Universidade do Minho

Escola de Engenharia

Departamento de Informática

# Determinação de Frequência de Recolhas aplicando Técnicas de Mineração de Dados

**Mónica Patrícia Freire do Vale**

Dissertação de Mestrado

2008

# Determinação de Frequência de Recolhas aplicando Técnicas de Mineração de Dados

**Mónica Patrícia Freire do Vale**

Dissertação apresentada à Universidade do Minho para obtenção do grau de Mestre em Informática, elaborada sob orientação do Professor Doutor Orlando Manuel de Oliveira Belo.

2008

*Ao Nuno.*

# Agradecimentos

Começo por agradecer a toda a minha família, pelo apoio desde sempre, em especial aos mais pequenos pela inspiração.

Agradeço também a todos os meus colegas de trabalho, em especial ao Eng.º Cândido Martins, pela partilha de conhecimento e por todo o apoio prestado. Agradeço também à Cachapuz, pelo financiamento deste projecto e pela disponibilização de toda a informação necessária.

Agradeço à Resulima pela disponibilização da informação utilizada na realização desta dissertação.

Agradeço ao Professor Doutor Orlando Belo, por toda a ajuda prestada na elaboração deste trabalho e pela compreensão.

Finalmente, mas não em último, ao Nuno, por tudo.

# Resumo


## Determinação de Frequência de Recolhas aplicando Técnicas de Mineração de Dados

A separação de resíduos é, nos dias de hoje, uma preocupação das populações e das empresas, pela necessidade de reduzir o consumo de recursos naturais do planeta em que vivemos. Essa preocupação gerou um aumento da produção de resíduos recicláveis e aumentou as dificuldades das empresas que efectuam a recolha desses resíduos. Estas empresas aperceberam-se da necessidade de se munirem de ferramentas de automatização dos processos, ferramentas essas que lhes permitem registar e armazenar grandes quantidades de informação, úteis para a gestão do negócio. No entanto, devido ao facto de a quantidade de informação existente ser bastante elevada, as ferramentas tradicionais não têm a capacidade para responder às questões dos gestores.

A utilização de ferramentas de mineração de dados poderá facilitar o acesso à informação e permitir extrair conhecimento dessa informação, identificando relações insuspeitas entre os dados que poderão ser um factor diferenciador na tomada de decisão [Hand et al., 2001]. No caso estudado nesta dissertação irão ser aplicados mecanismos de mineração de dados para determinar a frequência "ideal" de recolha das rotas, analisando os dados históricos quer das quantidades recolhidas, quer dos enchimentos dos contentores registados.



**Palavras-Chave:** Recolha selectiva, mineração de dados, previsão, optimização

# Abstract


## Determination of the Collection Frequency through the application of Data Mining Techniques

The waste separation is, nowadays, a concern of populations and enterprises, because of the need to reduce the use of natural resources of the planet we live on. This concern has increased the production of recyclable waste and the difficulties of the companies that collect these products. These companies have recognized the need to acquire tools that provide the automation of their processes, which allow them to register and store large amounts of information, useful for business management. However, due to the fact that the existing information is considerably large, the traditional tools don't have the potential to answer to the manager's questions.

The use of data mining tools may assist the access to the information and allow the knowledge discovery within the data, identifying unsuspicious relationships between the data that may be a determinant aspect on decision making [Hand et al., 2001]. In the case addressed in this dissertation, data mining mechanisms will be used to determine de "ideal" collection frequency of the routes, through the analysis of historical data of the quantities collected, as well as the filling levels of the registered bins.



**Keywords:** Selective gathering, data mining, prediction, optimization

# Índice

**Introdução ..... 1**

1.1 O processo de Recolha Selectiva ..... 1

1.2 Casos de aplicação da Mineração de Dados ..... 3

1.3 Motivação e objectivos ..... 6

1.3.1 Motivação ..... 6

1.3.2 Objectivos ..... 8

1.4 Organização da dissertação ..... 8

**A Recolha Selectiva ..... 11**

2.1 Modo de operação ..... 11

2.2 Análise de uma ferramenta de apoio à Recolha Selectiva ..... 13

2.2.1 SPAR - Sistema de Planeamento e Análise da Recolha ..... 14

2.2.2 Dados armazenados ..... 19

2.2.3 Limitações para a tomada de decisão ..... 23

2.2.4 Áreas de aplicação ..... 26

**Análise de alguns mecanismos de Mineração de Dados ..... 31**

3.1 Mecanismos de Mineração de Dados ..... 31

3.1.1 Classificação ..... 33
3.1.2 Segmentação (*Clustering*) ..... 34
3.1.3 Estimativa ..... 34
3.1.4 Previsão ..... 34
3.1.5 Regras de associação ..... 35
3.1.6 Séries temporais ..... 35
3.2 Áreas de aplicação da Mineração de Dados ..... 35
3.2.1 Transacções comerciais ..... 36
3.2.2 Dados de negócio e comércio electrónico ..... 37
3.2.3 Meteorologia ..... 38
3.2.4 Simulação ..... 38
3.2.5 Assistência médica ..... 39
3.2.6 Web ..... 39

**Estudo de casos de aplicação da Mineração de Dados ..... 41**

4.1 Estudo do *Vehicle Routing Problem* e variantes ..... 41
4.1.1 Abordagem à resolução de *Vehicle Routing Problem with Stochastic Demands and Time Windows* ..... 46
4.1.2 Abordagem à resolução de *Vehicle Routing Problem with Time Windows* ..... 47
4.2 Estudo da Previsão da Procura (*Demand Forecasting*) ..... 48
4.2.1 Abordagem à previsão da procura de papel ..... 49
4.3 O caso da Recolha Selectiva ..... 49

**Aplicação de Mineração de Dados na Recolha Selectiva ..... 51**

5.1 Casos de estudo ..... 51
5.2 Desenvolvimento ..... 52
5.2.1 Optimização da frequência de recolhas com base em séries temporais ..... 54
5.2.2 Optimização da frequência de recolhas com base em *clustering* ..... 63

5.2.3 Análise crítica da aplicação de mineração de dados na recolha selectiva..... 72

**Conclusões e Trabalho Futuro ..... 75**

6.1 Conclusões..... 75

6.2 Trabalho futuro..... 76

**Bibliografia ..... 79**

**Referências WWW..... 83**

# Índice de Figuras

Figura 1: Fases da metodologia CRISP-DM (adaptado de [WWW CRISPDM]) ..... 6
Figura 2: Esquema da Base de Dados do SPAR ..... 20
Figura 3: Exemplo de *Vehicle Routing Problem* ..... 42
Figura 4: Exemplo de *Vehicle routing problem with Time Windows* ..... 43
Figura 5: Exemplo de *Distance-Constrained Vehicle Routing Problem* ..... 43
Figura 6: Exemplo de *Vehicle Routing Problem with Backhauls* ..... 44
Figura 7: Exemplo de *Vehicle Routing Problem with Pickup and Delivery* ..... 45
Figura 8: Exemplo de *Vehicle Routing Problem with Stochastic Demands and Time Windows* ..... 46
Figura 9: Exemplo de resolução de *Vehicle Routing Problem with Time Windows* em três fases ..... 47
Figura 10: Exemplo de mecanismo de séries temporais ..... 49
Figura 11: Dados da vista vRotasPesos ..... 56
Figura 12: Estrutura da tabela M_PesosRotasSemana ..... 57
Figura 13: Dados da tabela M_PesosRotasSemana ..... 58
Figura 14: Configuração dos dados de treino para o modelo de séries temporais ..... 59
Figura 15: Configuração dos parâmetros do algoritmo de séries temporais ..... 60
Figura 16: Resultados obtidos com o modelo de séries temporais ..... 61
Figura 17: Exemplo de previsões do modelo de séries temporais próximas das reais ..... 62
Figura 18: Exemplo de previsões do modelo de séries temporais distantes das reais ..... 62

Figura 19: Dados da vista vEnchimentos ..... 65
Figura 20: Dados da vista vContentores ..... 65
Figura 21: Configuração dos dados de treino para o modelo de *clustering* ..... 68
Figura 22: Resultados obtidos com o modelo de *clustering* ..... 69
Figura 23: Elementos de um cluster do modelo de *clustering* ..... 70
Figura 24: Resultado da renomeação dos *clusters* ..... 70

# Índice de Tabelas

Tabela 1: Número de ecopontos associados a várias empresas de recolha no território português 24
Tabela 2: Exemplo de calendarização de rotas ..... 25
Tabela 3: Algoritmos disponíveis na ferramenta Microsoft SQL Server 2005 Analysis Services..... 53
Tabela 4: Sugestão de organização de rotas pelo modelo de *clustering*..... 71

# Índice de Fórmulas

Fórmula 1: Cálculo da taxa de enchimento de um contentor..... 21
Fórmula 2: Cálculo do peso por contentor (estimado)..... 23

# Capítulo 1

# Introdução

## 1.1 O processo de Recolha Selectiva

Nos dias que correm a separação de resíduos para posterior reciclagem e reutilização tornou-se numa tarefa de extrema importância. Há muito que as populações e as empresas, em especial nos países mais desenvolvidos, começaram a preocupar-se com o consumo dos recursos naturais, alertados para as consequências da diminuição desses recursos e para os efeitos causados pelo seu depósito depois de ultrapassada a sua utilidade.

Em 1992 o professor e ecologista William Rees [WWW WilliamRees] criou o conceito de Pegada Ecológica [WWW EcologicalFootprint], uma medida que calcula a quantidade de terra e água necessárias para sustentar uma população em termos dos recursos materiais e energéticos consumidos e dos recursos necessários para absorver os resíduos produzidos [WWW GlobalFootprint]. Na sociedade actual é difícil reduzir a dependência material e energética das

populações, a redução da quantidade de recursos naturais utilizados para a sua produção só pode ser feita recorrendo à reciclagem.

A separação de resíduos é feita, numa primeira fase, pelo cidadão comum, que separa os resíduos que produz por categorias e os deposita em recipientes apropriados. Em Portugal o local de deposição de resíduos recicláveis designa-se de ecoponto e consiste de um ou mais recipientes, sendo mais comuns os seguintes [WWW PontoVerde]:

- O papelão, de cor azul, para depósito de jornais, revistas, papel de escrita ou de embrulho, sacos de papel e embalagens de cartão.
- O embalão, de cor amarela, para depósito de embalagens de plástico (de produtos que não sejam tóxicos nem perigosos), sacos de plástico, pacotes de leite e bebidas e embalagens de metal.
- O vidrão, de cor verde, para depósito de garrafas, garrafões, frascos e boiões de vidro.

Devido ao sucesso conseguido na mobilização das populações para a separação dos resíduos e às vantagens da reciclagem, algumas entidades começam a apostar na reciclagem de novos produtos. É o caso dos óleos alimentares usados, que podem ser reutilizados para a produção de biodiesel, um combustível renovável e biodegradável que pode abastecer algumas viaturas [WWW Biodiesel].

O número de Ecopontos distribuídos pelo território nacional tem vindo a aumentar, sendo que em 2006 eram já cerca de 25000 equipamentos. O aumento do número destes equipamentos, em conjunto com as campanhas de sensibilização das populações lançadas pelas entidades e organismos nacionais, provocam naturalmente um aumento das quantidades de resíduos recicláveis a recolher [WWW PontoVerde]. Para as empresas que efectuam a recolha desses resíduos as dificuldades são acrescidas: mais pontos de recolha (em grande número nas áreas urbanas e em menor número nas áreas rurais) e mais produto a recolher.
Outra dificuldade é o facto de a produção de resíduos recicláveis não ser homogénea. A estação do ano influencia muito a produção de resíduos em geral, e a de resíduos recicláveis em particular,

nomeadamente devido à movimentação das populações no período de verão para as zonas costeiras do território português. Outras condicionantes, como o estado do tempo, a existência de festividades ou eventos, entre outros, podem ter um impacto significativo. Também a quantidade produzida de cada um dos produtos (papel, embalagem ou vidro) é muito desigual de produto para produto e é fortemente influenciada por estas e outras condicionantes, por exemplo, a proximidade de escolas (onde poderá ser produzida mais quantidade de papel) ou de restaurantes e cafés (onde poderá ser produzida mais quantidade de vidro).

Estas empresas necessitam, por isso, de ferramentas que lhes permitam automatizar os processos de recolha destes resíduos, registar a informação associada em suporte digital e, posteriormente, analisar esta informação e fazer dela uma base para decisão. Como é evidente, quanto maior e mais rica for a informação registada melhor será a base para uma correcta tomada de decisão, no entanto, também maiores são as dificuldades para analisar toda essa informação.

Os mecanismos tradicionais de extracção de informação (sob a forma de relatórios) não são suficientes para permitir analisar e retirar conclusões em tempo útil, pelo que se torna necessária a utilização de outras ferramentas que permitam tirar partido dessa informação, no sentido de optimizar os processos.

## 1.2 Casos de aplicação da Mineração de Dados

A proliferação das novas tecnologias em praticamente todas as vertentes da sociedade contemporânea – a Sociedade da Informação na qual a informação desempenha um papel primordial [WWW InformationSociety] – originou elevadas quantidades de informação, muitas vezes dispersa, disponibilizada em diversos formatos e acessível a uma grande quantidade de utilizadores um pouco por todo o mundo.

O desenvolvimento de bases de dados tão variadas e em tão elevada quantidade só foi possível através da disseminação de ferramentas tecnológicas, que facilitam a sua introdução e fornecem um suporte para as mesmas. Também a extracção de informação útil (conhecimento) a partir dessas bases de dados tem que ser apoiada em ferramentas capazes de relacionar, comparar e extrair de toda a informação existente aquela que tem de facto utilidade para a tomada de decisão, que muitas vezes é apenas uma pequena parte [WWW TheEarling].

Um dos processos de extracção de conhecimento existentes actualmente – e amplamente utilizado – é a mineração de dados. A mineração de dados é um mecanismo de extracção de conhecimento a partir de grandes quantidades de informação, através da utilização de modelos estatísticos, de aprendizagem, de inteligência artificial, entre outros [Sumathi & Sivanandam, 2006]. Para além de permitir extrair a informação relevante e de a organizar de forma a ser facilmente compreendida e aplicada à tomada de decisão, muitas vezes leva à descoberta de relacionamentos insuspeitos, permitindo alargar a perspectiva de análise [Hand et al., 2001].

É actualmente evidente que a mineração de dados deve ser tomada em conta quando a quantidade de dados existente numa empresa ou organização não permite que ela seja analisada em tempo útil, de forma a contribuir para a tomada de decisão. A extracção de relatórios de actividade, desde os simples aos mais complexos, apenas permite que se retire da informação existente exactamente aquilo que se pretende retirar, nomeadamente:

- quantidades vendidas de um determinado produto durante o mês passado;
- quantidade de produto em stock;
- preços médios, mínimos e máximos praticados para um produto.

Todo o conhecimento à partida imprevisível ou complexo de extrair, mas muitas vezes de elevado valor, não é tido em conta:

- relação da venda de um produto com a venda de outro produto;

- relação da diminuição das vendas de um produto com o aumento do preço de outro produto;
- estimativa da venda de um produto nas próximas semanas.

No sentido de auxiliar as organizações a implementar mais eficazmente e obter melhores resultados da mineração de dados surgiu, nos anos 90, o modelo CRISP-DM (CRoss-Industry Standard Process for Data Mining) [WWW CRISPDM]. Este modelo foi iniciado por três especialistas na área de mineração de dados e refinado ao longo de vários anos com a contribuição de mais de 300 organizações. A metodologia a utilizar, independente do negócio a analisar e das ferramentas a utilizar, consiste numa série de fases que devem ser seguidas de forma a tornar um projecto de mineração de dados mais rápido, mais barato, mais fiável e mais fácil de utilizar.

Este modelo define um conjunto de fases (Figura 1) para as quais existe um conjunto de tarefas a executar:

- **Compreensão do negócio:** compreender as necessidades de análise para o negócio em questão; definir os objectivos da mineração de dados e um plano de trabalho;
- **Compreensão dos dados:** analisar os dados disponíveis e avaliar a sua qualidade;
- **Preparação dos dados:** reunir todos os dados disponíveis; limpar e transformar os dados consoante as necessidades;
- **Modelação:** escolher o modelo de mineração, de acordo com as necessidades; construir o modelo, analisar e refinar;
- **Avaliação:** avaliar o modelo desenvolvido e os resultados obtidos, em relação aos objectivos;
- **Implementação:** avaliar como devem ser disponibilizados os resultados e implementar.

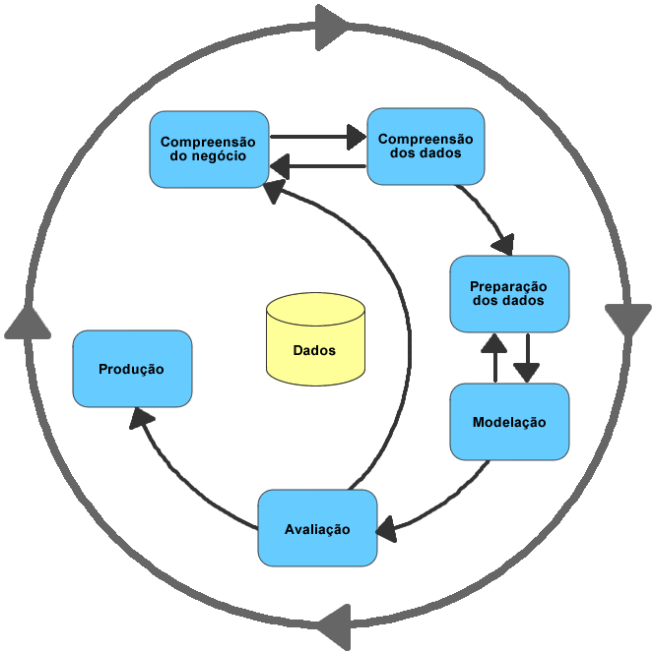


Figura 1: Fases da metodologia CRISP-DM (adaptado de [WWW CRISPDM])

As fases descritas, quando aplicável, irão ser seguidas no caso de estudo proposto nesta dissertação, nos capítulos seguintes deste relatório.

