Deteção de *outliers* e previsão de vendas numa empresa de distribuição farmacêutica em Portugal

Augusto Carlos Pereira Alves Ribeiro

Tese de Mestrado em Informática Especialização em Engenharia de Software

Orientação:

Professora Doutora Natércia Felgueiras Seabra Durão

Professora Doutora Maria Isabel Calapez Cabrita Leal Seruca

Janeiro, 2016

Agradecimentos

Às minhas orientadoras, Professora Doutora Natércia Durão e Professora Doutora Isabel Seruca, pela sua orientação, conselhos prestados, e acima de tudo, pelo apoio e incentivo, fundamental para a elaboração desta dissertação e para o resultado final deste projeto.

Aos Professor Doutor Fernando Moreira, Professora Doutora Filomena Lopes, Professora Doutora Maria João Ferreira, Dr. Jorge Pereira, Dr. João Carlos Monteiro e Dr. António Sá.

À minha família.

Aos meus colegas de mestrado Luís, Rui, Marco e Ana.

À Administração da “minha” empresa.

A todos, que participaram direta ou indiretamente na realização deste projeto e que contribuíram para o sucesso do trabalho final.

A todos o meu sincero obrigado.

"Não sabendo que era impossível, ele foi lá e fez."

Mark Twain / Jean Cocteau

Resumo

Deteção de *outliers* e previsão de vendas numa empresa de distribuição farmacêutica em Portugal

As ruturas de *stock* no abastecimento de medicamentos a farmácias podem ter origem em diversos fatores nomeadamente problemas fabris, falta de matéria-prima, fim de comercialização de produtos, surtos de doenças e epidemias. A estes fatores acresce a venda de medicamentos por parte de algumas farmácias a mercados externos, que tem aumentado nos últimos anos, e é considerada umas das principais causas das falhas de abastecimento de medicamentos em Portugal.

Este trabalho retrata o caso de estudo de uma empresa de distribuição farmacêutica em Portugal e tem como objetivo dar resposta a dois problemas essenciais. O primeiro consitiu na deteção de clientes (farmácias) e produtos (medicamentos) que podem ser considerados *outliers* e no rateamento de stock quando esses *outliers* são detetados, a fim de evitar a venda anormal e a rutura de *stocks* nas farmácias. O segundo consistiu em efetuar uma previsão de vendas da empresa de distribuição farmacêutica, para um melhor controlo e gestão dos níveis de *stocks* de medicamentos, de forma a evitar custos excessivos e simultaneamente garantir uma satisfação adequada dos clientes, diminuindo a possibilidade de perda de clientes devido a falhas de *stock*.

Na deteção de *outliers* (clientes e produtos) foram usados os métodos *Box-plot* e *Z-score* modificado e utilizado o *software* estatístico SPSS. O método de *data mining* de séries temporais Pegels amortecido foi utilizado no cálculo da previsão de vendas e a implementação foi feita em SQL, estando os dados a analisar armazenados numa base de dados ORACLE.

Palavras-chave: *outliers*, medicamentos, rutura *stock*, *data mining*, séries temporais, previsão de vendas

Abstract

Detection of outliers and sales prediction for a pharmaceutical distribution company in Portugal

Stock unavailability in the supply of medicines to pharmacies can be caused by several factors including manufacturing problems, lack of raw materials, end of product selling, disease and epidemics outbreaks. Furthermore, the sale of medicines by some pharmacies to foreign markets has increased in recent years, and is considered one of the main causes of medicine supply failures in Portugal.

This thesis depicts the case study of a pharmaceutical distribution company in Portugal and aims to address two main research issues. The first one consisted in detecting customers (pharmacies) and products (medicines) which may be considered outliers and perform stock proration when these outliers are detected, in order to avoid abnormal sales and out-of-stocks in pharmacies. The second one targeted the sales prediction for the pharmaceutical distribution company, in order to better control and manage the levels of stock of medicines, so as to avoid excessive inventory costs while guaranteeing customer demand satisfaction, and thus decreasing the possibility of loss of customers due to stock outages.

In outliers detection (customers and products) we used the *Box-plot* and *modified Z-score* methods as well as the SPSS statistical software. For sales prediction, the time series data mining method smoothed Pegels was used, while the implementation was done in SQL and the analyzed data was stored in an Oracle database.

Keywords: outliers, medicines, stock unavailability, data mining, time series, sales prediction

Índice

Índice de Figuras

Índice de Tabelas

Índice de Equações

1. Introdução

1.1. Contextualização

Uma das responsabilidades dos distribuidores grossistas de medicamentos em Portugal é serem obrigados por lei[1] a ter um *stock* mínimo de medicamentos, de modo a garantir o abastecimento no mercado nacional e, desta forma, evitar possíveis situações de rutura nas farmácias.

Segundo o Infarmed - Autoridade Nacional do Medicamento e Produtos de Saúde I.P. (Infarmed, 2012), as ruturas de *stock* de medicamentos, ocasionadas quando não existe quantidade disponível de determinados medicamentos para satisfazer os pedidos dos clientes (farmácias), podem ter como origem diversos fatores, tais como: problemas fabris, falta de matéria-prima, fim de comercialização de produtos, surto de doenças, epidemias, etc.

