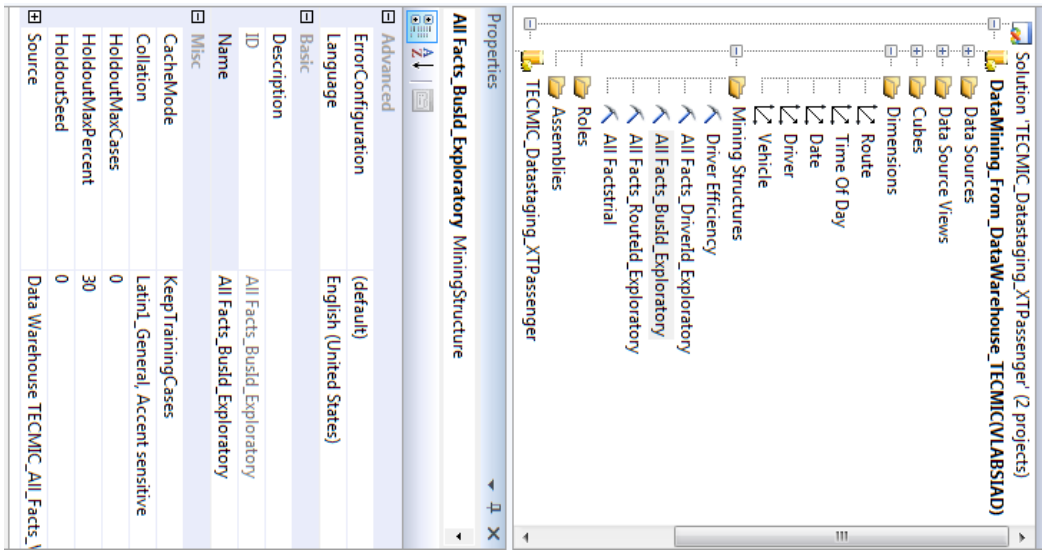

(c) Por veículo.

Figura 3.7: Casos segundo perspectivas

A Figura 3.8 ilustra a elaboração da estrutura do ponto de vista do condutor, sendo as restantes semelhantes.




(a) atributos.




(b) propriedades.

Figura 3.8: Estrutura de dados para análise exploratória do caso condutor

Todas as estruturas de dados elaboradas têm por base a extracção dos factos do armazém de dados realizada por vista sobre o cubo multidimensional cuja consulta do código se remete para o anexo G.

### 3.2.7 Modelos de dados

Decidiu-se realizar a prospecção com a elaboração de modelos baseados no algoritmo Naive-Bayes, fundamentando-se tal escolha no facto de que

com todos os atributos disponíveis discretizados em intervalos e considerados um a um como o atributo alvo são calculadas todas as tabelas de probabilidade condicional, para ”perceber melhor os dados e assim preparar a elaboração de outros melhores modelos” [25, pág. 217].

Esta escolha não impede e é mesmo vantajosa para elaborar outros modelos de análise com base noutros algoritmos de entre os listados na taxonomia apresentada em 3.6, tendo desejavelmente em atenção a progressiva complexidade de interpretação de resultados e realizando a selecção de entre os algoritmos constantes da lista de algoritmos mais usados em prospecção de dados segundo Rexer [41], mas reduzindo o número de atributos de entrada aos identificados pelos modelos com base em Naive-Bayes como factores de influência do atributo alvo.

Assim descreveremos seguidamente os modelos exploratórios elaborados em cada uma das estruturas referidas em 3.2.6.

## 3.3 Resultados

Apresentam-se nesta secção alguns dos modelos de prospecção de dados elaborados, resultados por eles obtidos, apresentando as matrizes de confusão respectivas e *lift-charts*.

### 3.3.1 Modelo exploratório com Naive Bayes

Dada a natureza da implementação Microsoft do classificador Naive Bayes foi necessário discretizar todos os atributos contínuos e atendendo ao resultado da matriz de correlação dos atributos disponíveis, decidindo-se seleccionar o método de discretização para os atributos operacionais e deixar ao automatismo da ferramenta [27] a discretização dos meteorológicos.

Para todos os atributos operacionais excepto ”Accelerator Usage Per

100Km" e "Brake Usage Per 100Km" seleccionou-se a discretização por *Clusters* para divisão dos casos em cinco *Buckets*. No caso das duas excepções utilizou-se a técnica de *EqualAreas* devido à presença de picos pronunciados nos histogramas destes atributos e porque este método secciona os intervalo dos *Buckets* de forma a conterem iguais quantidades de casos.

A implementação Microsoft do algoritmo de agrupamento por maximização de expectativa está descrita no relatório técnico [28] e é a base da técnica de discretização por *Clusters*, que conforme descrito na documentação da ferramenta [27], selecciona aleatoriamente mil casos.

Elaboraram-se três modelos sobre a cada uma das estrutura de dados referida em 3.2.6, agregando os atributos com origem nas dimensões do armazém de dados 2.3.2 primeiro apenas com atributos operacionais, segundo apenas com atributos meteorológicos e por último com todos. Apresentam-se nas subsecções subsequentes os resultados obtidos por perspectiva de análise, condutor, rota e veículo.

A parametrização do classificador Naive-Bayes foi estabelecida como se mostra na Figura 3.9, apenas se alterando a cardinalidade de estados a considerar o que permitiu evitar a redução automática de cardinalidade para todos os atributos excepto *Time* [7]. :

[7] no formato HH:MM

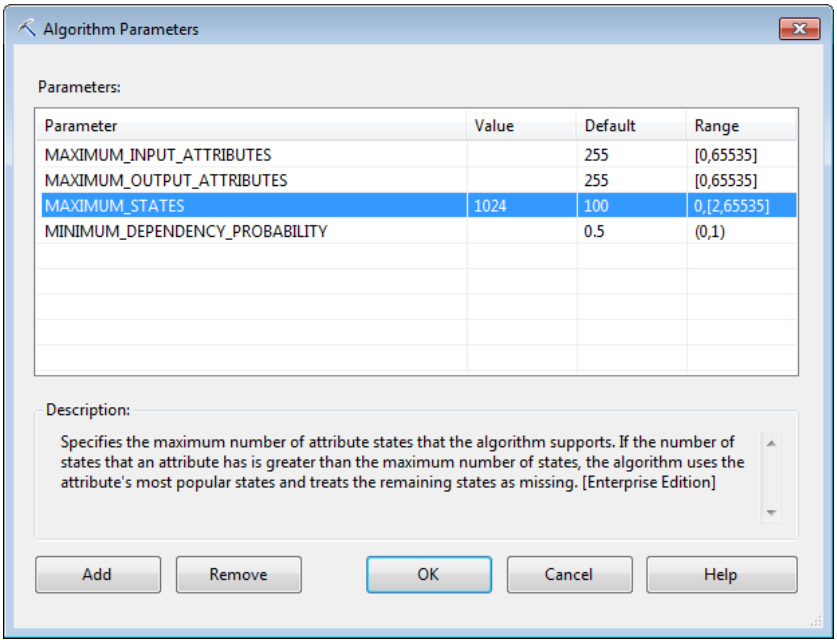



Figura 3.9: Parametrização do classificador Naive-Bayes.