## 1.3 Motivação e objectivos

### 1.3.1 Motivação

Ao processo de recolha selectiva de resíduos está associada uma grande quantidade de informação: existem ecopontos e contentores que é necessário recolher; existem rotas de recolha que definem o trajecto a efectuar para recolher um conjunto de ecopontos; e existem equipas de trabalho, normalmente compostas por uma viatura, um motorista e um ou dois ajudantes. A cada recolha efectuada está ainda associada a informação relativa à mesma: os ecopontos e contentores visitados ou recolhidos; o estado de cada contentor (em termos de enchimento, higiene, avarias, entre outros); a equipa de trabalho que os recolheu; as datas de início e fim de trabalho; os

quilómetros da viatura no início e no fim do trabalho; as quantidades de combustível consumidas; as quantidades de produto recolhidas; entre outras informações. Todas estas informações são importantes para as empresas que efectuam a recolha dos ecopontos, que as utilizam de forma a gerir o seu negócio, determinando as rotas a efectuar por cada equipa de trabalho em cada dia da semana na tentativa de recolher a maior quantidade de resíduos possível, optimizando ao mesmo tempo os recursos dispendidos com esse processo.

Essas empresas tradicionalmente baseavam-se no registo manual de informação de negócio. As equipas de trabalho registavam todas as informações do seu dia de trabalho em folhas de papel e posteriormente a informação registada era analisada na tentativa de encontrar informação que permitisse planear melhor o trabalho. Contudo, o processamento dessa informação, não sendo apoiado por ferramentas informáticas, era ineficaz e não permitia que fosse dado o devido tratamento a toda a informação existente. Face a estas dificuldades, há muito que essas empresas reconheceram a necessidade de se munirem de ferramentas que lhes permitam automatizar a recolha e análise da informação. Desta forma, as equipas de trabalho passaram a introduzir a informação directamente em suporte digital, a introdução passou a estar menos sujeita a erros e a ficar imediatamente disponível para análise, através de ferramentas de consulta de informação.

Actualmente estas empresas pretendem extrair da informação registada o conhecimento que lhes permita determinar com mais precisão o estado global do seu negócio e, mais importante ainda, prever qual será o estado futuro. A informação histórica armazenada nas bases de dados operacionais pode ser analisada em detalhe, utilizando ferramentas de extracção de conhecimento, na tentativa de encontrar informação que permita prever acontecimentos futuros com base nos acontecimentos passados, potenciando a descoberta da solução ideal para a gestão e optimização do negócio.

### 1.3.2 Objectivos

O objectivo deste trabalho consistiu no estudo e aplicação de técnicas de mineração de dados para descoberta de conhecimento na área da recolha selectiva, que proporcione o apoio à tomada de decisão e à optimização dos processos de empresas que efectuam a recolha de ecopontos. Em particular, pretendeu-se elaborar um mecanismo eficaz para determinação da frequência ideal de recolha das rotas, que garantam a recolha de grandes quantidades de produto, diminuam o consumo de recursos e evitem a reclamação por parte dos cidadãos.

## 1.4 Organização da dissertação

No capítulo 2 é descrito o modo de operação das empresas que efectuam a recolha selectiva de resíduos, é analisada uma ferramenta utilizada nestas empresas para apoio ao negócio, os dados recolhidos e as limitações existentes para a tomada de decisão. Estas fases correspondem às fases de compreensão do negócio e de compreensão dos dados descritas na metodologia CRISP-DM, seguida no desenvolvimento deste projecto.

No capítulo 3 são analisados alguns mecanismos de mineração de dados e algumas áreas de aplicação.

No capítulo 4 são analisados exemplos de aplicação de mineração de dados, em contextos com algumas semelhanças à recolha selectiva de resíduos.

No capítulo 5 são descritos dois modelos de mineração de dados elaborados para determinar a frequência ideal de recolha de ecopontos, em diferentes contextos, utilizando os mecanismos de séries temporais e *clustering*. Neste capítulo são seguidas as fases de preparação dos dados, modelação e avaliação da metodologia CRISP-DM.

Finalmente no capítulo 6 são descritas as conclusões a retirar do trabalho elaborado e o trabalho futuro a desenvolver.

# Capítulo 2

# A Recolha Selectiva

## 2.1 Modo de operação

A recolha selectiva de resíduos consiste na deposição dos materiais recicláveis, devidamente separados por tipo de material, em locais apropriados para o efeito, designados de ecopontos, que são locais de deposição constituídos por um ou mais contentores. Os resíduos recicláveis mais comuns estão divididos em três categorias: papel e cartão; embalagens de plástico e metal; e embalagens de vidro, que correspondem aos contentores azul, amarelo e verde, respectivamente.

Os ecopontos estão distribuídos por todo o território português, sendo que existem em maiores quantidades nas cidades. A sua distribuição tem em conta a densidade populacional, a proximidade de determinados elementos (como escolas e restaurantes), entre outros factores. Também a quantidade de contentores de cada um dos tipos existentes em cada ecoponto pode variar consoante as necessidades. Por exemplo, junto a escolas é presumível que os ecopontos tenham de

ser reforçados com um contentor de papel, enquanto que junto a restaurantes poderá ser necessário mais um contentor de embalagens ou vidro.

Os cidadãos são responsáveis pela deposição dos materiais nos ecopontos e existem empresas, geralmente entidades públicas como aterros sanitários e câmaras municipais, responsáveis pela recolha destes materiais. Cada empresa ou entidade é responsável por recolher os ecopontos localizados dentro da sua zona de actuação e por transportar esses resíduos para centros de triagem, que posteriormente irão efectuar a triagem de todo o material e encaminhá-lo para os centros de valorização.

As empresas que efectuam a recolha dos ecopontos geralmente definem rotas de recolha, constituídas pelo conjunto de ecopontos que as equipas de trabalho devem "visitar" e recolher quando essa recolha se justificar, ou seja, quando a quantidade de produto existente nos contentores for superior a um determinado valor (geralmente acima de 75% da capacidade do contentor). As rotas são construídas tendo em conta a distribuição dos ecopontos, naturalmente procura-se que as viaturas não percorram muitos quilómetros entre ecopontos, sempre que isso seja possível. A recolha é feita por viaturas concebidas para o efeito, com uma caixa para onde é recolhido o material e uma grua que permite a elevação e a descarga do contentor. Geralmente cada viatura recolhe apenas um produto de cada vez, mas existem casos em que as empresas optam por dividir a caixa do camião em duas ou três partes, de forma a poderem recolher mais que um produto ao mesmo tempo. Dependendo das características das viaturas, as rotas podem ser mono ou multi-produto, isto é, para cada ecoponto podem conter apenas os contentores de um dos produtos ou os contentores de vários produtos. Também pode acontecer o mesmo contentor estar associado a várias rotas, nos casos em que os contentores enchem com muita frequência.

A definição do conjunto de rotas de recolha é feita, geralmente, quando existem alterações significativas na distribuição dos ecopontos ou quando se prevê que a alteração das rotas permitirá melhorar o desempenho da empresa. Pode também esporadicamente alterar-se ligeiramente a constituição das rotas existentes, adicionando ou removendo ecopontos e contentores ou mudando-

os para outras rotas, por diversas razões. No entanto, a redefinição completa de todas as rotas causa alguns problemas, pois obriga a que as equipas de trabalho tenham de se habituar a um novo traçado, provocando alguma ineficiência nas primeiras execuções das novas rotas. A atribuição das rotas de recolha às equipas de trabalho deve ter em conta, ente outros, os seguintes factores:

- Os ecopontos devem ser recolhidos de modo a optimizar a sua recolha, isto é, evitando a deslocação de uma viatura a um ecoponto que não tem (ou tem pouco) material para recolher.
- Os ecopontos devem ser recolhidos de modo a evitar perturbar os cidadãos, seja porque estão muito cheios e impedem a deposição de produto ou porque o produto já está depositado há demasiado tempo e começa a libertar cheiros desagradáveis.

Para fazer esta atribuição correctamente as empresas necessitam de dados relativos a todos os seus ecopontos e respectivos contentores. Quer sejam dados reais, recolhidos no terreno recentemente, quer sejam previsões, com base no histórico de recolhas de cada um dos contentores. Só tendo esta informação as empresas podem avaliar quais as rotas mais adequadas a serem recolhidas em cada momento, de modo a maximizar a quantidade de material recolhido, minimizar os recursos utilizados e satisfazer as expectativas das populações.

## 2.2 Análise de uma ferramenta de apoio à Recolha Selectiva

Na tentativa de optimizar os processos de recolha dos ecopontos, as empresas que actuam nesta área verificaram que necessitavam de recolher o máximo de informação possível relativamente a todos os processos da recolha. Era essencial que as empresas tivessem a noção clara do comportamento de todos os seus ecopontos, de forma a poderem melhor agrupá-los em rotas de recolha e a recolhê-los na altura certa.

Desde há alguns anos que estas empresas fazem a recolha desses dados, recorrendo às equipas de trabalho, constituídas pelo motorista da viatura de recolha e, em alguns casos, por um ou dois ajudantes. Estes trabalhadores eram responsáveis por anotar, em folhas de papel criadas para o efeito, todas as informações relacionadas com a recolha: ecopontos visitados, estado de enchimento dos contentores, avarias que eventualmente existiam, entre outras informações.

A análise de toda esta informação registada em papel era complexa e muitas vezes não permitia tirar conclusões em tempo útil, isto é, antes da recolha seguinte. Também a análise do comportamento de cada um dos contentores de cada um dos ecopontos era complicada, devido à enorme quantidade de equipamentos e à enorme quantidade de dados recolhida. Muitas empresas optavam por digitalizar toda esta informação, registando-a manualmente em folhas de cálculo, mas depressa verificaram que era um trabalho moroso e com poucos resultados práticos, pois dificilmente a informação era digitalizada e analisada em tempo útil.

### 2.2.1 SPAR - Sistema de Planeamento e Análise da Recolha

Surgiu a necessidade de criar uma ferramenta que permitisse registar estes dados em suporte digital e que fornecesse mecanismos eficazes de tratamento dessa informação. Foi nesse sentido que a empresa Cachapuz[1] desenvolveu a ferramenta SPAR – Sistema de Planeamento e Análise da Recolha, que permite registar a informação relacionada com a recolha em suporte digital, analisar toda essa informação e planear o trabalho das equipas de recolha com base na informação registada [WWW Cachapuz]. Disponibiliza ainda mecanismos para que os cidadãos possam contribuir com informação importante para o sistema. Os principais módulos desta ferramenta serão descritos de seguida.

[1] Cachapuz, Equipamentos para Pesagem, Lda., Parque Industrial de Sobreposta, Braga

### BackOffice

Este é o módulo principal e consiste numa ferramenta *desktop*, com uma base de dados relacional para suporte à informação, que permite gerir todas as entidades do processo, analisar a informação registada e atribuir as rotas às equipas de trabalho.

Neste módulo são configuradas as entidades base, que determinam a informação que pode ser registada pelas equipas de trabalho, tais como:

- Produtos que podem ser recolhidos, geralmente papel, embalagem e vidro.
- Níveis de enchimento em que os contentores podem estar na altura de recolha, por exemplo, a 0%, 25%, 50%, 75% ou 100% da sua capacidade.
- Estados de higiene possíveis para um contentor, por exemplo, limpo, médio ou sujo.
- Avarias que podem existir num contentor, por exemplo, partido, incendiado ou sem tampa.

No BackOffice são catalogados todos os motoristas, viaturas e ainda todos os ecopontos, aos quais é associada a sua localização, coordenadas geográficas e os respectivos contentores. A cada contentor está associado o tipo de contentor (papelão, embalão ou vidrão), os últimos dados registados (relativamente ao seu enchimento, higiene, data da última recolha, etc.) e ainda uma taxa de enchimento, que é calculada com base nas recolhas efectuadas. É também neste módulo que os ecopontos e contentores são agrupados em rotas, que são depois atribuídos às equipas de trabalho a cada dia. Neste módulo existe ainda uma secção de consultas, que permite configurar um conjunto de *queries* para extrair da base de dados a informação necessária, como por exemplo as quantidades de produto recolhidas por mês, os quilómetros efectuados pelas viaturas, o último estado registado para cada um dos contentores de cada uma das rotas, etc.

## Cartografia

Este é um módulo que permite representar a informação presente no sistema sobre um sistema de informação geográfica (SIG). Este módulo permite identificar num mapa a localização dos ecopontos do sistema, o nível de enchimento dos contentores de cada um dos produtos e os percursos tomados pelas viaturas quando efectuaram as recolhas.

A disponibilização da informação em modo cartográfico facilita a sua visualização, uma vez que permite identificar rapidamente, por exemplo, as zonas onde é necessário efectuar recolhas (são visíveis no mapa os ecopontos mais cheios, identificados com cor vermelha) e as zonas que podem ser recolhidas mais tarde (as zonas no mapa onde os ecopontos estão identificados com cor branca).

## Mobilidade

Este módulo consiste numa componente para dispositivos móveis (PDAs), para ser utilizada no terreno pelas equipas de trabalho. Esta ferramenta indica às equipas de trabalho quais os ecopontos e contentores a recolher e permite registar toda a informação relativa aos equipamentos (níveis de enchimento, estados de higiene, eventuais avarias, entre outros dados). Embora as equipas normalmente só recolham os contentores de um dos produtos, este o módulo permite-lhes registar a informação relativa aos contentores dos outros produtos, possibilitando assim às empresas ter a informação mais actualizada.

É também possível registar outros dados relacionados com o trabalho, como os quilómetros e as horas de início e fim de trabalho, as pesagens para cada um dos produtos recolhidos (tara, bruto e líquido) e os litros de combustível abastecidos durante o dia de trabalho. Estando os PDAs equipados com uma antena GPS, o módulo regista periodicamente as coordenadas geográficas da viatura, permitindo posteriormente traçar o percurso efectuado.

A informação é sincronizada para o PDA no início do dia de trabalho, para que o PDA receba a informação relativa à(s) rota(s) a recolher, e no final do dia de trabalho, para que a informação registada pela equipa fique disponível para análise no BackOffice.

### Portal do Cidadão

O SPAR disponibiliza um portal para comunicação com o cidadão. Este portal disponibiliza informação de interesse para o cidadão, como os ecopontos da sua área de residência, representados sobre um mapa cartográfico, para ser mais atractivo para a população. Disponibiliza ainda algumas estatísticas relativas à recolha, como as quantidades de produto recolhidas, o número de viaturas e funcionários da entidade, entre outras.

Este portal possibilita ainda ao cidadão dar a sua opinião sobre a actuação da empresa e, mais importante ainda, informar sobre a necessidade de recolher um ecoponto ou contentor. Esta informação, enviada pelo cidadão, é uma mais-valia para a empresa, pois permite-lhe programar as rotas com base na informação actualizada enviada pela população.

### SMS

Complementarmente ao portal do cidadão, o módulo de SMS também permite ao cidadão informar a empresa sobre a necessidade de recolher ou reparar um ecoponto, mas através de uma mensagem SMS (*Short Message Service*).

Este módulo é particularmente importante em zonas remotas onde existem ecopontos, distantes dos restantes ecopontos de uma rota (por exemplo, inseridos em parques naturais). A empresa poderá designar um responsável por enviar um SMS quando o ecoponto precisar de ser recolhido, evitando assim que as viaturas percorram diversos quilómetros sem que seja necessário.

### *Business Viewer*

O módulo *Business Viewer* consiste numa aplicação de visualização da informação de negócio em alto nível. Nesta aplicação são configurados uma série de alertas e indicadores, que podem ser agrupados em vistas, e que são executados previamente para que a informação esteja imediatamente disponível quando é pretendida, sem ser necessário esperar que a aplicação processe a informação contida na base de dados. Por exemplo, pode ser criada uma vista que representa o comportamento das rotas, contendo um gráfico com o número de ecopontos de cada rota, um gráfico com as quantidades médias de produto recolhidas para cada rota e uma tabela com o número de vezes que cada rota foi recolhida no último mês.

Tendo em conta as características actuais dos gestores de empresas, com pouco tempo para a utilização de ferramentas para consulta de informação, com pouco tempo para interpretação de informação e em constante movimento, a aplicação permite a programação da entrega de relatórios (por e-mail ou SMS), simples de interpretar. É possível, por exemplo, configurar a aplicação de forma a enviar automaticamente no final de cada mês um e-mail com o gráfico de quantidades recolhidas por produto, o gráfico de quilómetros e horas dispendidos por motorista e a tabela com o total de combustível consumido.

### *Business Intelligence*

Este módulo, ainda em fase de desenvolvimento, agrega a informação do SPAR num *Data Warehouse*, uma base de dados multi-dimensional que agrupa a informação do sistema e que pode inclusivamente incluir dados provenientes de outras fontes [Rainardi, 2008]. Esta tecnologia permite que os dados históricos sejam preservados, ao contrário das bases de dados tradicionais em que a informação se perde à medida que é alterada.

O módulo de *Business Intelligence* facilita o acesso aos dados, pois a sua consulta é imediata uma vez que toda a informação já está previamente relacionada e calculada. Possibilita aceder a dados

históricos que permitem, por exemplo, comparar as quantidades recolhidas e a prestação da empresa ao longo dos meses. Esta ferramenta, devidamente explorada, constitui uma importante fonte de informação de decisão.