Para além destes fatores, a venda de medicamentos por parte de algumas farmácias a mercados externos tem aumentado nos últimos anos, e é considerada umas das principais causas das falhas de abastecimento de medicamentos em Portugal, conforme notícias vindas a público em diversos meios de comunicação social (Jornal Público, 2013).

Associadas a esta prática de exportação de medicamentos são apontadas as descidas dos preços dos fármacos em Portugal e a difícil situação financeira das farmácias, que têm tornado cada vez mais apetecível e lucrativa a venda de medicamentos para mercados externos.

Para as empresas distribuidoras de medicamentos é, assim, fundamental detetar os clientes (farmácias) e os produtos (medicamentos) *outliers* (valores

[1] As disposições conjugadas da alínea d) do n.º 1 do artigo 29.º com a da alínea c) do n.º 1 do artigo 100.º do Decreto-Lei n.º 176/2006, de 30 de Agosto, consagram para os titulares de autorização de introdução no mercado de medicamentos e para os distribuidores por grosso dos mesmos produtos uma obrigação particular na mesma matéria, que consiste em dispor permanentemente de medicamentos em quantidade e variedade suficientes para garantir o fornecimento adequado e contínuo do mercado geográfico relevante, de forma a garantir a satisfação das necessidades dos doentes.

que apresentam um grande afastamento ou são inconsistentes com os restantes) e ratear (dividir proporcionalmente) o *stock* quando esses outliers são detetados, no sentido de impedir a venda anormal e de evitar rutura de stocks nas farmácias.

Acresce a esta necessidade, o desfasamento entre a periocidade das entregas dos medicamentos nas farmácias, que podem ter várias entregas por dia e, a obtenção de *stock* por parte dos distribuidores, que pode demorar cerca de dois dias.

Por outro lado, para as distribuidoras farmacêuticas é essencial conseguir obter uma boa previsão das necessidades de medicamentos, devido ao curto prazo de validade de muitos medicamentos e à necessidade de controlar os níveis de *stock*, de forma a evitar custos excessivos de stock e simultaneamente a perda de clientes devido a falhas de stock.

Uma boa previsão de vendas está, geralmente, associada a conseguir obter um bom equilíbrio entre os custos de *stock* e uma adequada satisfação da procura dos clientes (Gupta, Maranas, & McDonald, 2000). Para o caso específico da indústria de distribuição farmacêutica, o problema adquire particular importância devido ao curto ciclo de vida de muitos dos produtos e da importância da qualidade dos produtos que está, por sua vez, fortemente ligada a aspetos de saúde pública (Doganis, Alexandridis, Patrinos, & Sarimveis, 2006; Zadeh, Sepehri, & Farvaresh, 2014).

1.2. Objetivos do estudo

Tendo em conta as já referidas dificuldades sentidas pelas empresas de distribuição farmacêutica, este trabalho pretende dar respostas às seguintes questões de investigação, utilizando no estudo efetuado, os dados de uma empresa de distribuição farmacêutica em Portugal:

(Q1) Será possível usando a deteção de *outliers*, controlar o *stock* mínimo de medicamentos, por parte das empresas de distribuição farmacêutica, de forma a impedir ruturas no abastecimento normal do mercado (farmácias)?

(Q2) Será possível utilizando um método de *data mining* para previsão de vendas, prever, de uma forma fiável, os valores a encomendar de cada produto, pelas empresas de distribuição farmacêutica, de forma a obter um bom controlo dos níveis de *stock* e respetivos custos?

De forma a dar resposta às questões de investigação identificadas foram definidos os seguintes objetivos principais:

(O1) Detetar clientes (farmácias) e produtos (medicamentos) que podem ser considerados *outliers,* (isto é, observações que apresentam um grande afastamento das restantes ou são inconsistentes com as mesmas, pelas quantidades transacionadas) e ratear (dividir proporcionalmente) o *stock* do produto (medicamentos) quando estes são detetados, no sentido de evitar a venda anormal e a rutura de stocks nas farmácias;

(O2) Efetuar uma previsão de vendas para a empresa, para um melhor controlo e gestão dos níveis de *stocks* e satisfação adequada dos clientes (farmácias), diminuindo a possibilidade de falta dos medicamentos.

Por último, de modo a sustentar o estudo desenvolvido, foram ainda estabelecidos objetivos secundários (OS) que permitiram a concretização do trabalho proposto:

(OS1) Caracterizar e rever os métodos de deteção de *outliers* disponíveis;

(OS2) Caracterizar e rever os métodos de *data mining* existentes.

1.3. Metodologia Adotada

A metodologia adotada para validação das respostas às duas questões de investigação identificadas é o caso de estudo de uma empresa de distribuição farmacêutica em Portugal.

A seleção da metodologia teve em consideração as características do trabalho de investigação a realizar, nomeadamente o âmbito e os objetivos do estudo a desenvolver e de como poderiam ser avaliados os resultados obtidos.