De notar que a execução destes três modelos de dados apenas demora quarenta e seis segundos, como a Figura 3.10 demonstra.

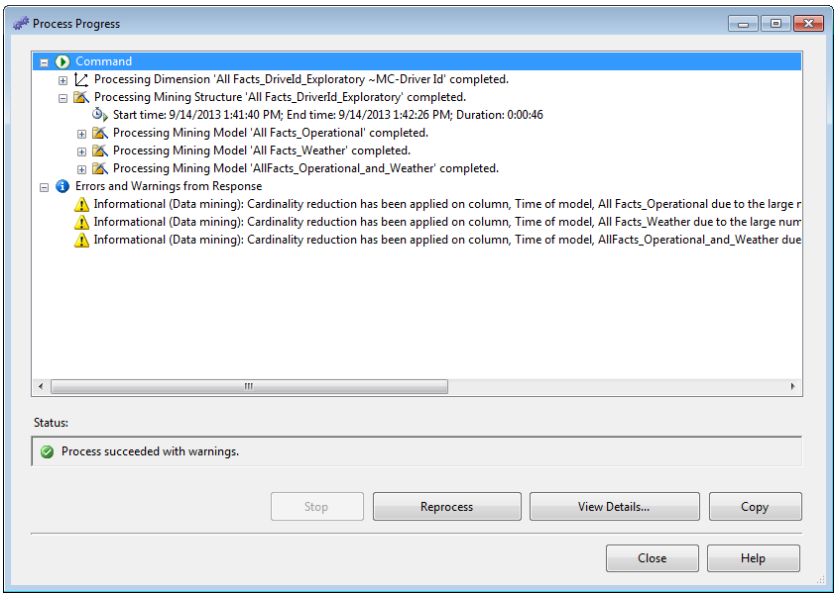



Figura 3.10: Execução do classificador Naive-Bayes.

#### 3.3.1.1 Perspectiva condutor

Nesta perspectiva de análise, procurar-se-á a identificação do conjunto de padrões de condução que caracterizam as classes de eficiência de cada um dos 1485 condutores. Tal como já se indiciava pela análise de correlação e agora se confirma por consulta à rede de dependência do modelo executado com todos os atributos, como ilustrado pela Figura 3.11, apenas existem relações entre alguns dos atributos operacionais e o atributo alvo, nomeadamente e por ordem decrescente de influência:

1. Percentagem do tempo com o motor em rotação na banda amarela;

2. Parte do dia;

3. Percentagem do tempo com o motor em rotação ao ralenti;

4. Percentagem de tempo viajado com movimento por inércia;

5. Quantidade de utilizações de embraiagem por 100Km.

Deste ponto em diante, concentraremos a análise apenas no modelo sobre atributos operacionais retirando não só os meteorológicos mas também os atributos oriundos das dimensões de análise OLAP, apenas assinalando como atributo alvo ”Average Fuel Comsumption In Litres per 100Km”, e repondo os valores por omissão dos parâmetros do classificador, e assim evitar consequências da *maldição da dimensionalidade*, que muitas vezes se manifestaram pelo disparo de excepções por falta de memória.