Em suma, o SPAR permite gerir, registar e analisar toda a informação relacionada com o processo de recolha selectiva. A quantidade de informação registada diariamente pelas equipas de trabalho é considerável e é analisada por três tipos de pessoas, com objectivos e necessidades distintas:

- o responsável pelo processo de recolha, que necessita de analisar o histórico e o estado actual dos seus equipamentos para planear o trabalho dos dias seguintes;
- o gestor ou administrador da empresa, que necessita de saber se a sua empresa está a cumprir com as expectativas, em termos de quantidades a recolher e recursos a despender;
- o cidadão, que pretende saber informações gerais e estatísticas sobre a empresa que recolhe na sua área de residência e enviar informações sobre a necessidade de recolher ou reparar o seu ecoponto.

É particularmente importante neste sistema que a informação flua correctamente para o responsável pela recolha, de modo a que possa tomar as decisões certas em tempo útil e optimizar os processos da empresa.

### 2.2.2 Dados armazenados

A informação registada no SPAR é armazenada numa base de dados relacional. As principais tabelas (Figura 2) contêm a informação relativa aos equipamentos da empresa (ecopontos e respectivos contentores) e relativa ao trabalho realizado pelas equipas (turnos de trabalho).

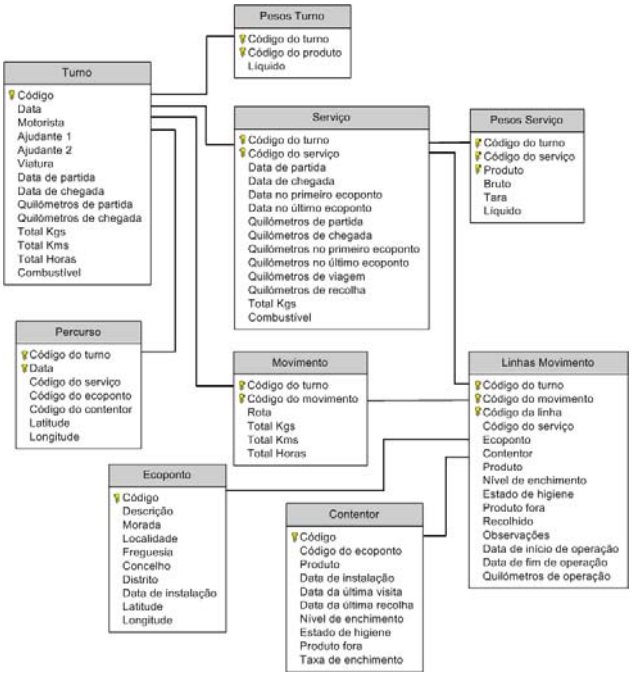


Figura 2: Esquema da Base de Dados do SPAR

Duas tabelas essenciais são os ecopontos e contentores, que correspondem aos equipamentos que a empresa tem que recolher e gerir. Um ecoponto contém um código e uma descrição, que permitem identificá-lo, e ainda a localidade, freguesia, concelho e distrito onde está localizado, a data em que foi instalado naquele local e as coordenadas geográficas. A cada ecoponto estão associados contentores, que têm um código que o identifica, o código do ecoponto a que pertence, o produto a que corresponde (que identifica o contentor como sendo um papelão, embalão ou vidrão), a data de instalação e os últimos dados registados pelas equipas de trabalho (a data da última visita e da última recolha efectuada, o nível de enchimento, o estado de higiene e se tinha produto fora) e ainda a sua taxa de enchimento. A taxa de enchimento é a percentagem que o contentor enche por dia, calculada através da média de dias que o contentor demora a encher uma determinada quantidade e que, a cada visita, é actualizada (Fórmula 1).

$$\text{Taxa de enchimento} = \frac{\text{Taxa de enchimento anterior} + \frac{\text{Enchimento actual} - \text{Enchimento anterior}}{\text{Dias passados}}}{2}$$

Fórmula 1: Cálculo da taxa de enchimento de um contentor

Na base de dados são também armazenados os turnos de trabalho, cada turno tem associada a equipa de trabalho (viatura, motorista e eventuais ajudantes), a data de início e fim de trabalho, os quilómetros de início e fim de trabalho, um resumo do total de quilómetros, horas e quantidades de produto recolhidas e ainda do total de litros de combustível abastecidos durante o dia de trabalho.

A cada turno estão associados serviços, que representam a saída da equipa de trabalho das instalações da empresa, a recolha de uma série de ecopontos e o regresso às instalações para descarregar o produto. Por cada turno de trabalho podem existir vários serviços, uma vez que as viaturas podem ficar sem capacidade de recolha por mais que uma vez e necessitar de ser descarregadas de forma a permitir continuar a recolha dos ecopontos. Em cada serviço ficam registadas as datas e quilómetros de início e fim, são também registadas as datas e quilómetros de chegada ao primeiro ecoponto e de partida do último ecoponto (para que possam ser calculados os tempos e quilómetros gastos sem que sejam recolhidos ecopontos) e ainda os quilómetros efectuados em viagem, em recolha, os quilogramas de produto recolhidos e os litros de combustível eventualmente abastecidos.

A cada turno estão também associados os movimentos, que correspondem às rotas a recolher, e que contêm o resumo do total de quilogramas de produto recolhidos, quilómetros percorridos e horas consumidas. Os movimentos contêm as linhas de movimento, que são todos os contentores dos ecopontos que fazem parte das rotas. A cada linha está associado o código do serviço no qual a linha foi registada, o ecoponto, contentor e o produto correspondentes e os dados relativos ao estado do contentor: nível de enchimento verificado, estado de higiene, se tinha ou não produto

fora, se foi ou não recolhido, outras observações, data de início e de fim de operação e quilómetros da viatura.

No final de cada serviço, quando a viatura se dirige às instalações para descarregar, são registadas as pesagens efectuadas, contendo cada uma o turno e o serviço a que pertencem, o produto que está a ser pesado e os pesos bruto, tara e líquido.

A cada turno são associadas as pesagens totais para cada um dos produtos pesados nos serviços do turno, neste caso fica apenas registado o produto e o peso líquido total. De referir que a cada turno podem estar associadas uma ou mais rotas e a recolha dos contentores dessas rotas pode ser efectuada em um ou mais serviços. Como os pesos dos produtos são registados por cada serviço, pode não existir uma correspondência directa entre o peso registado no serviço e o peso associado a uma rota, pois depende se todos os contentores dessa rota foram recolhidos nesse serviço. Para resolver este problema o SPAR efectua uma distribuição ponderada dos pesos por todos os contentores recolhidos em cada um dos serviços, permitindo calcular um valor aproximado ao peso recolhido na respectiva rota. O peso de cada contentor é calculado segundo uma regra de três simples (Fórmula 2): o peso total no serviço está relacionado com a soma dos enchimentos do serviço, da mesma forma que o peso do contentor está relacionado com o enchimento do contentor; logo, segundo a regra de três simples, o peso do contentor corresponde à multiplicação do peso total no serviço pelo enchimento do contentor, tudo dividido pela soma dos enchimentos no serviço.

Este valor apresenta apenas uma estimativa, pois não é possível saber com exactidão o peso existente em cada contentor. O facto de o enchimento de um contentor ser registado a 75% da sua capacidade não implica necessariamente que o seu peso seja superior ao de um contentor com registo de 50% de enchimento.

Durante a recolha dos ecopontos, o SPAR vai registando, através da antena de GPS do PDA, as coordenadas geográficas correspondentes aos pontos de passagem da viatura. A esta informação

estão associados o turno e o serviço respectivos, a data a que se refere aquele ponto de passagem, o ecoponto e o contentor que estavam a ser registados no momento (caso a equipa esteja naquele momento a registar a recolha de um contentor) e as coordenadas geográficas: latitude e longitude.

Peso total no serviço —— Soma dos enchimentos no serviço

Peso do contentor —— Enchimento do contentor

$$\text{Peso do contentor} = \frac{\text{Peso total no serviço} \times \text{Enchimento do contentor}}{\text{Soma dos enchimentos no serviço}}$$

Fórmula 2: Cálculo do peso por contentor (estimado)

A base de dados permite manter alguns dados históricos, pois para cada contentor visitado ou recolhido são registados os dados que representam o seu estado na altura. Estes dados permitem analisar historicamente a evolução do ecoponto em termos de enchimento. Também no caso das rotas, é possível analisar a evolução da configuração das rotas e a evolução dos pesos recolhidos em cada fase.

Estes e outros dados que podem ser extraídos da base de dados fornecem informações importantes para os gestores das empresas que operam nesta área, nomeadamente os responsáveis pelo planeamento do trabalho para as diferentes equipas, com vista a minimizar os recursos dispendidos com as recolhas.

### 2.2.3 Limitações para a tomada de decisão

O sistema SPAR permite registar, em formato digital, toda a informação relacionada com o processo de recolha dos ecopontos, sendo que a quantidade de informação registada diariamente é considerável. Uma empresa de média dimensão numa zona não densamente populacional terá

cerca de 700 ecopontos, se tivermos em conta os dados da Tabela 1, relativos a 5 empresas relevantes no território nacional [WWW ADP].

| **Empresa** | **Municípios Abrangidos** | **N.º Ecopontos** |
|---|---|---|
| Rebat | Amarante, Baião, Cabeceiras de Basto, Celorico de Basto, Marco de Canaveses e Mondim de Basto | 360 |
| Resat | Boticas, Chaves, Montalegre, Ribeira de Pena, Valpaços e Vila Pouca de Aguiar | 350 |
| Resioeste | Alcobaça, Alenquer, Arruda dos Vinhos, Azambuja, Bombarral, Cadaval, Caldas da Rainha, Lourinhã, Nazaré, Óbidos, Peniche, Rio Maior, Sobral de Monte Agraço e Torres Vedras | 1370 |
| Resulima | Arcos de Valdevez, Barcelos, Esposende, Ponte da Barca, Ponte de Lima e Viana do Castelo | 832 |
| Valorlis | Batalha, Leiria, Marinha Grande, Ourém, Pombal e Porto de Mós | 750 |

Tabela 1: Número de ecopontos associados a várias empresas de recolha no território português

Cada ecoponto tem, em média, 3 contentores, o que para uma empresa de média dimensão perfaz 700 * 3 = 2100 contentores. Analisando os dados recolhidos pelo SPAR, verifica-se que em média as rotas de recolha contêm cerca de 50 contentores, o que para uma empresa de média dimensão origina cerca de 42 rotas (2100 / 50 = 42). Os contentores são recolhidos, em média, de 3 em 3 dias, o que implica que existam 14 equipas de trabalho. Se a recolha for efectuada de forma cíclica, a calendarização das recolhas poderia ser efectuada tal como representada na Tabela 2.

| Dia 1 | Dia 2 | Dia 3 | Dia 4 | Dia 5 | Dia 6 | ... |
|---|---|---|---|---|---|---|
| Rota 1 | Rota 15 | Rota 29 | Rota 1 | Rota 15 | Rota 29 | ... |
| Rota 2 | Rota 16 | Rota 30 | Rota 2 | Rota 16 | Rota 30 | ... |
| Rota 3 | Rota 17 | Rota 31 | Rota 3 | Rota 17 | Rota 31 | ... |
| Rota 4 | Rota 18 | Rota 32 | Rota 4 | Rota 18 | Rota 32 | ... |
| Rota 5 | Rota 19 | Rota 33 | Rota 5 | Rota 19 | Rota 33 | ... |
| Rota 6 | Rota 20 | Rota 34 | Rota 6 | Rota 20 | Rota 34 | ... |
| Rota 7 | Rota 21 | Rota 35 | Rota 7 | Rota 21 | Rota 35 | ... |
| Rota 8 | Rota 22 | Rota 36 | Rota 8 | Rota 22 | Rota 36 | ... |
| Rota 9 | Rota 23 | Rota 37 | Rota 9 | Rota 23 | Rota 37 | ... |
| Rota 10 | Rota 24 | Rota 38 | Rota 10 | Rota 24 | Rota 38 | ... |
| Rota 11 | Rota 25 | Rota 39 | Rota 11 | Rota 25 | Rota 39 | ... |
| Rota 12 | Rota 26 | Rota 40 | Rota 12 | Rota 26 | Rota 40 | ... |
| Rota 13 | Rota 27 | Rota 41 | Rota 13 | Rota 27 | Rota 41 | ... |
| Rota 14 | Rota 28 | Rota 42 | Rota 14 | Rota 28 | Rota 42 | ... |

Tabela 2: Exemplo de calendarização de rotas

Tendo em conta estes dados, pode-se verificar que por dia são registados os dados de 50 * 14 = 700 contentores, o que significa diariamente 700 linhas na base de dados apenas para a tabela LinhasMovimento, que representa cada contentor de cada rota.

Cada dia de trabalho corresponde a cerca de 8 horas por equipa, nas quais são registadas coordenadas de posicionamento geográfico a cada 5 minutos. Um dia de trabalho corresponde, por isso, a 8 * 60 = 480 minutos, o que significa 480 / 5 = 96 registos na tabela Percurso por cada equipa de trabalho, o que multiplicado por todas as equipas resulta em 96 * 14 = 1344 registos nesta tabela.

Esta elevada quantidade de informação que é registada diariamente pelo SPAR é de grande utilidade, pois permite ter a noção exacta da forma como o trabalho é efectuado e de como se comporta todo o sistema. É possível retirar da informação relatórios de actividade, como as rotas efectuadas e respectivas quantidades recolhidas, horas dispendidas e quilómetros percorridos. É ainda possível analisar em detalhe o comportamento de um ecoponto ou contentor ao longo de um período e, com isso, tomar algumas medidas no sentido de melhorar a recolha desse contentor ou satisfazer a população servida por ele.

No entanto, devido ao facto de o volume de informação ser muito elevado, a análise completa do comportamento de todos os ecopontos e contentores e de todas as rotas é muito difícil, principalmente porque este sistema apenas disponibiliza ferramentas de consulta directa à base de dados, que apresentam os dados resultantes de uma série de *queries*. A utilização do SPAR permite optimizar os processos em determinadas situações, por exemplo, permite determinar os contentores que terão maiores quantidades de produto, quer pela análise dos últimos dados registados, quer pela estimativa de enchimento calculada com base na taxa de enchimento média do contentor. No entanto, a actividade de recolha selectiva não é linear e o enchimento real de um contentor pode ser influenciado por uma série de factores externos, que não podem ser facilmente deduzidos dos dados existentes através de um conjunto de *queries* à base de dados.

### 2.2.4 Áreas de aplicação

O SPAR actualmente está a ser usado em empresas que efectuam a recolha de ecopontos no território nacional, na maior parte dos casos aterros sanitários que possuem frota de veículos e equipas de trabalho para a recolha dos ecopontos, e por uma empresa em Angola para a recolha de ecopontos e de contentores de resíduos sólidos urbanos. Em algumas empresas o SPAR está a ser usado para registar a recolha de produto não em ecopontos, mas em particulares que produzem muito produto reciclável que, pelas elevadas quantidades produzidas, não pode ser depositado no ecoponto. Estes particulares estabelecem contratos com as empresas que efectuam a recolha, para

que esse produto seja recolhido. Algumas das empresas que usufruem deste serviço são, por exemplo, grandes superfícies comerciais, hipermercados, restaurantes, entre outros.

Devido às necessidades de registo de informação existentes noutras áreas semelhantes, como a recolha porta-a-porta, a recolha de óleos alimentares usados, a recolha de biomassa, entre outras, o SPAR poderá vir a ser adaptado às realidades destas empresas, que possuem algumas diferenças em relação à recolha de ecopontos, conforme descrito em seguida.

**Recolha porta-a-porta**

No caso da recolha porta-a-porta são recolhidos, tal como nos ecopontos, os produtos papel, embalagem e vidro. Mas neste caso em vez de existirem ecopontos com diversos contentores (em número mais ou menos fixo), existem pontos onde são recolhidos sacos com produto reciclável. Estes pontos podem ser, por exemplo, moradias, blocos de apartamentos e pequenos espaços comerciais. Neste sistema são recolhidos os sacos existentes em cada ponto e são fornecidos novos sacos, vazios, para que possam ser enchidos e entregues na próxima recolha.

O número de sacos em cada ponto não é fixo, depende da produção de cada entidade em cada ponto. Se a empresa que recolhe os resíduos verificar que a entidade terá capacidade para produzir mais resíduos pode optar por reforçar o número de sacos disponíveis para a entidade ou, no caso contrário, diminuir o número de sacos. Na próxima recolha não há garantia de que todos os sacos serão recolhidos, depende da produção efectuada pela entidade em questão.

Para este caso de gestão de recolhas é necessário, para além de registar as recolhas efectuadas, que as empresas tenham a noção de quantos sacos estão espalhados por quantos pontos e para cada um destes pontos é importante que se conheça a sua capacidade de produção, para se poderem planear recolhas com base nas quantidades totais espectáveis, dividindo a recolha dos pontos pelas viaturas disponíveis para o fazer.

### Recolha de óleos alimentares usados

Este caso é muito semelhante ao anterior, mas neste caso o produto a recolher é óleo alimentar usado, os recipientes são contentores de plástico e geralmente os pontos correspondem a restaurantes. Também neste caso o número de contentores associados a cada ponto não é fixo, pode variar ao longo do tempo consoante a capacidade de produção registada em cada um dos pontos.

Para a gestão desta área de negócio é necessário, tal como no caso anterior, ter a noção das quantidades recolhidas e do número de contentores espalhados por todos os pontos de recolha. Mas neste caso surge mais uma necessidade, registar contentores que têm que ser lavados e que, por isso, deixarão de estar disponíveis para entrega. Estes dados são essenciais para que o gestor possa planear devidamente o trabalho das suas equipas e também para que possa gerir os seus equipamentos, nomeadamente o número de contentores de que deverá dispor para suprir as necessidades dos clientes, tendo em conta as lavagens que são necessárias efectuar e que tornam os contentores indisponíveis durante um determinado período.