Com vista à deteção dos *outliers* foram analisadas e tratadas as encomendas de três Medicamentos (não sazonais) Sujeitos a Receita Médica (MSRM), efetuadas pela totalidade dos clientes (1325 farmácias) da empresa de distribuição farmacêutica, nos períodos de outubro de 2013 e outubro de 2014. Na deteção de *outliers* (clientes e produtos) foram usados os métodos *Box-plot* e *Z-score* modificado e utilizado o *software* estatístico SPSS (Statistical Package for the Social Sciences, versão 20).

Para a determinação das encomendas da empresa de distribuição farmacêutica foi efetuada a previsão de vendas de 357 medicamentos (de entre um total de 20000 produtos) comercializados pela empresa para o mês atual à data da realização deste trabalho (janeiro de 2016) e os dois meses posteriores (fevereiro e março de 2016). A previsão foi realizada com base na análise do histórico das quantidades desses medicamentos pedidas pelos clientes entre janeiro de 2014 e dezembro de 2015 (24 meses). O método de Pegels amortecido (Taylor, 2003) foi utilizado no cálculo da previsão e a implementação foi feita em SQL, estando os dados a analisar armazenados numa base de dados ORACLE versão 11.2.

1.4. Estrutura do Trabalho

O presente documento encontra-se organizado em 6 capítulos que refletem o percurso do trabalho realizado.

No primeiro capítulo, é feito um enquadramento geral do tema a abordar e é contextualizado o estudo a desenvolver. Neste capítulo são identificadas as duas questões de investigação a responder, os objetivos a atingir e a metodologia a seguir para a concretização do trabalho.

O segundo capítulo é dedicado à análise de dados com deteção de *outliers,* apresentando o conceito de *outlier* e uma revisão dos métodos mais relevantes de identificação de *outliers*.

O terceiro capítulo visa a caracterização dos conceitos associados à atividade de *Data Mining*, incluindo uma panorâmica das categorias, técnicas e principais métodos utilizados.

No capítulo quatro é efetuada a aplicação prática da deteção de *outliers* (clientes e produtos), comparando os resultados obtidos com a aplicação dos métodos de identificação Box-plot e Z-score modificado, implementados no SPSS. Neste capítulo, é ainda criada uma regra prática para melhor discriminar os *outliers* severos identificados na empresa de distribuição farmacêutica.

No capítulo cinco é apresentada a previsão de vendas efetuada para a empresa, enquadrando-se o problema e apresentando-se a fundamentação para o método de *data mining* de previsão utilizado. São ainda apresentados e interpretados os resultados obtidos.

No último capítulo são revisitadas as questões de investigação colocadas e são feitas considerações aos resultados obtidos. São ainda apresentadas sugestões de trabalho futuro, que permitam dar continuidade ao trabalho desenvolvido.

2. Análise de dados com deteção de outliers

O processo de análise dos dados está dividido em duas etapas: a análise preliminar e a análise do comportamento da série de dados (conjunto de resultados observados numa determinada sequência). A primeira etapa, está relacionada com a correção dos dados, de onde são retirados os valores atípicos (*outliers*). A segunda etapa está relacionada com a análise dos padrões inerentes à série de dados.

Este capítulo de índole teórico é dedicado apenas à análise preliminar dos dados. Para tal, apresentam-se algumas definições de *outliers* (Secção 2.1) e os métodos mais usuais de identificação de *outliers* (Secção 2.2).

2.1. Definição de *outlier*

As séries de dados históricos podem sofrer influências de eventos não usuais e não repetitivos (Chen & Lon-Mu, 1993): os *outliers*. Podem-se identificar dois tipos de *outliers* (Tolvi, n.d.): os erros grosseiros e os "verdadeiros" *outliers*. Os primeiros estão associados a erros de processamento, como é o caso de, por exemplo, a ocorrência de um erro no registo de uma venda que deve ser corrigido quando detetado. No caso dos "verdadeiros" *outliers*, após investigada a sua origem deve ser realizada uma das 3 opções: substituição do valor do *outlier* pela previsão; substituição pelo valor médio das observações imediatamente adjacentes; marcado para o futuro (caso se trate de uma campanha promocional).

Se as previsões são calculadas com base em séries de dados que incluem *outliers*, estas podem estar comprometidas devido ao impacto destes valores, sendo que a correção destes valores irá, de uma forma geral, melhorar os resultados obtidos nos cálculos das previsões (Duncan, Gorr, & Szczypula, 1998). Para que esta situação seja evitada, os dados devem ser analisados e, caso se detete a presença de um *outlier*, este deve ser substituído por um valor mais adequado e típico.

A análise de observações *outliers* é já um procedimento antigo e data das primeiras tentativas de analisar um conjunto de dados (Hodge & Austin, 2004).

Existem várias definições para *outlier*, entre as quais podemos citar: “An outlying observation, or outlier, is one that appears to deviate markedly from other members of the sample in which it occurs” (Grubbs, 1969).

A definição acima foi modificada por Barnett & Lewis (1994), à qual adicionaram: ”An observation (or subset of observations) which appears to be inconsistent with the remainder of that set of data”.

Outra definição mais recente para *outlier* é:

“An outlier is an observation which deviates so much from the other observations as to arouse suspicions that it was generated by a different mechanism” (Hawkins, 1980 conforme Aggarwal, 2013).