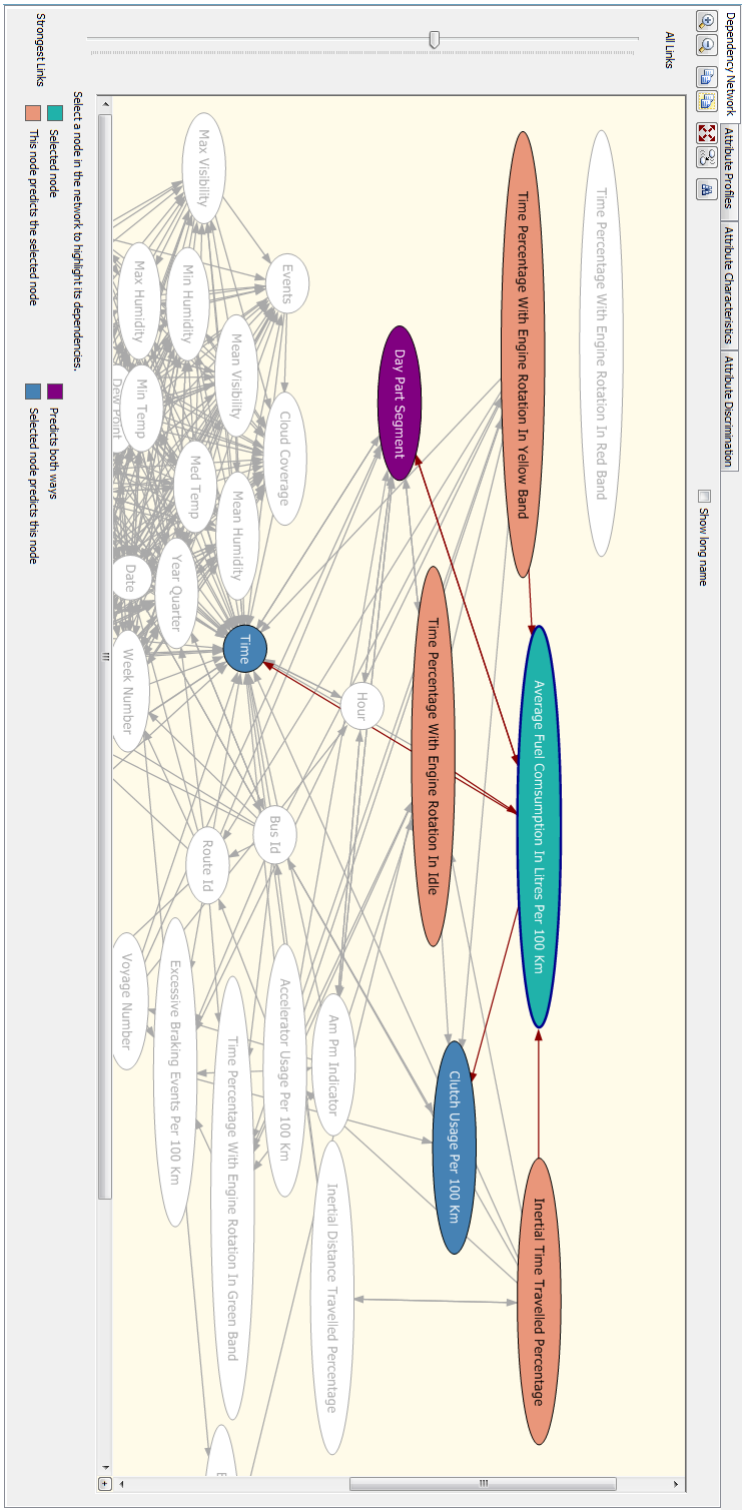



Figura 3.11: Rede de dependência de atributo alvo com classificador Naive-Bayes usando todos os atributos de entrada.

No contexto das recomendações do estudo referido em 1.2, poder-se-á utilizar a visualização de factores discriminatório para promover a mudança de hábitos de condução e o aumento de eficiência de condutores, por exemplo que actualmente realizam consumos entre 67 e 79,5 l/100Km motivando-os a atingir a classe de eficiência imediatamente superior, com consumos entre 55,7 e 67 l/100Km, como ilustrado pela Figura 3.12.

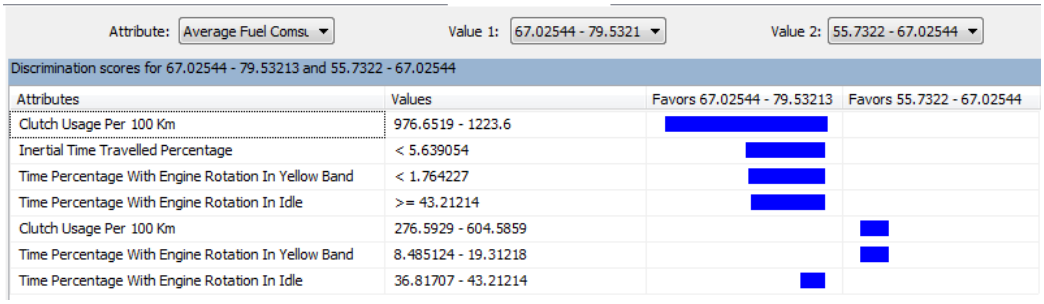



Figura 3.12: Como incentivar a melhoria da eficiência energética de um condutor.

A consulta à Figura 3.13 permite verificar qual a distribuição de cada atributo de entrada com influência no atributo alvo, por grupo de eficiência de condutores, por exemplo permitindo descobrir que os condutores mais eficientes utilizam menos vezes a embraiagem, aproveitam mais a inercia, utilizam o motor percentualmente menos tempo ao ralenti e na gama amarela de rotações.

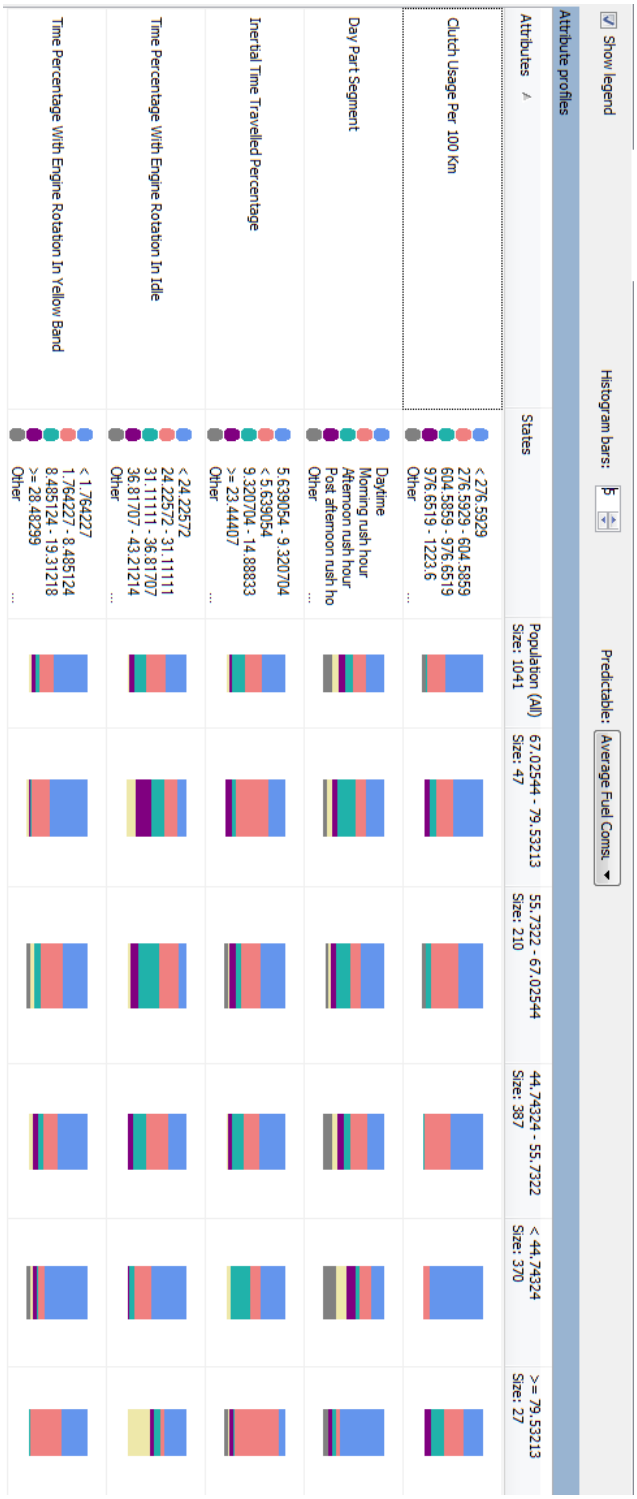



Figura 3.13: Distribuição de valores de atributos de entrada por grupo de eficiência de condutores.