### Recolha de biomassa

A recolha de biomassa consiste na recolha de resíduos de florestas, que posteriormente são encaminhados para pontos de produção de energia recorrendo a este tipo de combustível. Os resíduos de florestas são agrupados em montes, espalhados por uma área florestal e posteriormente uma viatura recolhe cada um desses montes. A construção dos montes é feita de modo a facilitar o seu processo, ou seja, são feitos montes junto às zonas onde estão os resíduos (estes apenas são amontoados para facilitar a recolha), o que implica que os pontos de recolha não são fixos. Também as quantidades de produto existentes nos montes não é homogénea, podendo haver montes em que a quantidade é superior à capacidade da viatura e montes em que essa quantidade é inferior.

Para as empresas que gerem esta área, é necessário registar a produção dos montes, identificado geograficamente onde se encontram e a quantidade estimada existente. Quando um monte é recolhido, é necessário registar que ele foi recolhido e, caso não tenha sido recolhido todo o produto existente, deverá ser indicado o produto remanescente no monte, para que possa ser recolhido posteriormente. Com esta informação o gestor pode planear as recolhas, tendo em conta os montes existentes, as quantidades respectivas, as distâncias entre os montes e as capacidades de cada um dos veículos da sua frota, que não são necessariamente homogéneas.

# Capítulo 3

# Análise de alguns mecanismos de Mineração de Dados

## 3.1 Mecanismos de Mineração de Dados

A mineração de dados surgiu pela necessidade de extrair de grandes quantidades de informação conhecimento útil para o negócio. Mais do que relacionar dados e confirmar suspeitas, como tradicionalmente efectuado através de *queries* directas a bases de dados, a mineração de dados pretende descobrir informação menos evidente, que possa ser transformada em informação de decisão [Sumathi & Sivanandam, 2006]. A mineração de dados é um dos passos da descoberta de conhecimento em bases de dados (*Knowledge Discovery in Databases*), descrita como "o processo não trivial de identificação de padrões nos dados válidos, novos, potencialmente úteis e em última instância compreensíveis" [Fayyad et al., 1996], no qual os mecanismos de mineração de dados são aplicados iterativamente.

A análise dos dados em busca de conhecimento começou por ser aplicada nos anos 60, sendo nessa altura apenas análise estatística, em que eram aplicadas técnicas estatísticas como correlação, regressão, chi-quadrado, entre outros [Kimball et al., 1998]. Mais tarde a estes métodos foram sendo adicionados outros como lógica difusa e redes neuronais, mas foi já nos anos 90 que as técnicas foram consolidadas e disponibilizadas em ferramentas de mineração de bases de dados empresariais e começaram a ser utilizadas efectivamente em sistemas de *data warehouse* entretanto construídos.

Alguns métodos e algoritmos utilizados pelos mecanismos de mineração de dados são descritos em seguida [Sumathi & Sivanandam, 2006].

### Estatística

Na mineração de dados é comum o uso de modelos estatísticos, construídos a partir de um conjunto de dados de treino. Estas técnicas permitem a criação de um modelo óptimo através da pesquisa na informação, com base numa medida estatística que se pretende obter, de modo a obter padrões e regras existentes entre os dados.
Alguns dos modelos estatísticos usados são a regressão, que permite definir um conjunto de atributos que dão origem a um determinado valor; a correlação, que analisa a correspondência de variáveis entre si; e a segmentação, que permite encontrar segmentos de entre um conjunto de dados, com base em determinadas medidas.

### Aprendizagem

Os métodos de aprendizagem pesquisam os dados na tentativa se encontrar o melhor modelo possível que coincide com os dados analisados, sendo usualmente utilizadas heurísticas no processo de pesquisa.

Os métodos mais comuns utilizam árvores de decisão, que são incluídas no conjunto de decisão e que permitem classificar um objecto com base no percurso efectuado desde a raiz da árvore até ao nodo, sendo que o percurso é determinado com base nas características do objecto.

**Visualização**

A exploração visual consiste na transformação de dados em objectos visuais, como pontos, linhas e áreas, o que permite que os dados sejam apresentados em espaços bi- ou tridimensionais, permitindo aos utilizadores explorar visualmente e interactivamente os dados. Um exemplo da utilização destas técnicas é a criação de gráficos, que permitem visualmente analisar a informação de forma mais rápida e eficaz.

A combinação de várias destas técnicas permite efectuar uma série de análises sobre os dados, tais como as que são descritas em seguida [Berry & Linoff, 2004]. As ferramentas comerciais nesta área incluem mecanismos de mineração que permitem efectuar algumas destas análises, com diversas aplicações práticas nos dias de hoje.

### 3.1.1 Classificação

A classificação consiste na análise das características dos elementos de análise, de forma a determinar a que classes pertencem. Neste método existe um número finito e pré-definido de classes, nas quais os elementos se podem incorporar. Exemplos da sua utilização.

- Classificar uma aplicação financeira segundo o seu risco: baixo, médio ou alto.
- Determinar, com base nas características dos clientes (idade, rendimentos, extracto social, etc.), se eles estarão dispostos a adquirir um determinado produto: sim ou não.

### 3.1.2 Segmentação (*Clustering*)

Consiste na segmentação (*clustering*) de um conjunto de dados heterogéneos em subgrupos (*clusters*) homogéneos, sem que estes subgrupos estejam pré-determinados.
Um exemplo da sua utilização:

- Agrupar um conjunto de alunos segundo as suas apetências (sem que o conjunto total de apetências, por exemplo, matemática ou línguas, sejam previamente conhecidas).

### 3.1.3 Estimativa

Consiste no cálculo de valores desconhecidos, para um determinado input. Este cálculo analisa as características conhecidas dos elementos e estima os valores para características desconhecidas. Alguns exemplos da sua utilização:

- Estimar os rendimentos dos clientes.
- Estimar a probabilidade de um cliente adquirir um produto.

### 3.1.4 Previsão

A previsão permite classificar um elemento ou estimar valores desconhecidos, tal como a classificação ou a estimativa, mas baseando-se numa previsão de acontecimentos. Este método, ao contrário dos anteriores, tem em atenção o relacionamento temporal entre os elementos, permitindo prever acontecimentos futuros com base nos acontecimentos passados. Alguns exemplos de previsões:

- Prever, de entre um conjunto de pessoas, quais as que irão frequentar o ensino superior.
- Prever quais os clientes que irão adquirir um determinado produto ou subscrever um determinado serviço.

### 3.1.5 Regras de associação

Permite determinar que "coisas" estão interligadas. Através desta técnica podem-se estabelecer regras a partir dos dados. Por exemplo:

- Determinar a probabilidade de um cliente comprar o produto X, quando também comprar o produto Y.
- Criar um cabaz de compras, constituído por um conjunto de produtos que possam ser adquiridos por um conjunto de clientes.

### 3.1.6 Séries temporais

Permitem prever dados futuros com base num conjunto de dados variáveis no tempo [Crows, 1999]. Estes mecanismos têm em conta as propriedades dos períodos temporais, como as hierarquias, a sazonalidade e certas ocasiões, como as festividades religiosas, períodos de férias, entre outros. Alguns exemplos da utilização desta técnica:

- Com base nas vendas de um produto nos anos passados, prever as vendas para o próximo ano, estimando a variação ao longo dos meses.
- Com base no histórico do valor de um conjunto de acções na bolsa, prever a evolução para os próximos meses.

## 3.2 Áreas de aplicação da Mineração de Dados

A mineração de dados tem sido usada em diversas áreas, como análise de vendas, apoio a campanhas de marketing e selecção de produtos, astronomia, geologia, medicina, aprovação de créditos, detecção de fraudes e lavagem de dinheiro e mesmo no desporto [Sumathi & Sivanandam, 2006]. Pode ser usada para melhorar o desempenho das empresas, aumentar as vendas, prever

quantidades de produto necessárias em stock, melhorar o relacionamento com os clientes, entre outras. A análise dos dados históricos dos clientes permite encontrar padrões e tendências que definem o perfil dos clientes, possibilitando campanhas publicitárias mais eficazes. Já a análise dos stocks e o seu comportamento permite identificar produtos com boas performances, nos quais poderá ser importante investir.

Algumas destas áreas de actuação e exemplos de utilização destes mecanismos na indústria são descritos em seguida.

### 3.2.1 Transacções comerciais

As grandes empresas negoceiam diariamente com milhares de clientes e efectuam milhares de transacções. Necessitam de tentar melhorar a satisfação dos seus clientes, prevenir transacções fraudulentas, definir os produtos que devem ser apresentados em campanhas de marketing, entre outros. A análise simples de um gráfico com a representação do volume de vendas de cada produto ao longo do ano pode indicar as datas preferenciais para reforçar os stocks ou fazer campanhas publicitárias sobre determinado produto. Mas esta análise não permite identificar outros factores que influenciam a aquisição de um determinado bem. Para conseguir atingir os seus objectivos, estas empresas podem analisar os dados relativos às suas transacções efectuadas sob diversas perspectivas, como por exemplo as seguintes:

- Clientes: em que circunstâncias os clientes compram determinados produtos (existe alguma relação com a estação do ano ou uma época festiva?); qual o comportamento dos clientes que efectuam diversas compras em alturas distintas (existe um padrão para o tipo de produtos que adquirem ou a época do ano em que são efectuadas as compras?).
- Produtos: de que forma a compra de um produto está relacionada com a compra de outro ou outros produtos; quais as circunstâncias que provocam um aumento no volume de vendas dos produtos (época do ano, campanha publicitária?).

A empresa American Express utilizou mecanismos de mineração de dados para extrair informação útil para marketing, tendo obtido bons resultados com a aplicação dessa informação e aumentando o uso de cartões de crédito em cerca de 10% [Sumathi & Sivanandam, 2006].

### 3.2.2 Dados de negócio e comércio electrónico

Os processos de negócio geram grandes quantidades de informação, usualmente gerada por aplicações de *back office*, *front office* e de rede. Nas aplicações de *back office* é gerida toda a informação de suporte aos sistemas e ainda a informação relativa à utilização do sistema, como os utilizadores autorizados, o registo das operações efectuadas por cada utilizador e dos erros ocorridos. Nas aplicações de *front office* são efectuadas as operações sobre o sistema, como a inserção, edição ou remoção de produtos, preços, quantidades em stock, vendas e encomendas. As aplicações de rede permitem geralmente utilizar uma pequena parte das funcionalidades do sistema, tal como a inserção de vendas ou de encomendas.

A análise integrada de toda esta informação pode revelar informações importantes sobre os processos nas empresas e auxiliar a melhoria dos mesmos e pode ser efectuada, por exemplo, sob as seguintes perspectivas:

- Picos de acesso: em que alturas do dia, da semana ou do mês ocorrem picos de acesso ou alteração dos dados e de que forma esses picos influenciam o desempenho do sistema.
- Erros: em que circunstâncias ocorrem erros (quais os processos ou os utilizadores que originam mais erros); quais dos erros existentes influenciam fortemente o desempenho do sistema.

### 3.2.3 Meteorologia

Existe uma grande quantidade de informação relativa a condições atmosféricas que é obtida de sensores espalhados por todo o planeta. Esta informação poderá permitir determinar as relações entre eventos meteorológicos dispersos, estabelecendo causas para determinados acontecimentos e antecipando outros, como por exemplo:

- De que forma a temperatura ou a corrente da água em determinados pontos dos oceanos influencia a formação de tempestades em determinadas zonas do globo.
- Com base na variação da temperatura ao longo dos últimos anos, qual é a previsão para a temperatura nos próximos anos.
- Qual a relação entre a temperatura, a pluviosidade, os níveis de dióxido de carbono na atmosfera e o degelo dos pólos.

O projecto Environmental Scenario Search Engine (ESSE) usa mecanismos de mineração de dados para pesquisar informação em bases de dados de registos meteorológicos [WWW ESSE]. Com este projecto será possível analisar informações importantes sobre condições meteorológicas, efectuando *queries* complexas aos dados recolhidos.

### 3.2.4 Simulação

Cada vez mais os cientistas usam modelos de simulação para testar ou provar as suas teorias. Estas simulações produzem geralmente quantidades enormes de informação, que tem de ser analisada utilizando técnicas eficazes. As técnicas tradicionais de consulta a bases de dados tendem a ser demoradas e a extrair apenas uma pequena parte do conhecimento existente em toda a informação.

O projecto AUTO-OPT [WWW AUTOOPT], desenvolvido para a indústria automóvel, possibilita a simulação dos comportamentos de um automóvel na fase de concepção. Aos dados numéricos da simulação são aplicados mecanismos de mineração de dados para permitir encontrar soluções para problemas que possam ser detectados na simulação.

### 3.2.5 Assistência médica

Na área de assistência médica existe uma grande quantidade de informação sobre patologias, pacientes, tratamentos, custos e resultados obtidos. A determinação de relações entre estas variáveis poderá ser de grande importância no sentido de indicar os melhores tratamentos a disponibilizar em determinadas circunstâncias, tal como nos seguintes exemplos:

- De entre um conjunto de pacientes que sofreram de uma determinada patologia, quais os tratamentos eficazes e em que circunstâncias o foram (idade do paciente, sexo, massa corporal, outras patologias existentes, entre outros dados).
- Qual a relação entre a ocorrência de uma patologia e as condições sociais dos pacientes (local de habitação, tipo de emprego, estado civil ou número de filhos).
- Qual a relação entre o aparecimento de uma série de patologias no mesmo paciente e em que condições ocorrem (devido a tratamentos a outras patologias, devido a hábitos alimentares, entre outros).

### 3.2.6 Web

Actualmente a Web facilita a partilha de informação, que varia desde texto, música, imagem ou vídeo. Com tão vasta quantidade de informação disponível, a busca e descarga da informação pretendida é dificultada, sendo cada vez mais evidente a necessidade de mecanismos que auxiliem este processo. Seria interessante se existisse um mecanismo que alojasse automaticamente a

informação que mais pessoas pretendem aceder em sites com melhores desempenhos (em termos de largura de banda de acesso e capacidade de tratamento de pedidos) e a informação menos importante em sites com menores desempenhos (que possivelmente não iriam deteriorar o acesso a essa informação, pois existiriam menos pessoas a aceder a ela). A escolha da informação mais "importante" a disponibilizar podia ser efectuada através da análise, por exemplo, dos seguintes factores:

- Quais as pesquisas mais efectuadas nos principais motores de busca nos últimos dias e qual a sua relação com os conteúdos descarregados.
- Quais as noticias de maior destaque nos principais jornais nos últimos dias e qual a influência de cada jornal (em termos de acesso a conteúdos presentes no jornal) nos diversos países.
- Qual a época do ano ou época festiva actual (é presumível que no verão se pretendam mais informações sobre viagens, enquanto que na época de natal sejam mais requisitadas sugestões de prendas).

# Capítulo 4

# Estudo de casos de aplicação da Mineração de Dados

## 4.1 Estudo do *Vehicle Routing Problem* e variantes

A recolha selectiva de resíduos exige a gestão da deslocação de viaturas a determinados pontos, denominados de ecopontos, para recolha dos resíduos existentes, configurando um problema que partilha algumas características com o *Vehicle Routing Problem*, descrito em seguida.

### *Vehicle Routing Problem* (VPR)

O *Vehicle routing problem*, proposto por Dantzig and Ramser em 1959, consiste na entrega de produtos a um conjunto de clientes utilizando uma frota de veículos [WWW VehicleRoutingProblem]. Os veículos partem de um depósito e o somatório dos pesos ou volumes dos produtos a entregar por uma viatura não pode exceder a sua capacidade. O objectivo é optimizar a entrega dos produtos, percorrendo o menor número de quilómetros possível. Para este problema, será

necessário definir a atribuição dos clientes às diferentes viaturas e as rotas a percorrer [Savelsbergh, 2002]. A Figura 3 representa um exemplo do problema, com a distribuição geográfica dos clientes, do depósito e a definição das rotas a efectuar pelas viaturas.

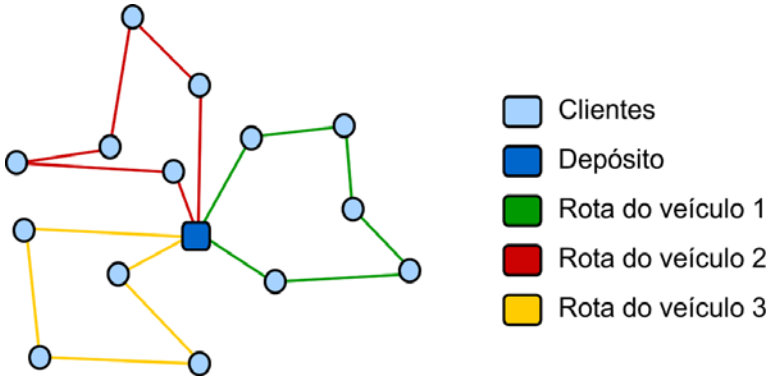


Figura 3: Exemplo de *Vehicle Routing Problem*

Existem algumas variantes para este problema, nas quais são adicionadas mais variáveis ou condicionantes e que aumentam a complexidade do problema. Algumas dessas variantes são descritas em seguida [Marković et al., 2005].

### ***Vehicle routing problem with Time Windows*** **(VRPTW)**

No VRPTW é adicionada uma janela temporal, determinada pela necessidade de atender os clientes num determinado período de tempo. Para este problema, para além de ser necessário definir a atribuição dos clientes às viaturas e as rotas, passa a ser necessário também definir o agendamento. A Figura 4 representa a configuração das rotas para um problema com estas características, sendo que neste caso a ordem com que as rotas são efectuadas não é indiferente.