Por último, existe ainda a possibilidade de um *outlier* ser uma observação normal, “surprising veridical data” (John, 1995) pelo que, antes de decidir-se o que deverá ser feito com as observações *outliers* é conveniente ter conhecimento das causas que levam ao seu aparecimento.

Em muitos casos, as razões da existência de *outliers* determinam a forma como devem ser tratadas estas observações. Segundo, Kriegel et al. (Kriegel, Kröger, & Zimek, 2009) a deteção de *outliers* é usada para: verificar erros de medição / execução / introdução valores, avaliar a variabilidade inerente dos elementos da população, detetar fraudes, conhecer o comportamento de gastos de consumidores, estudos médicos, pesquisa farmacêutica e marketing.

2.2. Métodos de identificação de *outliers*

Nesta secção, vai ser feita uma breve descrição de alguns dos vários métodos (selecionados como de maior interesse) para a detecção de *outliers*.

De acordo com a literatura existente (Oliveira, 2008) têm sido propostos um grande número de testes de detecção de *outliers.* A título exemplificativo, podem-se referir, os testes que têm como base o critério de “distância da média”, ou ainda, o teste de Dixon que se baseia num valor ser demasiado grande (ou

pequeno) em relação ao seu vizinho mais próximo. É ainda importante aqui salientar que, para a deteção de *outliers* deve ser usada a mediana (med) e não a média ($\bar{x}$) dos valores (Leys, Ley, Klein, Bernard, & Licata, 2013).

2.2.1. Método Box-plot

O método para deteção de *outliers* baseado na regra *box-plot* foi introduzido por Tukey (1977). Posteriormente, esta regra foi estudada por Hoaglin, et al. (Hoaglin, Iglewicz, & Tukey, 1986), e foi convertida, numa regra adequada de identificação de um *outlier* por Hoaglin e Iglewicz (Hoaglin & Iglewicz, 1987).

O gráfico *box-plot* tornou-se desde então um dos mais populares procedimentos estatísticos gráficos. Tukey incluiu ainda, uma regra simples para identificar observações como valores atípicos. Essa regra identifica observações discrepantes (*outliers*), quando estas estão fora do intervalo:

$$[\,(Q1 - g \times (Q3 - Q1), Q3 + g \times (Q3 - Q1)\,]$$

Equação 1 - Regra de Tukey para identificar *outliers*

sendo:

$Q1$ – 1º Quartil,

$Q3$ – 3º Quartil,

g – valor para diferenciar entre *outliers* “moderados” e “severos”.

As escolhas mais comuns para g são de 1,5 para sinalizar valores "moderados" e 3,0 para a sinalização de valores " severos ".

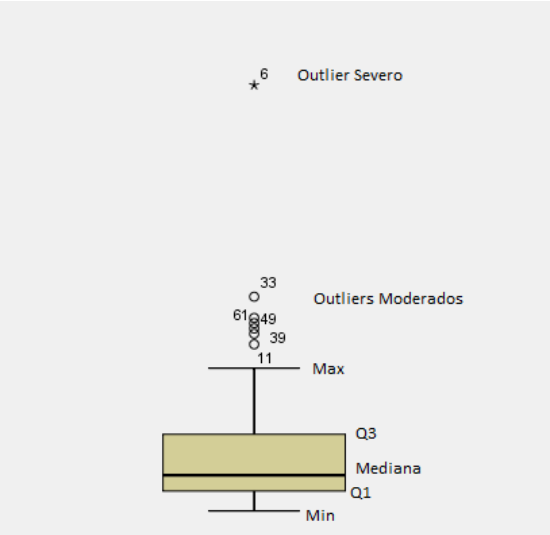


Figura 1 – Exemplo de *Box-plot* com identificação dos *outliers*

Os valores que estiverem fora do intervalo de $Q3 + 1{,}5 \times (Q3 - Q1)$ e $Q1 - 1{,}5 \times (Q3 - Q1)$ serão considerados *outliers* moderados (o), e os valores fora do intervalo $Q3 + 3 \times (Q3 - Q1)$ e $Q1 - 3 \times (Q3 - Q1)$ serão considerados *outliers* severos (*) (cf. Figura 1).

De referir ainda que, tanto para os *outliers* moderados como para os severos detetados é necessário investigar a sua origem (isto é, o porquê da sua existência).

2.2.2. Box-plot modificado

Vanderviere e Huber (Vanderviere & Huber, 2004) introduziram um *box-plot* modificado tendo em conta a medida *medcouple* (MC), que é uma medida robusta de assimetria de uma distribuição assimétrica.

Sendo $\boldsymbol{x} = \{x_1,\ x_2,\ \ldots,\ x_n\}$ um conjunto de dados obtidos de forma independente e ordenados crescentemente, isto é, $x_1 \leq x_2 \leq \cdots \leq x_n$, a *medcouple* (MC) dos dados é definida por:

$$MC_{(\ x_1,\ldots,x_n\ )} = med\ \frac{(x_j - med_k) - (med_k - x_i)}{x_j - x_i}$$

Equação 2 – Definição de *medcouple* (MC)

onde, med_k é a mediana de $\boldsymbol{x}$ e, i e j têm que satisfazer a condição $x_i \leq med_k \leq x_j$ e $x_j \neq x_i$.