É também trivial identificar quais as características dos condutores mais eficientes pela consulta das características de atributos de uma classe como ilustrado pela Figura 3.14.

Attribute: Average Fuel Consu Value: 44.74324 - 55.7322

Characteristics for 44.74324 - 55.7322

| Attributes | Values | Probability |
|---|---|---|
| Clutch Usage Per 100 Km | < 276.5929 | |
| Time Percentage With Engine Rotation In Yellow Band | < 1.764227 | |
| Clutch Usage Per 100 Km | 276.5929 - 604.5859 | |
| Inertial Time Travelled Percentage | 5.639054 - 9.320704 | |
| Time Percentage With Engine Rotation In Idle | 24.22572 - 31.11111 | |
| Time Percentage With Engine Rotation In Idle | < 24.22572 | |
| Day Part Segment | Daytime | |
| Day Part Segment | Morning rush hour | |
| Inertial Time Travelled Percentage | < 5.639054 | |
| Time Percentage With Engine Rotation In Yellow Band | 1.764227 - 8.485124 | |
| Time Percentage With Engine Rotation In Idle | 31.11111 - 36.81707 | |
| Inertial Time Travelled Percentage | 9.320704 - 14.88833 | |
| Day Part Segment | Post afternoon rush hour | |
| Day Part Segment | Afternoon rush hour | |
| Day Part Segment | Nightime | |
| Time Percentage With Engine Rotation In Yellow Band | >= 28.48299 | |
| Time Percentage With Engine Rotation In Idle | 36.81707 - 43.21214 | |
| Day Part Segment | Pre morning rush hour | |
| Time Percentage With Engine Rotation In Yellow Band | 8.485124 - 19.31218 | |
| Time Percentage With Engine Rotation In Yellow Band | 19.31218 - 28.48299 | |
| Inertial Time Travelled Percentage | >= 23.44407 | |
| Inertial Time Travelled Percentage | 14.88833 - 23.44407 | |
| Clutch Usage Per 100 Km | 604.5859 - 976.6519 | |
| Time Percentage With Engine Rotation In Idle | >= 43.21214 | |

Figura 3.14: Caracterização de uma classe de eficiência de condutores.

Atendendo aos modelos elaborados, considerando os trinta por cento de casos reservados para teste, realizou-se o teste da capacidade dos modelos preverem a classe de toda a população e do extracto de eficiência de consumo de combustível entre 21,4 e 44,7 l/100Km como se mostra nas Figuras 3.15 e 3.16:

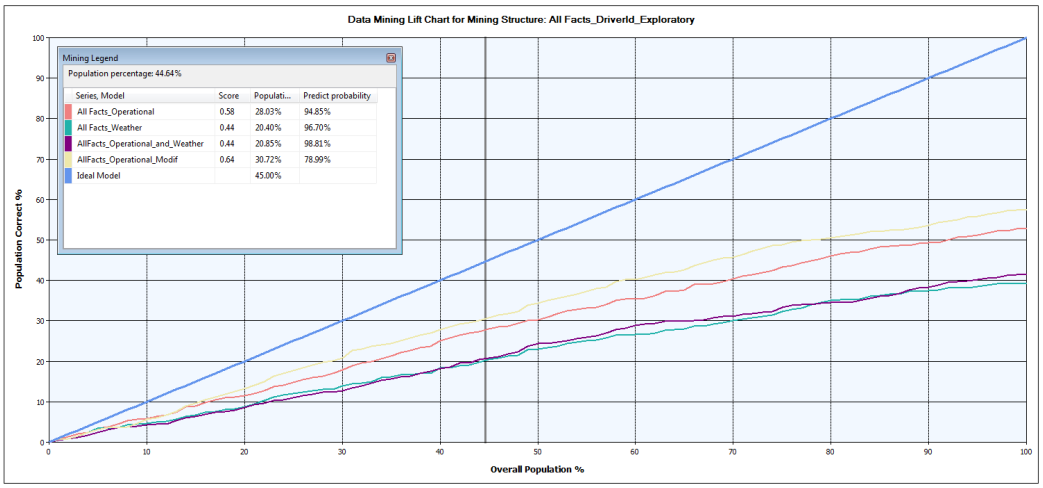



Figura 3.15: Comportamento de modelos Naive-Bayes para toda a população de teste.

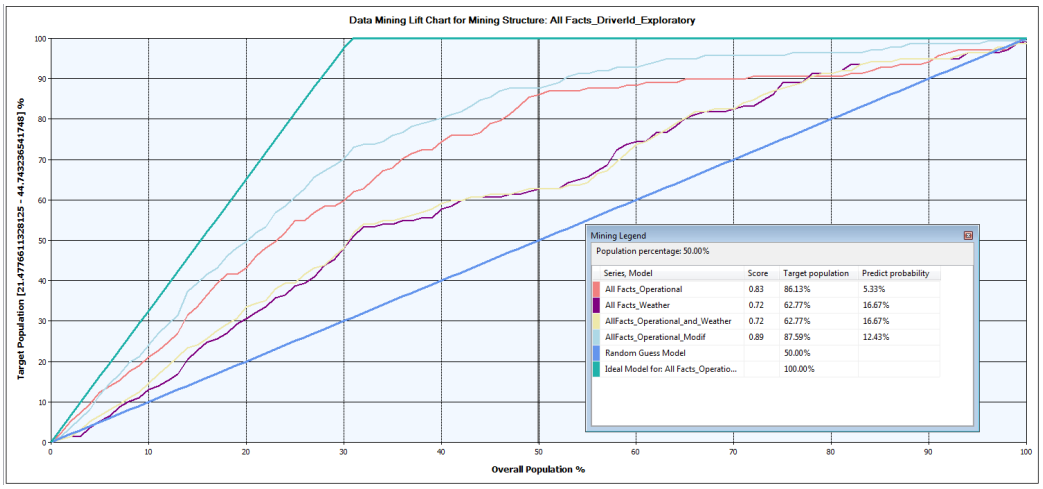



Figura 3.16: Comportamento de modelos Naive-Bayes para toda a população de teste de uma classe de eficiência.