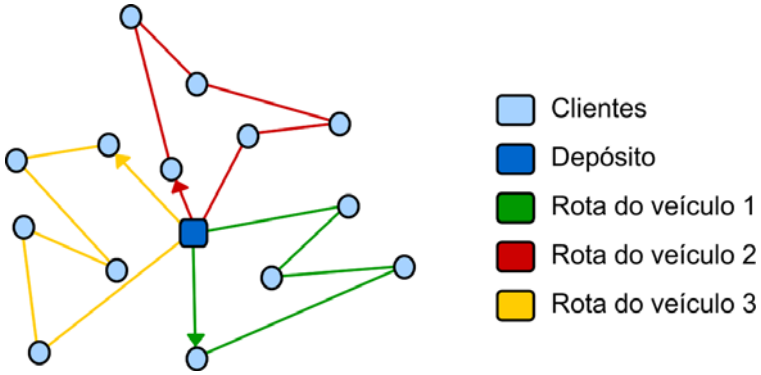


Figura 4: Exemplo de *Vehicle routing problem with Time Windows*

### ***Distance-Constrained Vehicle Routing Problem*** **(DCVRP)**

No DCVRP existe um limite de distância ou tempo que é necessário para cumprir uma rota. Isto pode acontecer, por exemplo, pela necessidade de cumprir com os horários de trabalho dos motoristas das viaturas. Em relação aos problemas anteriores, este poderá exigir que existam mais viaturas para o mesmo cenário de clientes de forma a cumprir as suas exigências, tal como ilustrado na Figura 5.

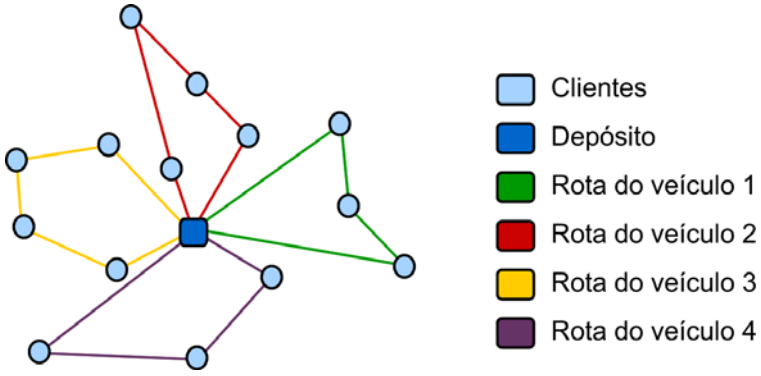


Figura 5: Exemplo de *Distance-Constrained Vehicle Routing Problem*

### ***Vehicle Routing Problem with Backhauls*** **(VRPB)**

No caso do VRPB, não se trata apenas de entregar produtos a um conjunto de clientes, mas também de recolher produtos noutro conjunto de clientes. Neste caso o conjunto de clientes é dividido em duas partes e a complexidade aumenta significativamente, pois para cada viatura é necessário ter em conta a sua capacidade para transportar os produtos para os clientes onde é necessário entregar, mas é preciso também garantir que a viatura não recolhe produtos junto de um cliente se a sua capacidade disponível não o permitir (Figura 6).

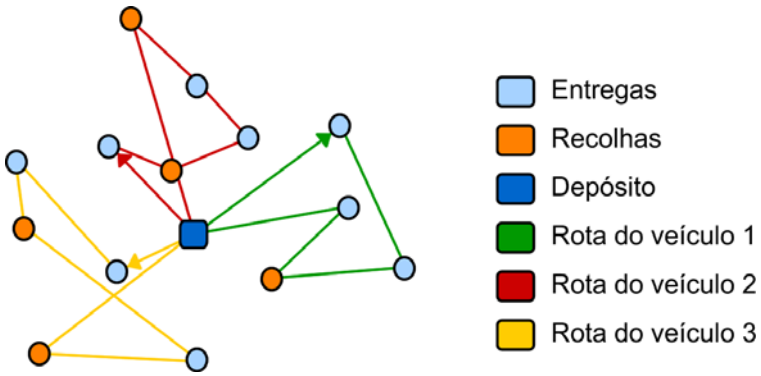


Figura 6: Exemplo de *Vehicle Routing Problem with Backhauls*

### ***Vehicle Routing Problem with Pickup and Delivery*** **(VRPPD)**

O VRPPD configura um caso em que para cada cliente existe simultaneamente a entrega e a recolha de produtos, não necessariamente homogéneos em termos de peso ou volume. Neste caso uma viatura só poderá deslocar-se a um cliente se tiver capacidade tanto para entregar, como para recolher os produtos respectivos àquele cliente (Figura 7).

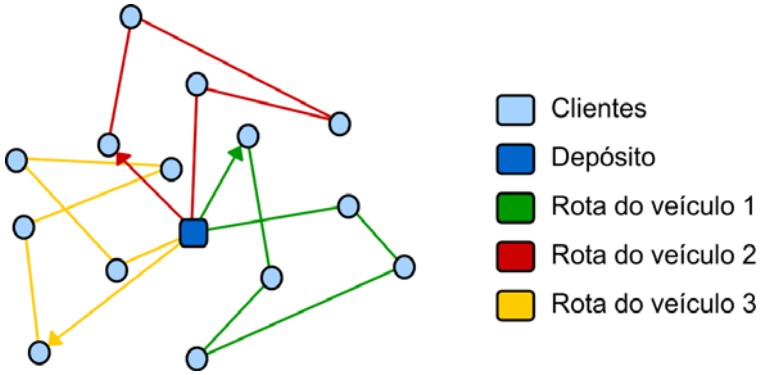


Figura 7: Exemplo de *Vehicle Routing Problem with Pickup and Delivery*

Existem várias aproximações para a resolução destes problemas, entre os quais a utilização de algoritmos exactos – que geram a solução exacta mas são demasiado complexos para problemas com um grande número de entradas – e de algoritmos heurísticos – algoritmos de aproximação rápidos que geram uma solução estimada, embora não havendo garantia de a solução gerada estar próxima da solução exacta.

Os métodos heurísticos geralmente consistem de duas fases, uma de *clustering* e uma de geração de rota (*route*) [Bramely & Simchi-Levi, 1993]. A execução destas fases pode seguir a ordem *clustering* seguido de geração de rota (*cluster first – route second*), ou a ordem geração de rota seguida de *clustering* (*route first – cluster second*). No primeiro caso, são gerados segmentos de clientes com base nas suas características e depois são criadas rotas eficientes para os vários grupos. No segundo caso é gerada uma única rota que passa por todos os clientes e depois essa rota é dividida em segmentos.

Dois exemplos de resolução dos algoritmos VRP, apresentados em seguida, usam mineração de dados em contextos diferentes [Marković et al., 2005] e [Hu & Huang, 2007].

### 4.1.1 Abordagem à resolução de *Vehicle Routing Problem with Stochastic Demands and Time Windows*

O primeiro caso [Marković et al., 2005] consiste na resolução de um problema do tipo *Vehicle Routing Problem with Stochastic Demands and Time Windows* (VRPSDTW). Esta é uma variação do VRPTW (descrito acima), mas em que à partida não são conhecidas as quantidades de produto pretendidas por cada cliente, o que configura um cenário de *stochastic demands*.

Neste caso, cada viatura terá de percorrer uma determinada rota para entregar produtos aos clientes, no entanto poderá ser necessário regressar ao local de depósito para recarregar a viatura, no sentido de satisfazer todos os clientes (Figura 8).

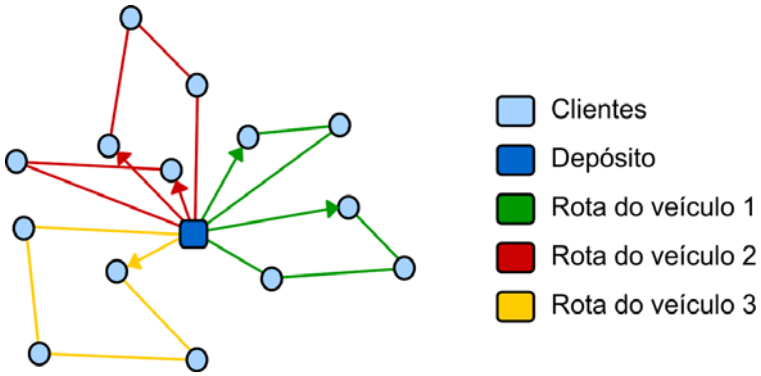


Figura 8: Exemplo de *Vehicle Routing Problem with Stochastic Demands and Time Windows*

Para a resolução deste problema, os autores optaram por usar um mecanismo de mineração de dados, no caso redes neuronais com retro-propagação, para prever as quantidades que irão ser necessárias para cada cliente, conseguindo com isto reduzir o problema VRPSDTW a um problema VRPTW (pois as quantidades encomendadas já são conhecidas). Com base nessa previsão, é aplicado um algoritmo de resolução de VRP, que determina as rotas a efectuar, tendo em conta a sequência dos clientes a visitar e as deslocações necessárias ao depósito.

A metodologia desenvolvida é, segundo os seus autores, eficaz na resolução do problema VRPSDTW, e pode ser aplicada na resolução de VRP com qualquer tipo de incerteza.

### 4.1.2 Abordagem à resolução de *Vehicle Routing Problem with Time Windows*

O segundo estudo [Hu & Huang, 2007] consistiu na resolução de um problema do tipo *Vehicle Routing Problem with Time Windows* (VRPTW), em que é usada uma extensão aos métodos heurísticos de duas fases. Neste caso são utilizadas três fases: classificação de clientes, geração de rotas e um modelo de programação inteira.

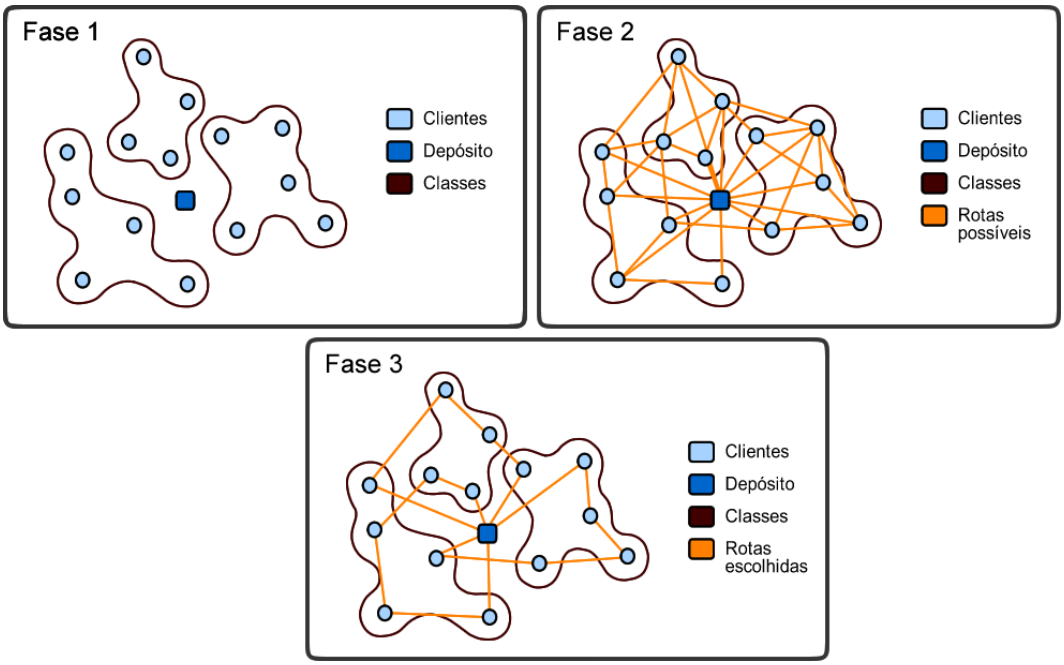


Figura 9: Exemplo de resolução de *Vehicle Routing Problem with Time Windows* em três fases

A fase de classificação de clientes é efectuada por ferramentas de mineração de dados, usando um mecanismo de *clustering*, e permite classificar os clientes por áreas com base em diversos atributos, como a distância, o nível do pedido, a sua densidade, a configuração da sua cidade, entre

outros. A fase de geração de rotas cria as rotas possíveis, tendo em conta as quantidades pedidas pelos clientes, a capacidade das viaturas disponíveis e os tempos máximos por rota. Finalmente é aplicado o modelo de programação inteira, que analisa todas as rotas possíveis e determina as mais eficazes, em termos de redução de custos de operação (Figura 9).

A metodologia desenvolvida neste trabalho, segundo os seus autores, permite melhorar o desempenho na resolução do problema VRPTW e permite a alteração dos parâmetros do problema, gerando novas soluções com facilidade.

## 4.2 Estudo da Previsão da Procura (*Demand Forecasting*)

A previsão da procura (*demand forecasting*) consiste na estimativa da quantidade de um produto ou serviço que os clientes irão adquirir [WWW DemandForecasting]. A previsão das vendas, mais comummente usada nas empresas para gerir os seus processos produtivos, permite definir as necessidades de produção, organizar e planear a produção, mas não permite explorar o verdadeiro valor da procura para os produtos [Oracle 2005].

Existem muitos outros factores que influenciam a procura de um produto, como a sua disponibilidade, o seu preço, o preço de produtos relacionados e a época do ano. Estes factores têm uma influência directa no valor real dos produtos e, consequentemente, nas quantidades a produzir, nas quantidades a manter em armazém, nas promoções a efectuar, nos preços a praticar, entre outros. É possível aplicar mecanismos de mineração para efectuar esta previsão no sentido de optimizar os processos de produção e comercialização das empresas, tal como efectuado num trabalho na área da produção de papel [Respício et al., 2002], apresentado em seguida.

### 4.2.1 Abordagem à previsão da procura de papel

Neste estudo [Respício et al., 2002] é aplicado o mecanismo de séries temporais para prever as necessidades de produção, com base na análise das encomendas efectuadas em períodos passados (Figura 10). No caso estudado foram analisados os nove meses anteriores ao estudo e, com base nesses dados, foram previstas as necessidades de produção para os três meses seguintes, com base em padrões que pudessem existir nos dados.

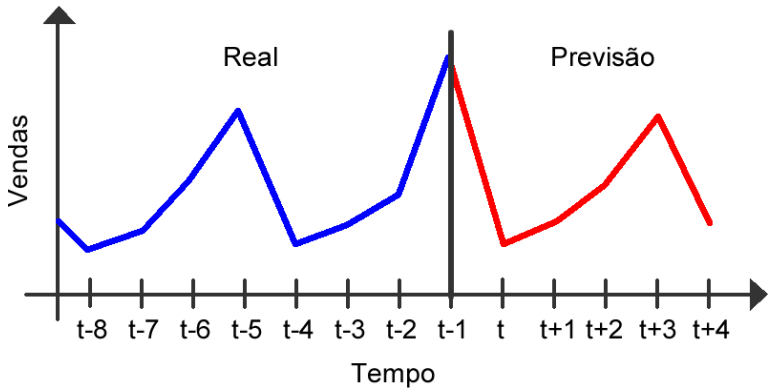


Figura 10: Exemplo de mecanismo de séries temporais

A metodologia desenvolvida neste estudo compreende outras fases, fora do âmbito da mineração de dados. A solução global, segundo os autores, permitiu optimizar os processos da empresa, reduzindo custos de operação e permitindo uma coordenação mais eficaz do processo de planeamento.

## 4.3 O caso da Recolha Selectiva

Algumas das necessidades encontradas nos problemas anteriores também são encontradas no problema da recolha selectiva de resíduos, pelo que a utilização de semelhantes mecanismos de mineração de dados poderá ser útil.

Fazendo o paralelismo com os casos de *Vehicle Routing Problem* analisados, a optimização da recolha selectiva poderá ser efectuada usando redes neuronais para prever o enchimento dos contentores, tal como foram usadas no primeiro exemplo para prever as quantidades de produto pretendidas pelos clientes; ou poderão ser usados mecanismos de *clustering* para agrupar contentores com características semelhantes num número desconhecido de classes, tal como foram agrupados os clientes do segundo exemplo.

Já no caso da previsão de procura, também existem algumas semelhanças entre as necessidades desta área de negócio e a deposição dos resíduos nos ecopontos, uma vez que esta deposição poderá seguir um padrão numa determinada escala temporal. O cálculo desse padrão poderá servir de base para permitir estimar futuras variações no tempo, pelo que, tal como no exemplo abordado, poderão ser usados mecanismos de séries temporais para prever o enchimento dos contentores com base nos padrões detectados nas recolhas anteriores.

Para além destes, outros mecanismos de mineração poderão ser usados para tarefas semelhantes, uma vez que alguns deles permitem efectuar o mesmo tipo de análise, embora utilizando metodologias diferentes e, presumivelmente, produzindo resultados diferentes. A associação de elementos a classes pode ser efectuada com mecanismos de associação ou *clustering* (conforme se conheçam ou não à partida as diferentes classes) e a previsão pode ser efectuada com mecanismos de previsão, regras de associação, entre outros.

# Capítulo 5

# Aplicação de Mineração de Dados na Recolha Selectiva

## 5.1 Casos de estudo

O objectivo deste trabalho consistiu na construção de modelos de mineração de dados que permitissem optimizar o processo da recolha selectiva de resíduos. Em particular, pretendia-se determinar a frequência "ideal" para recolher cada uma das rotas configuradas, proporcionando a recolha de grandes quantidades de produto, diminuindo o consumo de recursos e evitando a reclamação por parte dos cidadãos

No desenvolvimento deste projecto pressupõe-se que as rotas existentes numa determinada empresa de recolha de resíduos recicláveis estão pré-definidas e são fixas, durante o período de análise. A criação de rotas optimizadas para os ecopontos que a empresa recolhe (tendo em conta a quantidade de ecopontos existentes, as características geográficas e demográficas de cada

ecoponto, entre outras) é efectuada, normalmente por empresas especializadas, na fase de arranque do processo de recolha de resíduos naquela empresa. Estas rotas são revistas e optimizadas de tempos a tempos, quando se verifica um decréscimo de produtividade ou uma desadequação das rotas devido à reconfiguração dos ecopontos do sistema.