O intervalo *box-plot* modificado (Brys, Hubert, & Rousseeuw, 2005) é então definido pelos seguintes limites:

$$[L, U] =$$

$$[Q_1 - 1.5 \times \exp(-3.5\, MC) \times IQR\,, Q_3 + 1.5 \times \exp(4\, MC) \times IQR] \quad se\ MC \geq 0$$

$$[Q_1 - 1.5 \times \exp(-4\, MC) \times IQR\,,\ Q_3 + 1.5 \times \exp(3.5\, MC) \times IQR] \quad se\ MC \leq 0$$

onde, L é o limite inferior e U é o limite superior do intervalo. As observações que se enquadram fora do intervalo são consideradas *outliers*.

2.2.3. Teste de Dixon

Segundo Dean & Dixon (1951), o teste de Dixon é geralmente utilizado para a deteção de um pequeno número de valores extremos (*outliers*). Este teste, só pode ser usado convenientemente quando o tamanho da amostra está entre 3 e 25 observações. Este teste possui ainda algumas limitações, tais como, poder originar a ocultação de múltiplos *outliers*.

Mais tarde, no sentido de evitar a ocultação de dois *outliers*, do mesmo lado da distribuição, propuseram a seguinte decisão:

- estando os dados classificados por ordem crescente ($x_1 \leq x_2 \leq \cdots \leq x_n$) se x_1 é um *outlier*, deve ser detetado por comparação do valor calculado r com os valores críticos listados na tabela 1 (em que n é o número de observações e α é o nível de significância[2]), ou seja, se o valor calculado r é maior do que o valor crítico (cf. Tabela 1), a observação x_1 é considerada um *outlier*.

[2] De acordo com Maroco (2007, p.71) esta "evidência" é o já definido α (nível de significância). Sir Ronald Fisher (o autor de várias metodologias estatística, entre elas a ANOVA) sugeriu que se um determinado resultado ocorresse mais do que 1 vez em 20 tentativas ao acaso, então dever-se-ia considerar o resultado em causa como sendo o real e não uma mera coincidência. Como 1/20=0.05 (5%), este valor é geralmente usado como a probabilidade (5%) ou nível de significância (0.05) para decidir se algo é realmente representativo da população teórica ou não. Outros níveis de significância usados com frequência são 0.1 e 0.01.

Este teste dependendo do tamanho da amostra (n), utiliza as seguintes fórmulas para o cálculo de r:

1. para $3 \leq n \leq 7$

$$r = \frac{x_2 - x_1}{x_n - x_1}$$

Equação 3 – Valor calculado (r) de Dixon

2. para $8 \leq n \leq 10$

$$r = \frac{x_2 - x_1}{x_{n-1} - x_1}$$

Equação 4 – Valor calculado (r) de Dixon

3. para $11 \leq n \leq 13$

$$r = \frac{x_3 - x_1}{x_{n-1} - x_1}$$

Equação 5 – Valor calculado (r) de Dixon

4. para $n >= 14$

$$r = \frac{x_3 - x_1}{x_{n-2} - x_1}$$

Equação 6 – Valor calculado (r) de Dixon

Tabela 1 - Valores críticos de Dixon. Fonte: Kanji, 1993

	n	Nível de significância α						
		0.30	0.20	0.10	0.05	0.02	0.01	0.005
$r = \frac{x_2 - x_1}{x_n - x_1}$	3	0.684	0.781	0.886	0.941	0.976	0.988	0.994
	4	0.471	0.560	0.679	0.765	0.846	0.889	0.926
	5	0.373	0.451	0.557	0.642	0.729	0.780	0.821
	6	0.318	0.386	0.482	0.560	0.644	0.698	0.740
	7	0.281	0.344	0.434	0.507	0.586	0.637	0.680
$r = \frac{x_2 - x_1}{x_{n-1} - x_1}$	8	0.318	0.385	0.479	0.554	0.631	0.683	0.725
	9	0.288	0.352	0.441	0.512	0.587	0.635	0.677
	10	0.265	0.325	0.409	0.477	0.551	0.597	0.639
$r = \frac{x_3 - x_1}{x_{n-1} - x_1}$	11	0.391	0.442	0.517	0.576	0.638	0.679	0.713
	12	0.370	0.419	0.490	0.546	0.605	0.642	0.675
	13	0.351	0.399	0.467	0.521	0.578	0.615	0.649
$r = \frac{x_3 - x_1}{x_{n-2} - x_1}$	14	0.370	0.421	0.492	0.546	0.602	0.641	0.674
	15	0.353	0.402	0.472	0.525	0.579	0.616	0.647
	16	0.338	0.386	0.454	0.507	0.559	0.595	0.624
	17	0.325	0.373	0.438	0.490	0.542	0.577	0.605
	18	0.314	0.361	0.424	0.475	0.527	0.561	0.589
	19	0.304	0.350	0.412	0.462	0.514	0.547	0.575
	20	0.295	0.340	0.401	0.450	0.502	0.535	0.562
	21	0.287	0.331	0.391	0.440	0.491	0.524	0.551
	22	0.280	0.323	0.382	0.430	0.481	0.514	0.541
	23	0.274	0.316	0.374	0.421	0.472	0.505	0.532
	24	0.268	0.310	0.367	0.413	0.464	0.497	0.524
	25	0.262	0.304	0.360	0.406	0.457	0.489	0.516

2.2.4. Teste de Grubbs

De acordo com Grubbs (Grubbs, 1969), o método é utilizado para detetar *outliers* num conjunto de dados univariado, provenientes de uma população que segue aproximadamente uma distribuição Normal (o que implica verificar a normalidade antes de aplicar o teste). De referir ainda que, o teste de Grubbs é um teste que deteta um *outlier* de cada vez. Assim, após este ser detetado, é removido da amostra e o teste é aplicado novamente até que não sejam detetados mais *outliers*, ou seja, é um processo iterativo.