Constata-se que o melhor modelo é sistematicamente o que apenas utiliza os atributos de entrada supra-enumerados e que influenciam o atributo alvo, sendo que a taxa de acerto de classificação se afasta mais do modelo ideal para as classes de menor eficiência. A Tabela 3.1 apresenta a matriz de confusão do modelo com melhor comportamento predictivo, mais claramente evidenciando a progressiva degradação de precisão da

| Realidade / Predição | < 44,7 | 44,7 – 55,7 | 55,7 – 67,0 | 67,0 – 79,5 | >= 79,5 |
|---|---|---|---|---|---|
| < 44,7 | 107 | 46 | 8 | 2 | 3 |
| 44,7 – 55,7 | 23 | 95 | 24 | 1 | 2 |
| 55,7 – 67,0 | 5 | 37 | 49 | 11 | 9 |
| 67,0 – 79,5 | 2 | 2 | 4 | 1 | 2 |
| >= 79,5 | 0 | 2 | 4 | 2 | 5 |

Tabela 3.1: Matriz de confusão de classificador Naive-Bayes.

classificação, sugestiva de reduzida representação de casos com maior ineficiência e de reorganização em três classes de eficiência, agrupando quatro classes em duas: o conjunto das duas classes de maior e o conjunto das duas de menor eficiência.

#### 3.3.1.2 Perspectiva rota

A análise sob a perspectiva da rota apenas apresenta dependências do atributo alvo com média força de ligação ao atributo meteorológico ”Eventos”, conforme ilustrado pela Figura 3.17:

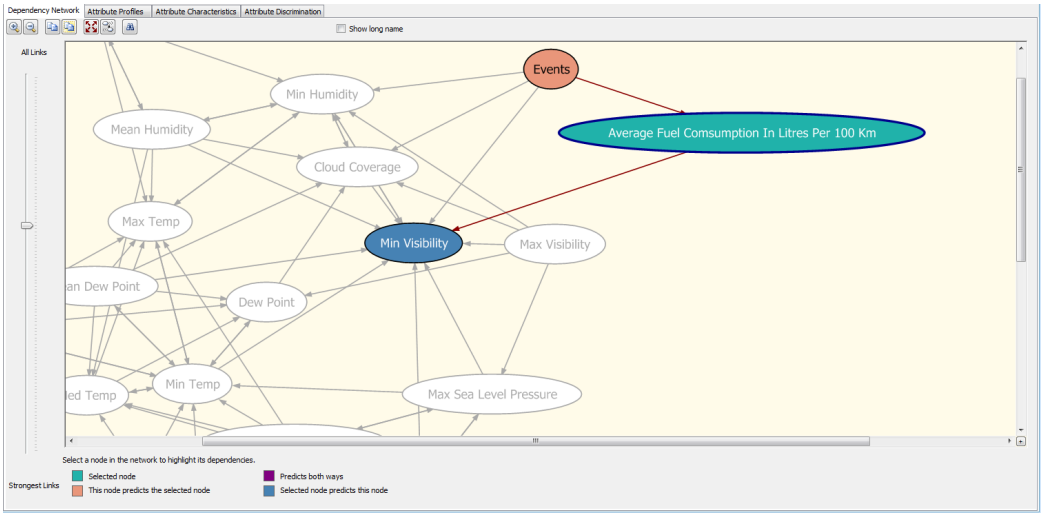



Figura 3.17: Rede de dependência do atributo alvo da perspectiva Rota.

Com ressalva pelo impacto da concentração de casos num número restrito de rotas identificada em 3.2.6, adensam-se ainda assim, as suspeitas da necessidade de escrutinar algumas rotas quanto à qualidade

do escoamento de águas pluviais porquanto se detectam nos dados disponíveis a relação causa efeito de alguns eventos conforme se ilustra na Figura 3.18:

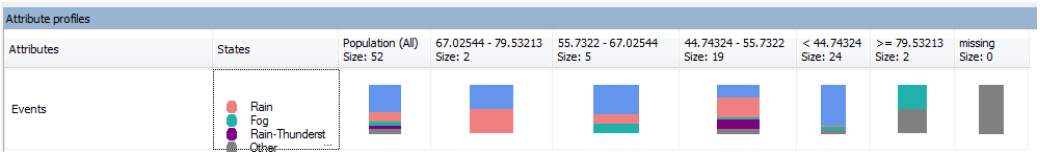



Figura 3.18: Dependência do atributo alvo na perspectiva Rota de Eventos.

Também por causa da concentração de casos em poucas rotas, salientando a necessidade de mais dados segundo esta perspectiva, considerou-se desadequado prosseguir numa ”análise” que face aos dados seria especulativa.

#### 3.3.1.3 Perspectiva veículo

A análise sob a perspectiva do veículo apenas apresenta uma fraca dependências do atributo alvo com o atributo operacional ”Accelerator Usage Per 100Km”, conforme ilustrado pela Figura 3.19:

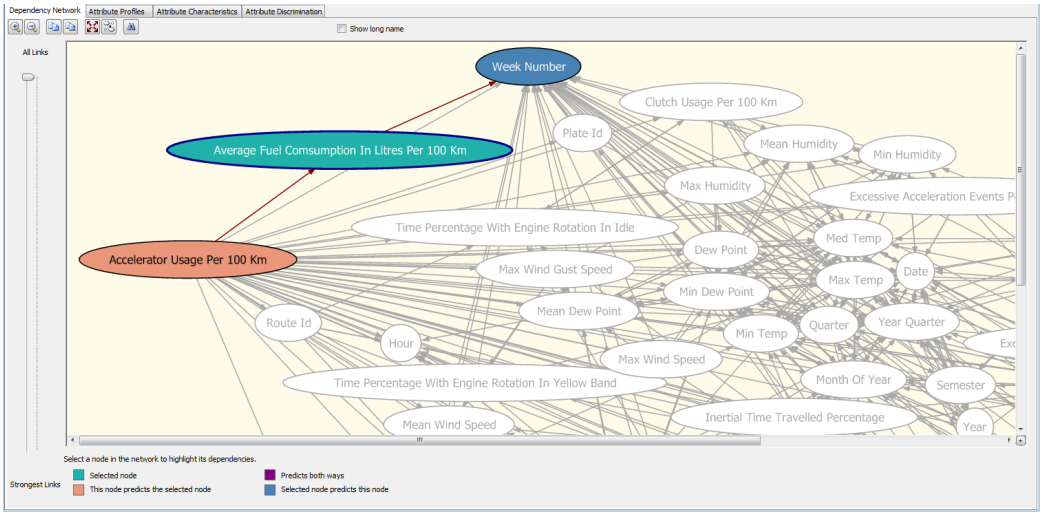



Figura 3.19: Rede de dependência do atributo alvo da perspectiva veículo.