Para este caso de estudo foi utilizada a base de dados de uma empresa que utiliza a ferramenta SPAR para registar as recolhas efectuadas desde o final de 2006, a empresa Resulima, SA[2], que é responsável por efectuar a recolha dos ecopontos em todos os concelhos do Distrito de Viana do Castelo e no concelho de Barcelos, distrito de Braga. A escolha desta empresa para servir de base de teste deve-se ao facto de ser, de entre as empresas que utilizam o SPAR, a que usa há mais tempo e que, por isso, possui uma base de dados mais rica.

Nesta empresa as rotas de recolha são fixas e cada rota contém todos os contentores de um ecoponto (rota multi-produto), embora as viaturas recolham apenas um ou dois produtos de cada vez (apenas vidro ou papel e embalagem). Para este estudo foram apenas tidos em conta os dados registados no ano de 2007 uma vez que não estão completos os dados de 2006, pois a ferramenta SPAR começou a ser usada apenas no mês de Dezembro desse ano, nem os de 2008, na altura em que este trabalho foi realizado. Nos dados de 2007 estão definidas 16 rotas fixas (códigos 01, 02, ..., 16) dos produtos papel, embalagem e vidro (códigos 01, 02 e 03).

## 5.2 Desenvolvimento

Este projecto foi desenvolvido no contexto da empresa Cachapuz, pelo que a escolha das ferramentas a utilizar foi limitada aos requisitos da referida empresa. O motor de base de dados de utilizado foi o Microsoft SQL Server 2005, a ferramenta de mineração de dados usada foi Microsoft

[2] Resulima - Valorização e Tratamento de Resíduos Sólidos, SA, Viana do Castelo

SQL Server 2005 Analysis Services e a ferramenta de desenvolvimento foi Microsoft Visual Studio 2005.

A ferramenta Microsoft SQL Server 2005 Analysis Services (SSAS) contém algoritmos para efectuar as operações mais usuais em mineração de dados, algumas das quais foram descritas nos capítulos anteriores deste relatório: classificação, regressão, *clustering*, associação e análise sequencial [WWW MSDN]. Na Tabela 3 estão representados todos os algoritmos disponíveis nesta ferramenta e as tarefas que permitem efectuar. De notar que alguns dos algoritmos podem ser usados para mais do que uma tarefa.

| **Tarefa** | **Algoritmos disponíveis** |
|---|---|
| Previsão de um valor discreto | Microsoft Decision Trees Algorithm<br>Microsoft Naive Bayes Algorithm<br>Microsoft Clustering Algorithm<br>Microsoft Neural Network Algorithm (SSAS) |
| Previsão de um valor contínuo | Microsoft Decision Trees Algorithm<br>Microsoft Time Series Algorithm |
| Previsão de uma sequência | Microsoft Sequence Clustering Algorithm |
| Encontrar grupos ou itens comuns em transacções | Microsoft Association Algorithm<br>Microsoft Decision Trees Algorithm |
| Encontrar grupos com itens semelhantes | Microsoft Clustering Algorithm<br>Microsoft Sequence Clustering Algorithm |

Tabela 3: Algoritmos disponíveis na ferramenta Microsoft SQL Server 2005 Analysis Services

Neste projecto foram desenvolvidos dois casos de estudo, que pretendem atingir os objectivos traçados segundo abordagens distintas. O desenvolvimento dos casos de estudo foi efectuado segundo a metodologia CRISP-DM. Nos capítulos seguintes são descritas as fases de preparação de

dados, modelação e avaliação, definidas na metodologia. A fase de implementação, descrita no modelo, não se aplica a este trabalho, uma vez que o modelo desenvolvido não foi aplicado na prática.

### 5.2.1 Optimização da frequência de recolhas com base em séries temporais

A quantidade de resíduos recicláveis que são depositados nos ecopontos não é linear ao longo do ano e tem-se verificado um aumento consecutivo, possivelmente devido a campanhas de fomento à reciclagem ou mesmo pela consciencialização das populações devido aos problemas correntes com o aquecimento global. Mesmo ao longo do ano existem variações nas quantidades, provocadas, por exemplo, pela estação do ano (é presumível que a quantidade de produtos no verão seja superior junto à costa e inferior no interior do país), pela proximidade de épocas festivas (a época do Natal e da Páscoa provocam um aumento generalizado da quantidade de resíduos produzidos, recicláveis ou não recicláveis) ou de feriados nacionais, municipais, entre outros.

Os mecanismos de mineração de dados de séries temporais (*time series*) permitem detectar determinados parâmetros que ocorrem ao longo do tempo e, com base nessa análise, prever acontecimentos futuros [Tange & MacLennan, 2005]. Estes mecanismos analisam dados que ocorrem a espaços temporais sucessivos, por exemplo eventos que ocorrem todos os dias (estado do tempo, como sol ou chuva), todas as semanas (total semanal de vendas de um determinado produto) ou qualquer outro espaço temporal adequado aos dados em análise. No caso da recolha selectiva, pode ser utilizado este mecanismo para detectar a evolução das quantidades recolhidas em cada rota ao longo do tempo, a intervalos regulares, e com base nisso prever a data "ideal" para recolher cada uma das rotas.

### Compreensão dos dados

Com a ferramenta SPAR são registadas todas as rotas recolhidas pelas equipas de trabalho da empresa e, para cada rota, são registados os pesos totais recolhidos de cada produto. Estes dados estão registados nas tabelas TbTurno, TbMovimento e TbPesosTurno.

Na tabela TbTurno é registada, entre outros dados, a data de partida do turno, que pode servir de referência para a data em que a rota foi recolhida (embora em alguns casos as rotas comecem a ser recolhidas num dia e acabem no dia seguinte, no caso das equipas de trabalho nocturnas). Na tabela TbMovimento é registada a rota recolhida naquele turno e na tabela TbPesosTurno são registados os pesos totais recolhidos para cada um dos produtos.

Na análise dos dados verificaram-se alguns casos em que os pesos registados não estariam correctos, estes pesos são registados manualmente no PDA e por vezes são erradamente registados e necessitam de ser corrigidos posteriormente. Na falta dos dados reais, estes pesos errados foram corrigidos, sendo inseridos pesos médios para os produtos em questão.

### Preparação dos dados

As tabelas que armazenam os dados necessários para este caso de estudo contêm demasiada informação, muita dela desnecessária para o caso. Para facilitar as operações foi criada na base de dados do SPAR uma vista que agrupa apenas os dados necessários, denominada vRotasPesos, que contém a informação sobre todas as recolhas efectuadas para todas as rotas, a respectiva data de recolha e os pesos recolhidos por produto (Figura 11).

| Data | Rota | Produto | Peso |
|---|---|---|---|
| 02-01-2007 | 15 | 01 | 3460 |
| 02-01-2007 | 15 | 02 | 1220 |
| 02-01-2007 | 13 | 01 | 1400 |
| 02-01-2007 | 13 | 02 | 800 |
| 02-01-2007 | 04 | 01 | 1200 |
| 02-01-2007 | 04 | 02 | 720 |
| 02-01-2007 | 01 | 01 | 1040 |
| 02-01-2007 | 02 | 01 | 1040 |
| 02-01-2007 | 01 | 02 | 420 |
| ... | ... | ... | ... |

Figura 11: Dados da vista vRotasPesos

Pela análise dos dados recolhidos percebe-se que as rotas não são recolhidas todos os dias e em algumas vezes, embora esporádicas, não são recolhidas sequer uma vez por semana; noutras vezes, no entanto, chegam a ser recolhidas mais que uma vez na mesma semana. Conforme foi referido anteriormente, o mecanismo de mineração de dados de séries temporais trabalha com intervalos regulares de tempo, pelo que será necessário decidir qual das seguintes opções a tomar:

- Inserir dados fictícios que preencham os dias em que as rotas não foram recolhidas, com valores médios para o peso de cada produto, para que a escala de análise temporal seja diária.
- Inserir dados fictícios que preencham as semanas em que as rotas não foram recolhidas e agrupar as recolhas efectuadas por semana, calculando a média dos pesos recolhidos em cada semana para cada produto de cada rota, para que a escala de análise temporal seja semanal.
- Agrupar as recolhas efectuadas por mês ou trimestre, para se obter uma escala temporal mensal ou trimestral, respectivamente.

Conforme foi referido, as rotas não são recolhidas todos os dias, na maior parte das vezes são recolhidas apenas uma ou duas vezes por semana, pelo que a escala temporal adequada para esta análise será a semanal. O resultado da mineração de dados proporcionará dados que permitam analisar qual será a melhor semana para recolher uma rota, o que se adequa à cadência de recolhas efectuada pelas equipas de trabalho.

## Modelação

Para este caso de estudo foi definido que iria ser usado o mecanismo de mineração Microsoft Time Series Algorithm, o único mecanismo de séries temporais disponibilizado no Microsoft SQL Server 2005 Analysis Services, pelo que será necessário construir uma tabela ou base de dados que contenha os dados adequados à utilização deste mecanismo. Foi previamente estabelecido que a escala temporal a utilizar é uma semana, por isso deverá existir uma tabela que contenha todas as semanas existentes no período de análise e, para cada semana, os dados relativos às quantidades recolhidas para cada rota e cada produto.

A vista vRotasPesos criada anteriormente já continha alguns dos dados necessários, no entanto foi necessário adicionar a esses dados uma coluna que identificasse a semana a que se referia cada recolha. Foi ainda necessário agrupar as recolhas das rotas quando ocorreu mais do que uma recolha na mesma semana.

| | Column Name | Data Type | Allow Nulls |
|---|---|---|---|
| 🔑 | IdSemana | float | ☐ |
| 🔑 | RotaProduto | nvarchar(50) | ☐ |
| | Peso | float | ☑ |
| | | | ☐ |

Figura 12: Estrutura da tabela M_PesosRotasSemana

Para simplificar o processo foi criada a tabela M_PesosRotasSemana, apenas com três colunas (Figura 12):

- IdSemana – contém o identificador da semana em que foi efectuada a recolha (valor entre 0 e 52, correspondente às 52 semanas de 2007).
- RotaProduto – contém uma sequência com o código da rota e o código do produto referentes à recolha (por exemplo, R05P01 para a recolha da rota 05 e o produto 01, papel).
- Peso – contém o peso médio recolhido para rota, o produto e a semana em questão.

Posteriormente foi gerada uma *script* que permite recolher os dados da vista vRotasPesos e inseri-los nesta tabela (Figura 13).

Finalmente foi criado o projecto de mineração de dados, escolhendo-se como técnica de mineração o Microsoft Time Series Algorithm. Como fonte de dados foi seleccionada uma vista da tabela criada anteriormente, M_PesosRotasSemana (previamente adicionada ao projecto). Nos tipos de tabelas foi seleccionada a vista da tabela PesosRotasSemana como *Case table*.

| IdSemana | RotaProduto | Peso |
|---|---|---|
| 1 | R01P01 | 596 |
| 1 | R01P02 | 220 |
| 1 | R01P03 | 2585 |
| 1 | R02P01 | 580 |
| 1 | R02P02 | 220 |
| 1 | R02P03 | 3020 |
| 1 | R03P01 | 530 |
| 1 | R03P02 | 180 |
| ... | ... | ... |

Figura 13: Dados da tabela M_PesosRotasSemana

Em relação aos dados de treino foi configurada a seguinte estrutura (Figura 14): a coluna IdSemana é uma chave, neste caso a chave temporal do algoritmo; a coluna RotaProduto é uma chave, que

será usada para identificar a rota e o produto de análise; a coluna Peso é ao mesmo tempo o campo de entrada e a prever.

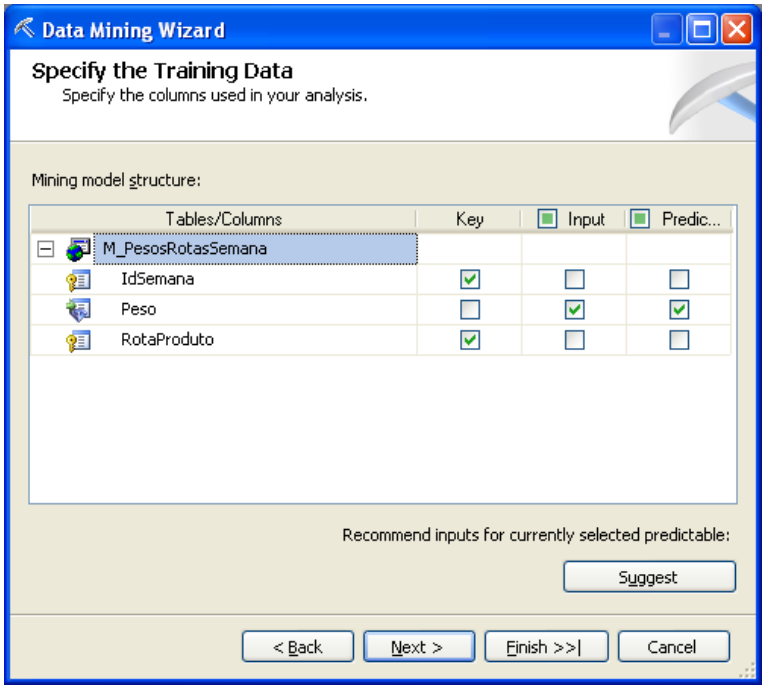


Figura 14: Configuração dos dados de treino para o modelo de séries temporais

Conforme referido anteriormente, os espaços temporais não preenchidos, ou seja, as semanas em que não foram recolhidas determinadas rotas, têm que ser preenchidos para que o mecanismo de séries temporais possa efectuar os cálculos necessários. Para isso foi configurado no algoritmo Microsoft Time Series o parâmetro "*Missing Value Substitution*" (onde é possível especificar o método a utilizar para preencher lacunas nos dados) com o valor "*Mean*" (média) de forma a substituir os valores de peso em falta em algumas semanas pela média dos valores registados (Figura 15).

Após a configuração o algoritmo foi processado, tendo demorado cerca de 15 segundos a processar os dados relativos às recolhas dos 3 produtos das 16 rotas em cada uma das 52 semanas do ano de 2007 (que corresponde a 3 * 16 * 52 = 2496 dados de recolhas).

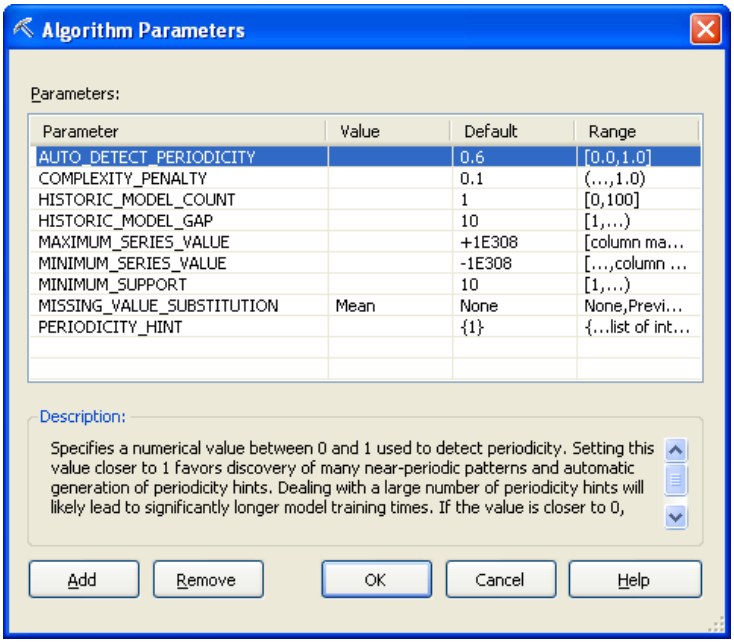


Figura 15: Configuração dos parâmetros do algoritmo de séries temporais

O mecanismo Microsoft Time Series Algorithm produz um gráfico que permite visualmente verificar as previsões que produziu (Figura 16). Neste gráfico existe uma zona que contém os dados reais, que foram submetidos para cálculo da previsão, e uma zona com as próximas previsões, em que é possível definir o número de previsões que pretendemos visualizar, a partir da última data submetida. Estas previsões só serão apresentadas se forem estáveis, o que no caso apresentado não se verifica visto termos poucos dados, pelo que só são apresentadas as próximas duas previsões (que correspondem às duas primeiras semanas de 2008).

Neste gráfico é possível definir que sejam mostradas as previsões históricas, neste caso o mecanismo tem em conta 80% dos dados para construir o modelo (42 semanas) e para os

restantes 20% apresenta uma previsão. Esta pode ser confrontada directamente no gráfico com os dados reais, para se avaliar se as previsões estão próximas da realidade.

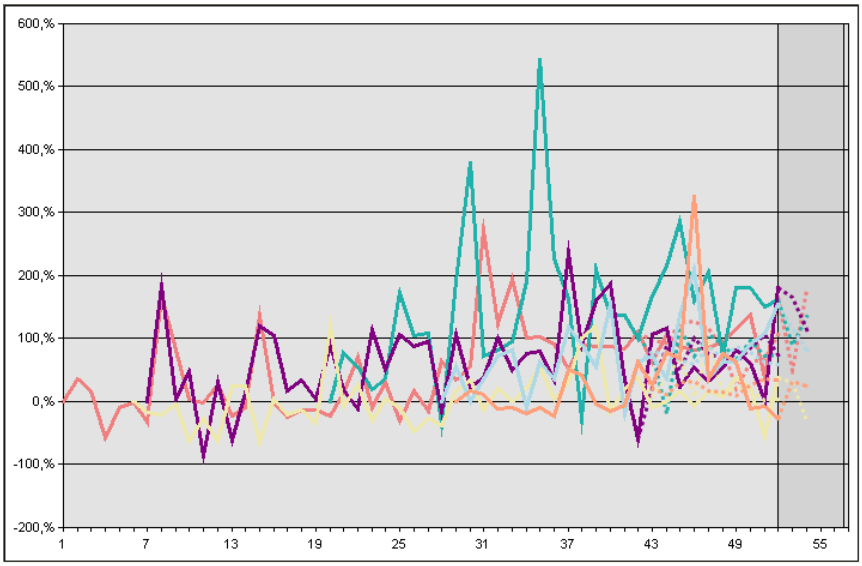


Figura 16: Resultados obtidos com o modelo de séries temporais

**Avaliação**

Uma vez que os dados disponíveis eram relativamente poucos, o mecanismo não foi capaz de produzir previsões estáveis para mais do que duas semanas após o período de análise. No entanto, analisando a previsão histórica apresentada pelo algoritmo para dados relativos aos últimos 20% da escala temporal, foi possível analisar se a previsão de facto se aproxima dos valores reais obtidos. Analisando os resultados, verificou-se que mesmo com poucos dados de treino em cerca de 70% dos casos a previsão dada pelo algoritmo ficou bastante próxima da real obtida (Figura 17).