O teste de Grubbs encontra, o valor que tem o maior desvio médio absoluto. Note-se que se um *outlier* for identificado e removido, o teste não deve ser repetido sem antes encontrar o novo valor crítico (cf. Tabela 2) para a nova dimensão da amostra.

A aplicação do teste é então muito simples: procura o máximo das diferenças absolutas entre o x_i, $i = 1,2,\dots,n$ (valor observado na amostra) e média da amostra $\bar{x}$. O resultado é dividido pelo desvio padrão da amostra (cf. Equação 7). Se o valor calculado g for maior do que o valor crítico (cf. Tabela 2), o valor x_i, pode ser considerado um *outlier*.

$$g = \frac{\max_{i=1..n} \left| x_i - \bar{x} \right|}{s}$$

Equação 7 - Valor calculado (g) de Grubbs

Tabela 2 – Valores críticos de Grubbs (α=0,05 e α=0,01)

n	g_{crit} α=0.05	g_{crit} α=0.01
3	1.1543	1.1547
4	1.4812	1.4962
5	1.7150	1.7637
6	1.8871	1.9728
7	2.0200	2.1391
8	2.1266	2.2744
9	2.2150	2.3868
10	2.2900	2.4821
11	2.3547	2.5641
12	2.4116	2.6357
13	2.4620	2.6990
14	2.5073	2.7554
15	2.5483	2.8061
16	2.5857	2.8521
17	2.6200	2.8940
18	2.6516	2.9325
19	2.6809	2.9680
20	2.7082	3.0008

n	g_{crit} α=0.05	g_{crit} α=0.01
25	2.8217	3.1353
30	2.9085	3.2361
40	3.0361	3.3807
50	3.1282	3.4825
60	3.1997	3.5599
70	3.2576	3.6217
80	3.3061	3.6729
90	3.3477	3.7163
100	3.3841	3.7540
120	3.4451	3.8167
140	3.4951	3.8673
160	3.5373	3.9097
180	3.5736	3.9460
200	3.6055	3.9777
300	3.7236	4.0935
400	3.8032	4.1707
500	3.8631	4.2283
600	3.9109	4.2740

2.2.5. Teste Z-score

Um teste Z-score é um teste estatístico para o qual a distribuição da estatística teste, pode ser aproximada por uma distribuição normal. Muitos testes estatísticos podem ser convenientemente realizados como Z-scores se o tamanho da amostra é grande (n≥30) ou a variância da população conhecida. Se a variância da população é desconhecida e o tamanho da amostra não é grande (n <30), o teste t de Student pode ser mais apropriado.

Assim, para realizar-se um teste Z-score deve-se começar por calcular os Z-scores observados para a amostra, isto é, os valores z-estandardizados para cada uma das observações (x_i, $i = 1,2,\dots,n$) da amostra. O valor Z-score para uma observação é definido como

$$zi = \frac{xi - \bar{x}}{s}$$

Equação 8 – Z-score observado

onde $\bar{x}$ e s são a média e desvio padrão da amostra, respetivamente.

De seguida, pode-se avaliar os valores obtidos para os Z-scores observados e classifica-los:

- Se a dimensão da amostra n<50, valores observados para os Z-scores inferiores a -2.5 ou superiores a 2.5 devem ser considerados *outliers*.

- Se a dimensão da amostra está compreendida entre 50 e 100, valores observados para os Z-scores inferiores a -3.3 ou superiores a 3.3 são tipicamente considerados *outliers*.

- Se dimensão da amostra é muito grande (1000 ou mais observações), valores mais extremos do que ±3.3 podem ser considerados dados normais e não *outliers*.

Por último, pode-se concluir que o teste Z-score é um critério muito usado na prática para a identificação de *outliers* quando os dados são provenientes de uma população que segue uma distribuição normal (Shiffler, 1988), o valor Z-score máximo obtido para uma dada amostra depende do tamanho da mesma. Outra limitação desta regra é que o desvio padrão pode ser inflacionado por valores extremos.

2.2.6. Teste Z-score modificado

Os valores observados ($\bar{x}$ e s) para os estimadores ($\bar{X}$ e S) utilizados no teste Z-score, podem ser afetados por valores extremos. Para evitar este

problema, substitui-se estas medidas pela mediana (med) e mediana do desvio absoluto (MAD) no teste Z-Score modificado (Iglewicz & Hoaglin, 1993):

$$mi = \frac{0{,}6745\ (xi - med\ )}{MAD}$$

Equação 9 – Z-score modificado observado

onde MAD é a mediana do desvio absoluto (isto é, *MAD* é a mediana dos desvios absolutos em torno da mediana {| $(xi - med)$ |}.