Considerou-se inadequado prosseguir com a análise pelas mesmas razões anteriormente apontadas para a perspectiva rota.

# Capítulo 4

# Conclusões

Neste capítulo conclui-se a elaboração do presente relatório confrontando o trabalho realizado contra os objectivo enunciado em 1.3, elencando-se as ideias a reter deste trabalho e concluindo com aspectos passiveis de desenvolvimento futuro.

## 4.1 Trabalho realizado versus objectivos

Recuperando a definição de objectivos enunciados em 1.3, constata-se que estes se agrupam quanto à sua natureza em dois domínios:

1. dos sistemas de informação para apoio à decisão, e;
2. dos sistemas de informação para prospecção de conhecimento em bases de dados.

De uma forma simplista o objectivo dos sistemas de decisão é permitir aos especialistas ”que observam as engrenagens dos processos de negócio” [13] a procura de respostas para questões do tipo Quem? Quando? Onde?, enquanto que os sistemas de prospecção procuram os padrões embebidos nos dados e que permitem aos mesmos actores explicar porquê.

A elaboração deste trabalho, aderindo de perto à metodologia seleccionada como a estrutura do presente documento evidencia, apoiando-se

”aos ombros de gigantes” respondeu efectivamente aos objectivos propostos como fica demonstrado na descrição produzida nos capítulos 2 e 3, respectivamente relativos aos dois domínios supra-mencionados

Merece particular destaque o facto de se ter conseguido chegar a um modelo que explica a eficiência energética com base no comportamento dos condutores 3.3.1.1, com um ponto de partida ”esmagador” [24] com cerca de um milhão e meio de registos com quarenta e quatro colunas 2.2.1.5 , passando pela capacidade disponibilizada pelo armazém de dados descrito em 2 em responder à análise *slice & dice* por cada dimensão cuja análise era requerida e como é típico em OLAP.

Fica também demonstrado em 3.3.1.1 ser possível de usar o conhecimento assim extraído para continuamente, o que é uma ferramenta útil para maximizar a retenção das práticas de condução ecologicamente responsáveis conforme sugerido no estudo [8, pág. 37].

A natureza cíclica da metodologia seleccionada, levou a incontáveis iterações que em espiral progressiva convergiu, pelo menos assimptóticamente, com o atingir de objectivos, mas simultaneamente levantando novas questões, situação típica de todas as actividades de descoberta de conhecimento que muitas vezes envolvem uma quantidade de acaso e felicidade, potenciada pela persistência e conhecimento dos princípios subjacentes aos algoritmos da ferramenta, adquirido por estudo da bibliografia.

Deixam-se pois de seguida sugestões de evolução do presente trabalho, optimistas quanto à capacidade da base estabelecida permitir a adaptação a outros cenários de análise, seja de tipo OLAP ou prospecção de conhecimento, por sucessivas iterações do ciclo CRISP-DM.

## 4.2 Trabalho futuro

Com a preocupação de evoluir na complexidade dos modelos construídos pelos diversos tipos de algoritmos, sugerimos a elaboração de modelos

preditivos com base em árvores de decisão, agrupamento e por último redes neuronais artificiais perceptrão multi-camada pois estas com ”duas camadas intermédias conseguem a aproximação de qualquer função” [22, pág. 136, 230] o que permitirá explorar padrões não separáveis por hiperplanos.

Considera-se que a elaboração de uma estrutura de dados que redistribua e reescale atributos de entrada é uma actividade desejável na preparação para aplicação de modelos de agrupamento baseados em distâncias para evitar a distorção do espaço introduzida pelas diferentes escalas dos atributos.

Continuar a iterar o ciclo CRISP-DM procurando a optimização dos parâmetros dos algoritmos disponibilizados pela ferramenta escolhida e utilizar o processo descrito por Crivat [30] para avaliar o impacto de mais dados de treino sobre a precisão dos modelos preditivos.

Considerar o cruzamento de dados com outros atributos, por exemplo condições de tráfego rodoviário.

Finalmente, o pragmatismo ditou em muitos momentos que se tomassem opções favorecendo o atingir de objectivos e deixando abertas avenidas de descoberta que são quase ilimitadas como é inerente dos processos de descoberta e fica patente no presente relatório.

# Apêndice A

# Script SQL para criação de DW

Favor consultar DVD do projecto directoria de anexos para consultar a listagem referida abrindo o ficheiro ”createDataWarehouseTECMIC.sql”.

# Apêndice B

# Vista SQL sobre dados fonte

Favor consultar DVD do projecto directoria de anexos para consultar a listagem referida abrindo o ficheiro ”View1.sql”.

# Apêndice C

# Script C# obter dados meteorologia

Favor consultar DVD do projecto directoria de anexos para consultar a listagem referida abrindo o ficheiro ”main.cs”.

# Apêndice D

# Resultados análise de perfil de dados

Favor consultar DVD do projecto directoria de anexos para consultar a listagem referida abrindo o ficheiro ”XTPassenger_Initial_Profiling.xml” usando o Microsoft Data Profile Viewer que é disponibilizado pelo Microsoft SQL Server 2008R2.

# Apêndice E

# Membros calculados cubo OLAP

Favor consultar DVD do projecto directoria de anexos para consultar a listagem referida abrindo o ficheiro ”CalculationsScript.mdx”.

# Apêndice F

# script Gnuplot

Favor consultar DVD do projecto directoria de anexos para consultar a listagem referida abrindo o ficheiro ”Mydata.gnu”.

# Apêndice G

# Fonte de dados para prospecção

Favor consultar DVD do projecto directoria de anexos para consultar a listagem referida abrindo o ficheiro ”AllFacts.sql”.