Noutros casos, no entanto, a variação registada no período de análise não permitiu produzir previsões próximas dos valores reais (Figura 18), o que poderá vir a ser minimizado com a inserção de novos dados.

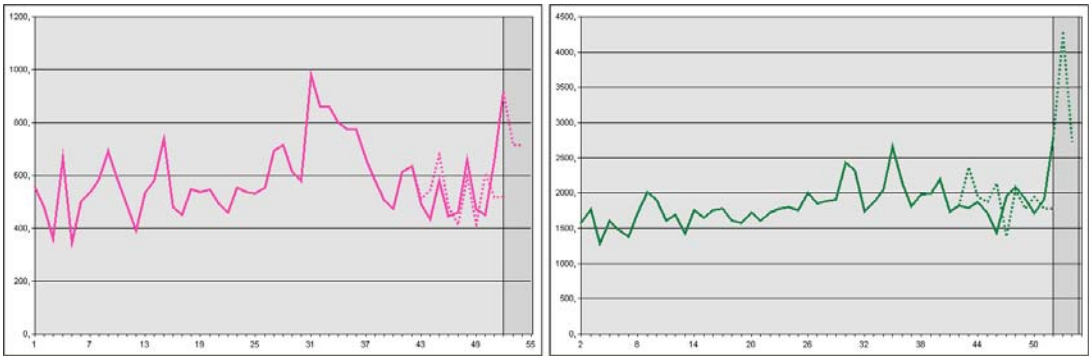

Figura 17: Exemplo de previsões do modelo de séries temporais próximas das reais

Com este desenvolvimento foi possível verificar que o mecanismo se consegue ajustar aos novos dados e apresentar previsões mais próximas das reais para as semanas seguintes. Nas empresas que efectuam a recolha, normalmente a atribuição de trabalho é efectuada semana a semana, pelo que não é de extrema importância que o algoritmo seja capaz de produzir previsões prolongadas no tempo, embora isso possa vir a ser possível com a introdução de novos dados.

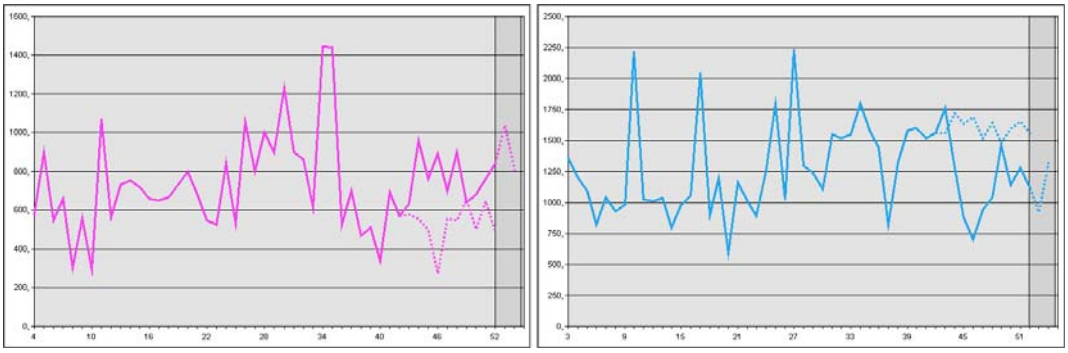

Figura 18: Exemplo de previsões do modelo de séries temporais distantes das reais

No caso estudado os dados existentes não permitiram ao algoritmo avaliar eventuais padrões sazonais, como por exemplo a ocorrência sistemática do aumento da produção junto à costa no período de verão, uma vez que apenas existiam dados coerentes relativos a um ano de recolhas.

Com a inserção de novos dados o mecanismo deverá ser capaz de detectar esses e outros padrões e proporcionar previsões mais aproximadas do estado real.

### 5.2.2 Optimização da frequência de recolhas com base em *clustering*

A recolha dos ecopontos é efectuada durante os cinco dias úteis de cada semana, de segunda a sexta-feira. O número de rotas dificilmente será proporcional ao número de dias da semana, como no caso da Resulima que são 16 rotas para 5 dias. Isto justifica-se pelo facto de umas rotas serem recolhidas com mais frequência de que as outras, devido às características dos contentores que fazem parte das rotas. O problema surge na definição da frequência com que cada rota deverá ser recolhida, essencial para evitar que os ecopontos sejam recolhidos cedo demais, quando não têm produto que justifique a sua recolha e utilizando recursos desnecessários, ou tarde demais, quando o produto existente já se encontra por fora dos contentores ou quando já se registaram queixas das populações.

Os mecanismos de mineração de dados de *clustering* podem auxiliar à definição do dia da semana "ideal" para recolher cada uma das rotas, com base no seu histórico. Estes mecanismos têm a capacidade de encontrar grupos entre os dados, que por vezes não são evidentes [Tange & MacLennan, 2005]. Os mecanismos de *clustering* normalmente são utilizados para segmentar os elementos num número desconhecido de grupos, neste caso sabia-se à partida que esse número varia entre 1 e 5 (um para cada dia útil da semana).

#### Compreensão dos dados

A ferramenta SPAR permite registar todos os dados relativos às recolhas efectuadas, incluindo os contentores que foram recolhidos, o respectivo estado de enchimento e a indicação se foi ou não recolhido. Normalmente em cada ecoponto é recolhido apenas um dos produtos (ou seja, todos os contentores desse produto), mas o SPAR permite que sejam registados também os enchimentos

verificados nos restantes contentores. Estes dados são muito valiosos para quem gere o negócio, pois permitem-lhes ter os dados mais actualizados e mais dados estatísticos para basearem as suas decisões. Todos os dias de semana do ano, excepto feriados, originaram registos de recolhas, embora os diferentes contentores do sistema não possuam registos para todos estes dias. Quando os registos não existem, será necessário decidir por qual das seguintes soluções optar:

- Ignorar o enchimento dos contentores para os dias em que ele não foi registado, diminuindo a eficácia da mineração pois existiriam muitos espaços por preencher entre os dados.
- Preencher os valores dos enchimentos para os dias em falta com um valor estimado, que permita uma análise mais correcta, com base no enchimento dos contentores em todos os dias da análise.

Estes dados são registados nas tabelas TbTurno, TbMovimento e TbLinhasMovimento. Na tabela TbTurno é registada a data de partida do turno de trabalho, que pode servir de referência para a data de recolha dos contentores, a tabela TbMovimento contém a rota respectiva e a tabela TbLinhasMovimento contém cada um dos contentores e o enchimento registado. Na análise destes dados não foram encontrados dados incoerentes, embora não haja certeza absoluta de que eles estão correctos, pois a sua inserção, através do PDA, está sujeita a erros por parte do elemento da equipa de recolha que os está a registar.

### Preparação dos dados

A preparação dos dados para que pudessem ser submetidos a um mecanismo de *clustering* começou pela criação de uma vista, contendo apenas os dados necessários, denominada vEnchimentos, que continha a data de partida da rota, o código da rota, o contentor, o enchimento e o registo de ter ou não sido recolhido (Figura 19).

| DataPartida | CodRota | CodContentor | Enchimento | Recolhido |
|---|---|---|---|---|
| 02-01-2007 | 15 | 02662 | 13 | False |
| 02-01-2007 | 15 | 02680 | 38 | True |
| 02-01-2007 | 15 | 02698 | 13 | False |
| 02-01-2007 | 15 | 02701 | 38 | True |
| 02-01-2007 | 15 | 02716 | 88 | True |
| 02-01-2007 | 15 | 02722 | 88 | True |
| 02-01-2007 | 15 | 02731 | 13 | False |
| 02-01-2007 | 15 | 02737 | 38 | True |
| 02-01-2007 | 15 | 02740 | 63 | True |
| 02-01-2007 | 15 | 02746 | 38 | True |
| ... | ... | ... | ... | ... |

Figura 19: Dados da vista vEnchimentos

Foi também criada uma vista que permitia identificar todos os contentores do sistema, vContentores, contendo apenas uma coluna com o código do contentor (Figura 20).

| CodContentor |
|---|
| 00004 |
| 00010 |
| 00013 |
| 00019 |
| 00022 |
| 00034 |
| 00037 |
| 00040 |
| 00043 |
| ... |

Figura 20: Dados da vista vContentores

Estas tabelas contêm todos os registos efectuados para todos os contentores, incluindo o enchimento que o contentor registava na altura. Com base nesses enchimentos deverá ser possível calcular a melhor altura para recolher cada um dos contentores, tendo em atenção os dias da semana em que eles mais vezes estiveram cheios, em termos históricos. Fazendo essa previsão

para cada um dos contentores será possível inferir as rotas que deverão ser recolhidas em cada um dos dias da semana, com o objectivo de optimizar o processo da recolha dos ecopontos.

## Modelação

Para este caso irá ser utilizado o mecanismo de mineração de dados Microsoft Clustering Algorithm, um dos algoritmos disponíveis para agrupar itens semelhantes. Para a utilização deste algoritmo foi criada uma tabela para suportar os dados de análise, M_EnchimentoContentores, contendo as seguintes colunas:

- IDData – identificador da data da recolha (entre 1 e 365, correspondendo aos 365 dias do ano de 2007).
- IDDiaSemana – identificador do dia da semana correspondente (entre 1 e 5, correspondendo aos dias entre segunda e sexta-feira).
- Rota – rota recolhida.
- Contentor – contentor registado.
- Enchimento – enchimento associado ao contentor, na data respectiva.
- Recolhido – indicador de recolha do contentor (verdadeiro ou falso, consoante foi ou não recolhido).

Conforme referido anteriormente, não existem dados concretos sobre o enchimento dos contentores para todos os dias do ano, pelo que se optou por preencher os dados intermédios com um valor médio, cujo cálculo é descrito em seguida.

Para o preenchimento da tabela M_EnchimentoContentores foi desenvolvida uma aplicação em VB.NET 2005 para efectuar a transformação dos dados, que contempla também o cálculo do valor médio do enchimento dos contentores. Uma vez que o número de dados existentes na tabela da análise era muito elevado (3323 contentores a multiplicar pelos 365 dias do ano correspondem a 1212895 registos), optou-se por reduzir a análise a apenas um dos produtos, neste caso o papel,

reduzindo significativamente o volume de dados a manipular pela aplicação de transformação (1115 contentores a multiplicar por 365 dias correspondem a 406975 registos), reduzindo-se assim o tempo necessário para efectuar esta transformação.

Esta função começou por carregar para a tabela M_EnchimentoContentores todos os dias possíveis do ano para todos os contentores, preenchendo o campo IDData, IDDiaSemana e o campo Contentor, deixando os restantes com valores por defeito para serem preenchidos posteriormente. Seguidamente, com base nos dados obtidos da vista vEnchimentos, esta tabela foi actualizada com os enchimentos registados para cada um dos contentores e a indicação se foram recolhidos. As linhas na tabela M_EnchimentoContentores onde estes dados foram preenchidos foram aquelas que continham dados para as datas e os contentores respectivos. Finalmente foi necessário preencher os "espaços em branco", com enchimentos aproximados aos reais, efectuando os seguintes passos:

- foram analisados todos os períodos em que existia uma falha, contando-se o número de dias do período em que existia essa falha;
- foram analisados os enchimentos existentes nos dias imediatamente anterior e imediatamente a seguir, encontrando-se uma variação de enchimento (de referir que se no dia imediatamente anterior o contentor tivesse sido recolhido, o enchimento tomado como referência foi 0, pois se o contentor foi recolhido ficou sem qualquer enchimento);
- foi calculado o valor de enchimento incremental por dia, que consistiu em dividir a variação do enchimento pelo número de dias passados;
- para cada um dos dias em falha, acrescentou-se ao valor anterior (começando pelo valor do dia imediatamente após o primeiro) o valor de enchimento incremental, sucessivamente até todos os espaços estarem preenchidos.

Após a execução desta transformação a tabela passou a conter os dados dos enchimentos de todos os contentores do sistema, ao longo dos 365 dias do ano. Para se poder analisar em que dias os contentores devem ser recolhidos e, respectivamente, em que dias as rotas deverão ser recolhidas,

foi necessário eliminar os dias em que os enchimentos dos contentores não justificavam a sua recolha. Usualmente um contentor é recolhido sempre que o seu enchimento seja igual ou superior a 75% da sua capacidade. Para obter estes dados foi criada uma vista, vEnchimentoContentores, que filtra os dados presentes na tabela M_EnchimentoContentores para conter apenas aqueles em que a coluna Enchimento tinha um valor maior ou igual a 75, ou seja, contendo todas as ocorrências de enchimento igual ou superior a 75% para todos os contentores, que estão associados à rota a que pertencem.

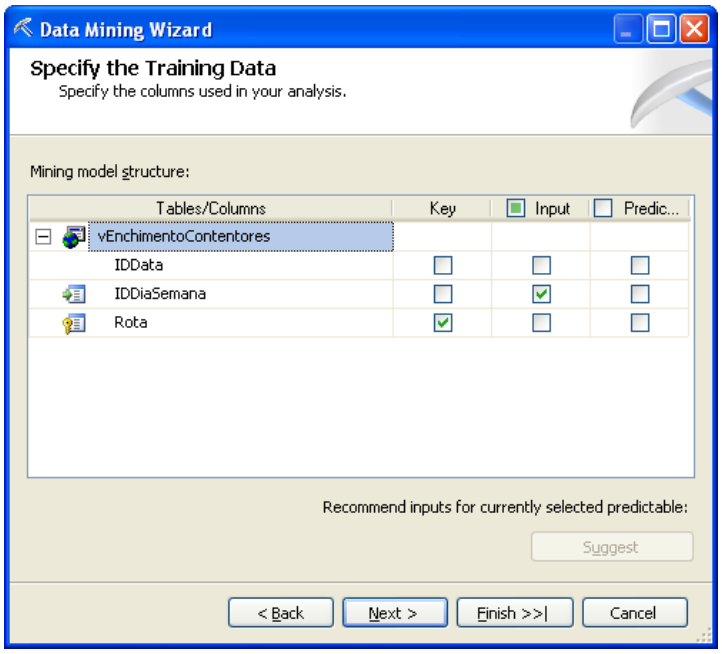


Figura 21: Configuração dos dados de treino para o modelo de *clustering*

Finalmente foi criado o projecto de mineração de dados, tendo sido escolhida a técnica de mineração Microsoft Clustering. Como fonte de dados para este projecto foi seleccionada a vista EnchimentoContentores, previamente criada no projecto (que corresponde à vista vEnchimentoContentores previamente criada na base de dados). Nos tipos de tabelas foi seleccionada a vista EnchimentoContentores como *Case table*.

Os dados de treino foram configurados da seguinte forma (Figura 21): a coluna Rota é uma chave, que irá ser usada como identificador das rotas nos resultados obtidos; a coluna IDDiaSemana é uma entrada e é com base neste valor que as rotas irão ser segmentadas.

Com o volume de dados existente o algoritmo demorou cerca de 3 segundos a processar. O mecanismo de Microsoft Clustering produz, entre outros, um diagrama de grupos (*Cluster Diagram*) contendo todos os grupos encontrados e, dentro de cada grupo, os elementos que lhe pertencem (Figura 22).

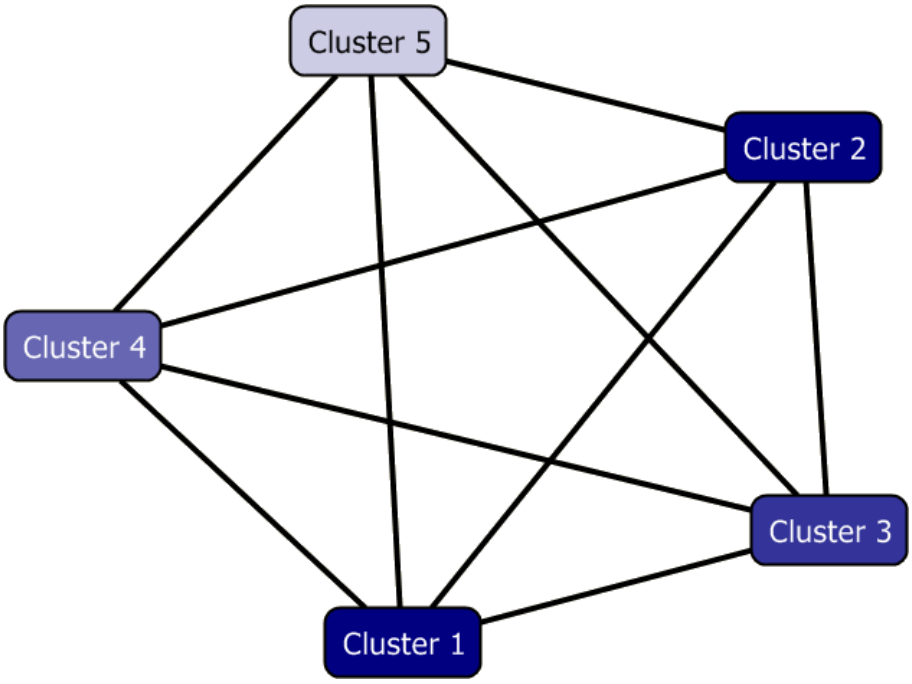


Figura 22: Resultados obtidos com o modelo de *clustering*

Analisando os dados presentes nos grupos é possível renomeá-los para serem mais facilmente visualizados. Neste caso, cada grupo continha uma tabela com as duas colunas adicionadas na modelação, IDDiaSemana e Rota (Figura 23).