O valor 0,6745 é usado para assegurar, o mesmo nível de significância (α) como no caso da distribuição Normal.

Iglewicz e Hoaglin, recomendam que os valores observados Z-scores modificados com um valor absoluto superior a 3,5 sejam considerados como potenciais *outliers*.

3. Data Mining

3.1. Conceito de Data Mining

O d*ata mining,* designação em inglês amplamente utilizada para a prospeção ou mineração de dados, consiste na procura de relacionamentos, padrões ou modelos que estão implícitos nos dados armazenados em grandes bases de dados (Santos & Ramos, 2009).

O *data mining* pode, assim, ser encarado como o processo de explorar grandes quantidades de dados com vista à identificação de padrões consistentes, como regras de associação ou sequências temporais, para detetar relacionamentos sistemáticos entre variáveis. Utiliza algoritmos para descobrir regras, identificar fatores e tendências-chave, descobrir padrões e relacionamentos ocultos em grandes bases de dados. Informação esta, que depois de interpretada é utilizada no suporte à tomada de decisão organizacional.

.A evolução das tecnologias de informação permitiu a recolha e armazenamento de grandes volumes de dados. As atividades envolvidas na análise destas grandes bases de dados, normalmente com o objetivo de extrair conhecimento útil para suportar o processo de decisão, são conhecidas como *data mining, knowledge discovery, pattern recognition* e *machine learning* (Vercellis C, 2009).

A expressão *data mining* surgiu pela primeira vez em 1990 (Han, Kamber, & Pei, 2011) em comunidades de bases de dados. A prospeção de dados é a etapa de análise do processo conhecido como *Knowledge Discovery in Databases (KDD)* ou Descoberta de Conhecimento em Bases de Dados.

O *data mining* é uma prática relativamente recente no mundo da computação, e utiliza técnicas de recuperação de informação, inteligência artificial, reconhecimento de padrões e de estatística para procurar correlações entre diferentes dados que permitam adquirir um conhecimento benéfico para uma empresa ou indivíduo.

O propósito do *data mining* é analisar e tirar algumas conclusões a partir de uma amostra de dados históricos em relação à população total, da maneira mais exata possível (Vercellis C, 2009).

3.2. O Data Mining como uma ferramenta de Business Intelligence

O *Data Mining* é frequentemente utilizado como umas das ferramentas de análise de dados de projetos de *Business Intelligence*. Quando utilizado nesse contexto, os dados utilizados na mineração estão armazenados num *data warehouse*.

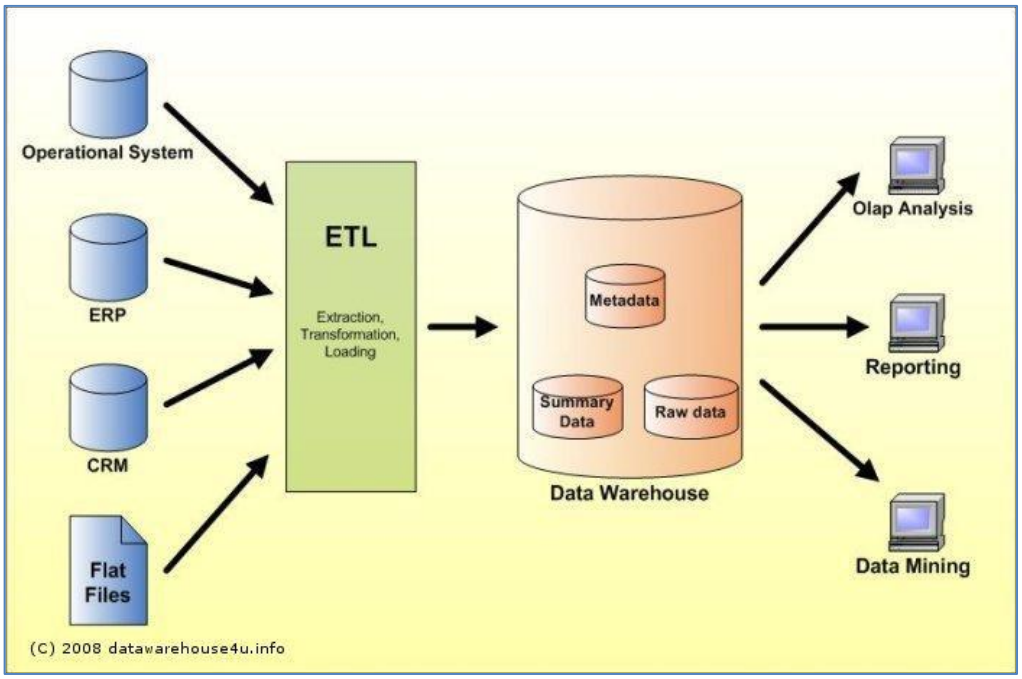


Figura 2 – Enquadramento do *data mining* num projeto de Business Intelligence. Fonte: STAT4U, 2008

Com os dados armazenados no *datawarehouse*, é possível utilizar as técnicas OLAP (*On Line Analytical Processing*) para efetuar análise multidimensionais dos dados, o *data mining* para identificar correlações e padrões nos dados ou factos desconhecidos, e as capacidades de emissão de relatórios geralmente existentes nas ferramentas de *software* de *Business Intelligence* (cf. Figura 2).