# Apêndice H

# Análise dados univariados

Favor consultar DVD do projecto directoria de anexos para consultar a listagem referida abrindo o ficheiro ”DadosUnivariados.pdf”.

# Apêndice I

# Backups de Bases de Dados

Favor consultar DVD do projecto directoria de anexos para consultar a listagem referida e restaurar o ficheiro de Backup pretendido.

# Apêndice J

# Solução Visual Studio

Favor consultar DVD do projecto directoria de anexos para consultar a listagem referida abrir o ficheiro ”TECMIC_Datastaging_XTPassenger.sln”.

# Apêndice K

# Cadeia mail envio proposta trabalho

**Subject:** RE: MEIC, aluno 14937, Submissão de proposta de trabalho de projeto
**From:** "Joao Ferreira" <jferreira@deetc.isel.ipl.pt>
**Date:** 10/16/2012 11:43 AM
**To:** <wv@deetc.isel.pt>, 'José Almeida' <jadcda@gmail.com>

Bom Dia,

Para informar que estou disponível para a orientação do projeto proposto.
Cumpts
JFerreira

-----Mensagem original-----
De: Walter Vieira [mailto:wv@deetc.isel.pt]
Enviada: terça-feira, 16 de Outubro de 2012 11:09
Para: 'José Almeida'
Cc: jferreira@deetc.isel.ipl.pt
Assunto: RE: MEIC, aluno 14937, Submissão de proposta de trabalho de projeto

Bom dia,

Acuso a recepção da sua proposta de projecto.
Dado tratar-se da continuação do projecto do ano anterior e sendo a decisão final de aceitação ou não da competência da CCMEIC, importa, no entanto, saber se o orientador está disponível para continuar a orientar o trabalho.

Cumprimentos,

Walter Vieira

-----Original Message-----
From: José Almeida [mailto:jadcda@gmail.com]
Sent: terça-feira, 16 de Outubro de 2012 10:54
To: Walter Vieira
Cc: Joao Ferreira
Subject: MEIC, aluno 14937, Submissão de proposta de trabalho de projeto

Ex.mo Prof Walter Viera,

Conforme indicado no moodle, página de informações para o ano letivo em curso, sou a submeter à V. consideração a proposta de trabalho de projeto apensa.

Com os melhores cumprimentos e saudações académicas,

--
José de Almeida
Aluno 14937

# Resumo da Proposta de Ideia Para Dissertação de Natureza Científica ou Trabalho de Projecto

**Código:** Mxxxx *(a definir posteriormente pela comissão de mestrado)*
**Designação:** Análise de eficiência energética em frota de transportes

**Orientador(es):** Prof. Dr. João Ferreira
**Contacto do orientador:** jferreira@deetc.isel.ipl.pt
**Local de contacto:** Gabinete na ADEETC

*(preenchimento opcional)* *remover o **X** que não interessa* *(Mestrado(s) onde é oferecida)*

| | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Dissertação | | Trabalho de Projecto | X | | MEIC | X | MEET | | MERCM | | |

**Resumo:** Pretende-se explorar com recursos a técnicas OLAM[1] os dados existentes numa base de dados de uma frota de transportes obtidos pela extração de dados por CAN BUS durante os percursos efetuados. Esta analise têm como objetivo de analisar a eficiência energética da operação de viaturas de uma frota de autocarros por:

- tipo de veículo;
- condutor;
- rota;
- data, dia e hora, e;
- condições meteorológicas.

Pretende-se facilitar a exploração de dados do tipo *Slice & Dice* clássica de sistemas OLAP, recorrendo depois a técnicas de Data Mining para por exemplo, detectar padrões de influência na eficiência de exploração, agrupar veículos ou condutores quanto à sua eficiência, e estudando a influência das condições meteorológicas e dos percursos feitos.

Do ponto de vista da tecnologia base, a plataforma de desenvolvimento do projecto basear-se-a na solução de arquitectura proporcionada pela plataforma de *Business Intelligence* da Microsoft sobre SQL Server 2008R2 com SSIS, a explorar em Visual Studio 2008, instalada numa máquina virtual correndo Windows 7 Professional.

Para desenvolver este trabalho de projecto está disponível por coordenação do Prof. Dr. João Ferreira uma parceria com uma empresa de soluções embarcadas de de frotas, a TECMIC.

O Aluno: José António Dias Correia de Almeida, número 14937.

**Constrições:**

1 on-line analytical processing (OLAP) com data mining

# Bibliografia


[1] Fayyad, U. M. et al,
*Advances in knowledge discovery and data mining*.
AAAI Press / The MIT Press,
1996.

[2] Usama Fayyad, Gregory Piatetsky-Shapiro, and Padhraic Smyth,
*From Data Mining to Knowledge Discovery in Databases*
AI Magazine Volume 17 Number 3 (1996),
(©AAAI)

[3] A. Azevedo, M. F. Santos,
*KDD, SEMMA AND CRISP-DM: A PARALLEL OVERVIEW*.
IADIS European Conference Data Mining,
2008.

[4] SAS Enterprise Miner – SEMMA.
SAS Institute.
Acedido pelo endereço http://www.sas.com/technologies/analytics/datamining/miner/semma.html em Dezembro de 2012.

[5] Chapman, P. et al,
*CRISP-DM 1.0 - Step-by-step data mining guide*.
Acedido pelo endereço http://www.crisp-dm.org/CRISPWP-0800.pdf em Janeiro de 2013

[6] Norma can BUS
*ISO 11898*.

Acedido pelo endereço http://www.iso.org/iso/iso_catalogue/catalogue_tc/catalogue_detail.htm?csnumber=33422 em Julho de 2013.

[7] *Site de acesso gratuito a repositório histórico de dados de meteorologia*.
Acedido pelo endereço http://www.wunderground.com/history... em Fevereiro de 2013

[8] Kazunori Kojima and Lisa Ryan,
*Transpot Energy Efficiency Information Paper*
Implementation of IEA Recommendations since 2009 and next steps
©OECD/IEA, September, 2010.

[9] Birol, F. et al,
*World Energy Outlook 2010*
ISBN 978 92 64 08624 1
©OECD/IEA, 2010.

[10] TECMIC, XTraN, Gestão de Frotas
Acedido pelo endereço http://www.tecmic.pt/por/xtran/xtran_intro.html em Março de 2013.