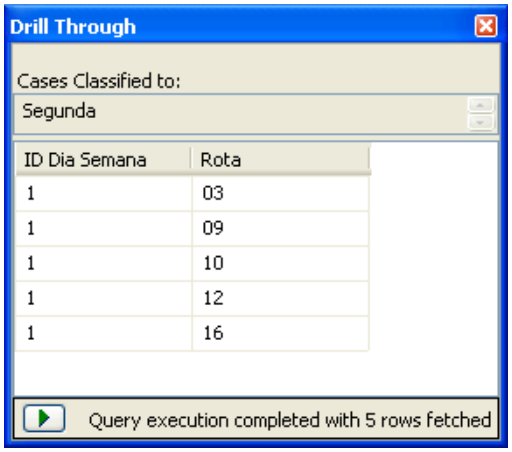


Figura 23: Elementos de um cluster do modelo de *clustering*

Cada grupo correspondia a um dia de semana específico, pelo que foi possível renomear os grupos com cada um dos dias úteis da semana (Figura 24).

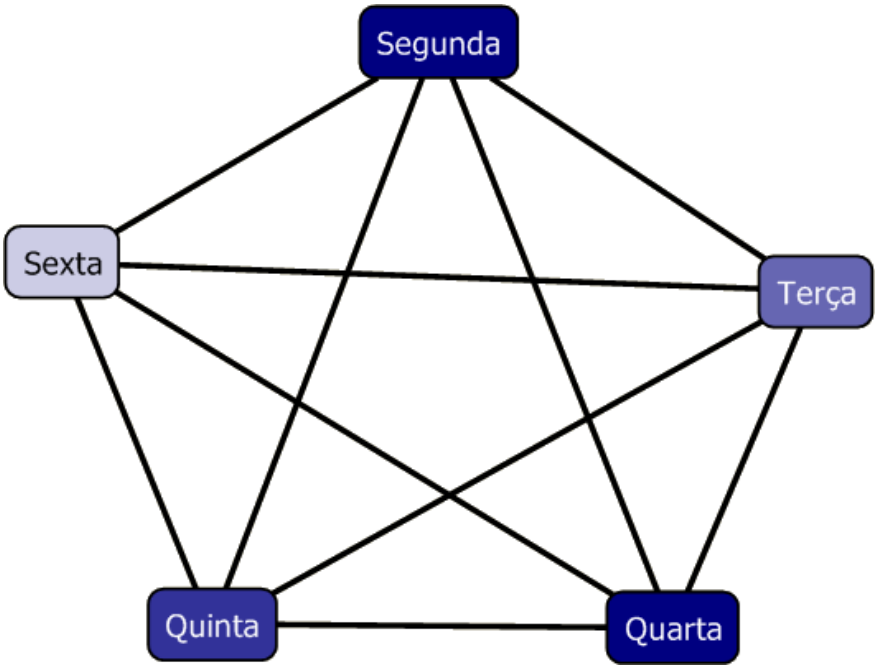


Figura 24: Resultado da renomeação dos *clusters*

## Avaliação

A aplicação do mecanismo de Microsoft Clustering aos dados de análise resultou na definição dos dias ideais para recolher cada uma das rotas, isto é, os dias em que mais vezes a soma dos ecopontos de cada uma das rotas foram registados como estando "cheios", ou seja, com enchimento superior a 75% (Tabela 4).

| **Dia da semana** | **Rotas** |
|---|---|
| Segunda-feira | 03 |
| | 09 |
| | 10 |
| | 12 |
| | 16 |
| Terça-feira | 08 |
| | 11 |
| Quarta-feira | 04 |
| | 06 |
| | 14 |
| | 15 |
| Quinta-feira | 01 |
| | 05 |
| | 07 |
| | 13 |
| Sexta-feira | 02 |

Tabela 4: Sugestão de organização de rotas pelo modelo de *clustering*

Com a análise destes dados verificou-se que o dia em que mais rotas deveriam ser recolhidas era segunda-feira. Esta previsão faz algum sentido uma vez que durante os dois dias do fim-de-semana os ecopontos não são recolhidos, sendo natural que o enchimento dos contentores à segunda-feira seja superior. No entanto, a distribuição das rotas não está adequada às necessidades do sistema, uma vez que o número de rotas a recolher por cada dia deverá ser relativamente igual para cada um dos dias (neste caso, 16 rotas a dividir por 5 dias deveriam dar origem a cerca de 3,2 rotas por dia).

Na análise efectuada foram apenas consideradas as rotas de um dos produtos, papel, pelo que as distribuições das rotas podem vir a ser mais equilibradas se forem adicionadas aos dias com menos rotas de papel rotas de outros produtos, produzindo assim um plano de trabalho semanal com um número de rotas semelhantes em cada um dos dias (neste caso 16 rotas a multiplicar pelos 3 produtos e a dividir pelos 5 dias da semana originam cerca de 9,6 rotas por dia).

Nos dados resultantes da mineração de dados não se verificou nenhum caso de uma rota que estivesse presente, simultaneamente, em mais do que um grupo. Havendo mais dados para análise essa situação poderia acontecer e, nesse caso, seria poderia optar-se por recolher a rota num ou no outro dia, alargando as possibilidades na escolha do plano semanal de recolhas, ou por recolher a rota em ambos os dias, caso se verifique que os seus contentores enchem com muita frequência.

### 5.2.3 Análise crítica da aplicação de mineração de dados na recolha selectiva

Os mecanismos de mineração de dados utilizados nos dois modelos elaborados, em contextos diferentes de aplicação, revelaram-se eficientes na análise da informação e na apresentação de resultados, pois embora na presença de uma quantidade considerável de dados processaram toda a informação em tempo reduzido.

Com os testes efectuados foi possível confirmar que a mineração de dados, devidamente aplicada, poderá constituir uma vantagem para a gestão dos processos da recolha selectiva de resíduos, uma

área com grandes necessidades de optimização e grande carência de ferramentas de apoio à decisão. A informação gerada neste negócio é bastante rica, deverá possibilitar a descoberta de conhecimento essencial à optimização de processos e diminuição de recursos, podendo por isso existir outras áreas em que a mineração de dados pode auxiliar à tomada de decisão.

A modelação dos mecanismos apresenta algumas dificuldades, nomeadamente na definição dos dados necessários para se obterem os resultados pretendidos. A mineração de dados oferece a possibilidade de extrair informação desconhecida entre os dados, o que é claramente um benefício no processamento de grandes quantidades de informação, no entanto, se os mecanismos não forem correctamente alimentados e configurados a informação resultante poderá não ter nenhum sentido prático, ou poderá não fazer mais do que permitir constatar o que é evidente, o que não representa informação de utilidade para a gestão de um negócio.

A metodologia CRISP-DM, utilizada no desenvolvimento deste projecto, permitiu minimizar os problemas com a construção dos modelos, pois define claramente as etapas a seguir de forma a permitir a correcta implementação do modelo. Esta metodologia, simples de implementar, permite evitar os erros associados a uma compreensão deficiente do negócio e dos dados disponíveis para análise.

# Capítulo 6

# Conclusões e Trabalho Futuro

## 6.1 Conclusões

O objectivo deste trabalho foi o de aplicar técnicas de mineração de dados na previsão da frequência de recolhas na área da recolha selectiva. Para cumprir este objectivo foram elaborados dois modelos de mineração de dados distintos:

- O primeiro, utilizando séries temporais, para prever as quantidades a recolher no futuro com base nas quantidades recolhidas no passado.
- O segundo, utilizando *clustering*, para determinar os dias ideais para recolher cada uma das rotas com base nos enchimentos dos contentores registados em recolhas anteriores.

Os testes foram efectuados numa base de dados com apenas um ano de utilização, limitando a possibilidade de previsão dos algoritmos. Nos dois casos estudados os resultados obtidos foram os seguintes:

- No algoritmo de séries temporais, o modelo efectuou a previsão dos acontecimentos ocorridos em 20% da escala temporal com base nos 80% anteriores; nas rotas em que as oscilações de pesos foram menos acentuadas as previsões ficaram bastante aproximadas dos valores reais; também foi possível analisar a adaptação da previsão aos dados, por exemplo se nos últimos períodos a tendência de recolha era de descida, o mecanismo mostrava uma tendência decrescente.
- No algoritmo de *clustering*, o modelo identificou claramente o dia de segunda-feira como o mais adequado para recolher grande parte das rotas, o que faz sentido devido à ausência de recolhas nos dois dias anteriores; este modelo permitiu verificar que, no caso analisado, as rotas são independentes, isto é, cada uma tem o seu dia ideal para recolha, o que, aplicado a um caso real, poderá ser informação valiosa para a gestão eficaz das recolhas.

Os resultados obtidos com este trabalho permitem identificar uma forte possibilidade de aplicação da mineração de dados em contexto real, utilizando estes ou outros modelos, que poderão trazer verdadeiros benefícios para as empresas no que diz respeito ao planeamento adequado das operações, à optimização dos recursos e à prestação de um bom serviço ao cidadão.

## 6.2 Trabalho futuro

Uma vez que os modelos elaborados produziram resultados indicativos de que a mineração de dados é aplicável em casos reais deste contexto, como continuação do trabalho efectuado os modelos deverão ser melhorados de modo a melhor se adaptarem às necessidades de cada empresa em particular.

Uma vez aplicados, os algoritmos deverão ser monitorizados e adaptados, de modo a produzirem melhores resultados. À medida que vão sendo gerados mais dados de processo os algoritmos deverão apresentar informação mais correcta, mas poderá haver a necessidade de reconfigurar alguns parâmetros no seu funcionamento.

A base de dados utilizada nos testes efectuados tinha dados de apenas um ano completo de utilização, o que não permitiu tirar partido de uma das possibilidades da mineração de dados, de elevada importância nesta área: a determinação de padrões sazonais. Deverão por isso ser construídos novos modelos que, aplicados a um maior conjunto de dados, poderão permitir determinar esses padrões, fundamentais numa área de negócio tão influenciada pelos contextos sociais.

Também deverá ser possível afinar a previsão dos dados se forem adicionadas mais variáveis ao processo, como condições meteorológicas, feriados e festas locais, entre outros, que têm uma influência muito grande na deposição de resíduos recicláveis nos ecopontos.

Finalmente, deverá ser analisada a possibilidade de adaptar estas técnicas a outras áreas semelhantes à recolha selectiva de resíduos, como a recolha de óleos alimentares usados ou de biomassa, que partilham muitas dificuldades e necessidades com a área de negócio aqui analisada.

# Bibliografia

[Berry & Linoff, 2004] Michael J.A. Berry e Gordon S. Linoff: "Data Mining Techniques For Marketing, Sales, and Customer Relationship Management, Second Edition". Wiley Publishing, 2004

[Bramely & Simchi-Levi, 1993] Julien Bramely e David Simchi-Levi: "A Location Based Heuristic for General Routing Problems". Graduate School of Business e Department of Industrial Engineering and Operations Research, Columbia University, NY, 1993

[Crows, 1999] Two Crows Corporation: "Introduction to Data Mining and Knowledge Discovery Third Edition". 1999

[Fayyad et al., 1996] Usama Fayyad, Gregory Piatetsky-Shapiro e Padhraic Smyth: "Knowledge Discovery and Data Mmining - Towards a Unifying Framework". Second International Conference on Knowledge Discovery and Data Mining, Portland, Oregon, 1996

[Hand et al., 2001] David Hand, Heikki Mannila e Padhraic Smyth: "Principles of Data Mining". The MIT Press, 2001

[Hu & Huang, 2007] Xiangpei Hu and Minfang Huang: “An Intelligent Solution System for a vehicle Routing Problem in Urban Distribution”

[Kimball et al., 1998] Ralph Kimball, Laura Reeves, Margy Ross, Warren Thornthwaite: “The Data Warehouse Lifecycle Toolkit – Expert Methods for Designing, Developing and Deploying Data Warehouses”. John Wiley & Sons Inc., 1998

[Marković et al., 2005] Hrvoje Marković, Ivana Ćavar e Tonči Carić: “Using Data Mining to Forecast Uncertain Demands in Stochastic Vehicle Routing Problem”. Faculty of Electrical Engineering and Computing e Faculty of Transport and Traffic Engineering, Zagreb, Croatia, 2005

[Oracle 2005] Oracle Corporation: “Forecasting Demand: Monitoring Demand Leads to More Profitable Decision-Making”. Oracle, 2005

[Rainardi, 2008] Vincent Rainardi: “Building a Data Warehouse – With Examples in SQL Server”. Apress, 2008

[Respício et al., 2002] A. Respício, M. E. Captivo e A. J. Rodrigues, “A DSS for Production Planning and Scheduling in the Paper Industry”. International Conference on Decision Making and Decision Support in the Internet Age, University College Cork, Cork, Ireland, 2002

[Savelsbergh, 2002] Martin Savelsbergh: “Vehicle Routing and Scheduling”. IMA Thematic Year on Optimization – Supply Chain and Logistics Optimization, September-December 2002

[Sumathi & Sivanandam, 2006] S. Sumathi e S. N. Sivanandam: “Introduction to Data Mining and its Applications”. Springer, 2006

[Tange & MacLennan, 2005] ZhaoHui Tang e Jamie MacLennan: “Data Mining with SQL Server 2005”. Wiley, 2005

# Referências WWW

[WWW ADP] http://www.adp.pt

Página do grupo Águas de Portugal (ADP), um grupo empresarial português na área do ambiente com o objectivo de "contribuir para a resolução dos problemas nacionais nos domínios de abastecimento de água, de saneamento de águas residuais e de tratamento e valorização de resíduos, num quadro de sustentabilidade económica, financeira, técnica, social e ambiental". Acedida a 12 de Outubro de 2008.

[WWW AUTOOPT] http://www.scai.fraunhofer.de/293.html?&L=1

Página com informações sobre a ferramenta AUTO-OPT, de apoio à indústria automóvel. Acedida a 27 de Novembro de 2008.

[WWW Biodiesel] http://pt.wikipedia.org/wiki/Biodiesel

Esta página faz parte da enciclopédia livre Wikipedia e fornece informações sobre o Biodiesel, o seu processo de fabrico, entre outras informações. Acedida a 14 de Setembro de 2008.

[WWW Cachapuz] http://www.cachapuz.com

Página da empresa Cachapuz - Equipamentos para Pesagem, Lda., com informações sobre a empresa e os seus produtos. Acedida a 28 de Setembro de 2008.

[WWW CRISPDM] http://www.crisp-dm.org

Página do projecto CRISP-DM (CRoss Industry Standard Process for Data Mining), que desenvolveu uma metodologia de trabalho para o desenvolvimento de projectos de mineração de dados, independente da indústria e da ferramenta. Acedida a 28 de Setembro de 2008.

[WWW DemandForecasting] http://en.wikipedia.org/wiki/Demand_Forecasting

Esta página faz parte da enciclopédia livre Wikipedia e fornece informações sobre a previsão da procura. Acedida a 27 de Novembro de 2008.

[WWW EcologicalFootprint] http://en.wikipedia.org/wiki/Ecological_footprint

Esta página faz parte da enciclopédia livre Wikipedia e fornece informações sobre o conceito de Pegada Ecológica. Acedida a 14 de Setembro de 2008.

[WWW ESSE] http://esse.wdcb.ru

Página do projecto Environmental Scenario Search Engine (ESSE), que aplica técnicas de mineração de dados para apoio à descoberta de informação em arquivos de dados de informação ambiental. Acedida a 27 de Novembro de 2008.

[WWW GlobalFootprint] http://www.footprintnetwork.org

Página do projecto Global Footprint Network, uma iniciativa criada para "proporcionar um futuro sustentável em que todas as pessoas têm a oportunidade de viver vidas satisfatórias dentro das capacidades de um planeta". Acedida a 21 de Setembro de 2008.

[WWW InformationSociety] http://en.wikipedia.org/wiki/Information_society

Esta página faz parte da enciclopédia livre Wikipedia e fornece informações sobre o conceito de Sociedade da Informação, a sua definição e algumas considerações. Acedida a 21 de Setembro de 2008.

[WWW MSDN] http://msdn.microsoft.com/en-us/library/ms175595(SQL.90).aspx

Página da empresa Microsoft com informações sobre os algoritmos de mineração de dados disponíveis na ferramenta SQL Server 2005 Analysis Services (SSAS) . Acedida a 21 de Setembro de 2008.

[WWW PontoVerde] http://www.pontoverde.pt

Página da Sociedade Ponto Verde S. A., uma entidade privada sem fins lucrativos com a missão de promover a recolha selectiva, a retoma e a reciclagem de resíduos no território português. Acedida a 14 de Setembro de 2008.

[WWW TheEarling] http://www.thearling.com

Página com informação sobre mineração de dados e tecnologias analíticas. Acedida a 21 de Setembro de 2008.

[WWW VehicleRoutingProblem] http://en.wikipedia.org/wiki/Vehicle_routing_problem

Esta página faz parte da enciclopédia livre Wikipedia e fornece informações sobre o vehicle routing problem (VRP), um problema proposto por Dantzig and Ramser em 1959. Acedida a 22 de Novembro de 2008.

[WWW WilliamRees] http://en.wikipedia.org/wiki/William_Rees_(academic)

Esta página faz parte da enciclopédia livre Wikipedia e fornece informações sobre William E. Rees, professor na University of British Columbia (UBC) e director da School of Community and Regional Planning (SCARP) na UBC. Acedida a 21 de Setembro de 2008.