[11] Jiawei Han,
*OLAP Mining: An Integration of OLAP with Data Mining*,
In Proceedings of the 7th IFIP 2.6 Working Conference on Database Semantics (DS-7)
1997, páginas 1–9

[12] Microsft *SQL Server Integration Services*
Acedido pelo endereço http://msdn.microsoft.com/en-us/library/ms141026(v=sql.105).aspx em Março de 2013.

[13] Ralph Kimball, Margy Ross,
*The data warehouse toolkit : the complete guide to dimensional modeling — 2nd ed*
Wiley Computer Publishing, ISBN 0-471-20024-7

[14] Ralph Kimball, Joe Caserta,
*The data warehouse ETL toolkit : practical techniques for extracting, cleaning, conforming,*
*and delivering data*
Wiley Publishing, Inc., ISBN 0-7645-7923-1

[15] W. H. Inmon,
*Building the Data Warehouse, Fourth Edition*
Wiley Publishing, Inc., ISBN 0-7645-9944-5

[16] Platão,
*Diálogo de Sócrates com Theaetetus*
Acedido pelo endereço http://www.gutenberg.org/files/1726/1726-h/1726-h.htm em Junho de 2013.

[17] Thomas Bayes
*An Essay towards solving a Problem in the Doctrine of Chances*
1763, Philosophical Transactions of the Royal Society of London 53 (1763), 370–418.

[18] Edmund Gettier,
*Is Justified True Belief Knowledge?*,
1967, Analysis. vol. 23 ( 966). Copyright @ by Edmund Gettier.

[19] SHANNON, C. E.
*A Mathematical Theory of Communication*
The Bell System Technical Journal,
Vol. 27, pp. 379–423, 623–656, July, October, 1948.

[20] Tom Mitchell,
*Machine Learning*
McGraw-Hill 1996, ISBN 0070428077

[21] Stuart Russell, Peter Norvig
*Artificial Intelligence: A Modern approach, 1st edition*
Pearson Education, Inc., ISBN 0-13-103805-2

[22] J. Gama, A. Carvalho, K. Faceli, A. Lorena, M. Oliveira,
*Extração de Conhecimento de Dados*,
2012, Edições Sílabo, ISBN 978-972-618-698-4

[23] Russell Ackoff,
*From Data to Wisdom*,
Journal of Applies Systems Analysis, Volume 16, 1989 p 3-9.

[24] Ian H. Witten, Frank Eibe, Mark A. Hall.,
*Data mining : practical machine learning tools and techniques.—3rd ed.*
2011, The Morgan Kaufmann series in data management systems, ISBN 978-0-12-374856-0

[25] Jamie MacLennan, Bogdan Crivat, ZhaoHui Tang,
*Data mining with Microsoft SQL server 2008*
2009, ISBN 978-0-470-27774-4

[26] Microsoft
*Implementação SQL Server de algoritmos de Data Mining*
Acedido pelo endereço http://technet.microsoft.com/en-us/library/ms175595(v=sql.105).aspx em Janeiro de 2013.

[27] Microsoft
*Discretização de dados em SQL Server*
Acedido por http://technet.microsoft.com/en-us/library/ms174512(v=sql.105).aspx em Janeiro de 2013.

[28] Paul S. Bradley Usama M. Fayyad Cory A. Reina,

*Scaling EM (Expectation-Maximization) Clustering to Large Databases* ,
Microsoft Research , November 1998 , Revised October 1999,
Technical Report MSR-TR-98-35, Microsoft Research, Microsoft Corporation.

[29] Microsoft
*Selecção automática de atributos*
Acedido por http://technet.microsoft.com/en-us/library/ms175382(v=sql.105).aspx em Janeiro de 2013.

[30] Crivat, B.,
*How much training data is enough?*,
Acedido por http://www.bogdancrivat.net/dm/archives/28#more-28 em Agosto de 2013.

[31] Pearson, Karl
*On lines and planes of closest fit to systems of points in space*
Philosophical Magazine Series 6, 1901, Vol.2(11), p.559-572
Taylor & Francis Group

[32] Geng, Liqiang and Hamilton, Howard J.,
*Interestingness measures for data mining: A survey*,
ACM Comput. Surv.,2006, volume = 38, number = 3, Acedido via http://doi.acm.org/10.1145/1132960.1132963 em Setembro 2013.

[33] Rigolli, M., Brady, M.,
*Towards a Behavioural Traffic Monitoring System*,
International Conference on Autonomous Agents,
Proceedings of the 4th International Joint Conference on
Autonomous Agents and Multiagent Systems, pp. 449-454, 2005.

[34] Ishibashi, M., Okuwa, M., Doi, S., Akamatsu, M.,
*Indices for Characterizing Driving Style and their Relevance to Car Following Behavior*,
SICE Annual Conf., pp. 1132-1137, 2007.

[35] O. Taubman-Ben-Ari, M. Mikulincer and O. Gillath,
*The multidimensional driving style inventory-scale construct and validation*,
Accident Analysis and Prevention, Vol. 36, pp. 323-332, 2004

[36] Hattori, Hiromitsu, Nakajima, Yuu and Ishida, Toru,
*Agent Modeling with Individual Human Behaviors*,
Proc. of 8th Int'l. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2009), pp. 1369-1470, 2009.

[37] Augustynowicz, A.,
*Preliminary Classification of Driving Style with Objective Rank method*,
International Journal of Automotive Technology, Vol. 10, No. 5, pp. 607-610, 2009.

[38] Chan, M., Herrera, A. and Andre, B.
*Detection of changes in driving behaviour using unsupervised learning*,
IEEE International Conference on Humans, Information and Technology, 1994, Vol. 2, pp. 1979–1982.

[39] Almeida, J.; Ferreira, J.,
*BUS Public Transportation System Fuel Efficiency Patterns*,
in proceedings of the 2nd International Conference on Machine Learning and Computer Science (IMLCS'2013) 24-25 August, Malaysia.

[40] Reiter, U.
*Modeling the driving behaviour influenced by information technologies. In Highway Capacity and Level of Service*,
(Ed.Brannolte), 1991, pp. 309–320 (Balkema, Rotterdam).

[41] Rexer, K.
*4th Annual Data Miner Survey 2010 Survey Summary Report*,
For more information contact Karl Rexer, PhD, krexer@RexerAnalytics.com, www.RexerAnalytics.com