3.3. Fases do Processo de Data Mining

O processo de descoberta de conhecimento em bases de dados inclui um conjunto de fases ilustradas na Figura 3 e descritas nas secções seguintes.

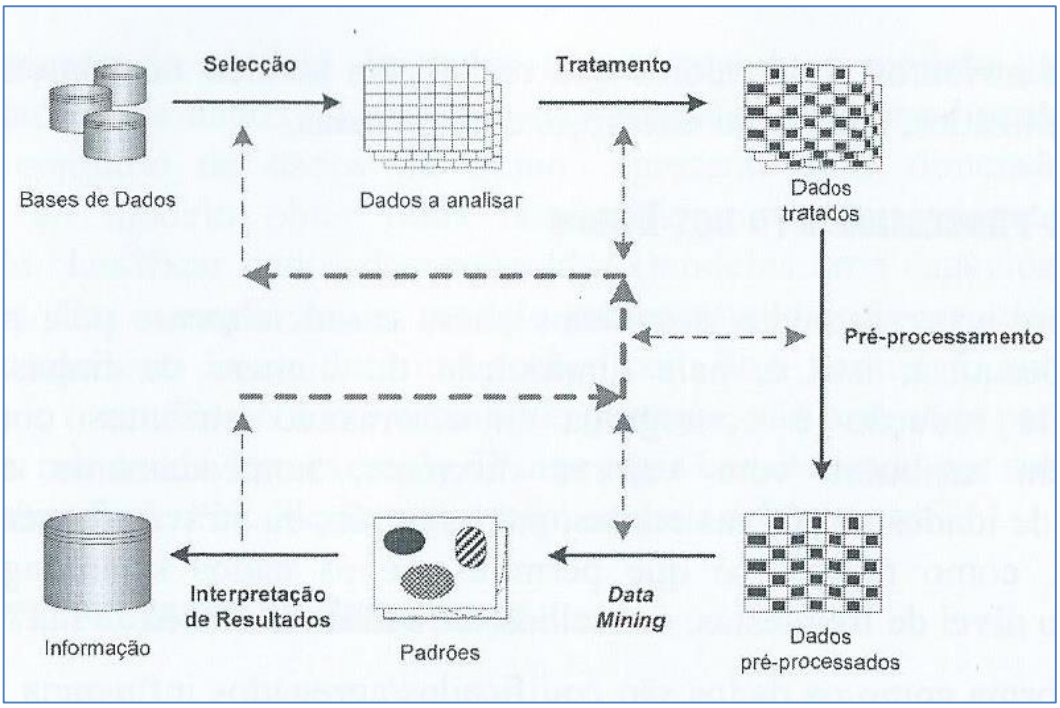


Figura 3 – Processo de descoberta de conhecimento em bases de dados. Fonte: Santos & Ramos, 2009

3.3.1. Seleção dos dados

Nesta fase são selecionados os dados armazenados nos diversos repositórios de dados (sistemas operacionais, *Data Warehouses,* etc*.)*, necessários aos algoritmos de *Data Mining*. A seleção de dados tem como principal objetivo limitar o espaço de pesquisa, eliminar atributos que não têm qualquer interesse no processo de descoberta de conhecimento (Santos & Ramos, 2009). Nessa categoria de atributos sem interesse para a mineração e que podem ser eliminados, incluem-se atributos de carácter meramente informativo, como, nomes de clientes, números de contribuinte, etc..

3.3.2. Tratamento dos dados

No tratamento dos dados, procede-se à limpeza e verificação dos dados. São detetadas situações como registos duplicados (usualmente originadas por erros na introdução dos dados) ou ainda outras inconsistências como atributos com valores fora da gama de valores aceitável. A verificação de inconsistências nos dados, permite a identificação de registos que possuem atributos com valores que não fazem sentido no contexto em que estão a ser utilizados, como, por exemplo, datas erradas (Santos & Ramos, 2009).

3.3.3. Pré-processamento dos dados

Esta fase visa essencialmente a redução do espaço de pesquisa, isto é, a diminuição do número registos e colunas a analisar. Esta redução é conseguida transformando atributos com valores contínuos em atributos com valores discretos ou através da generalização de atributos. Assim, por exemplo, idades podem ser substituídas por faixas etárias e atributos de localização podem ser generalizados como regiões, o que poderá permitir a análise dos dados ao nível de freguesias, concelhos e distritos. A forma como os dados são codificados/agregados influencia fortemente os resultados que poderão ser obtidos na fase de *Data Mining*. Assim nesta fase, os dados provenientes da fase anterior são agrupados ou transformados, de forma a tornar a pesquisa mais eficiente (Santos & Ramos, 2009).

3.3.4. Data mining

Esta é a fase da mineração propriamente dita, na qual os dados são analisados. Nesta fase, o tipo de problema a resolver e tipo de resultados pretendido, permite a identificação da categoria e técnica de *data mining* a utilizar. Para atingir os objetivos propostos pode ser necessário utilizar mais do que uma técnica, já que a qualidade e tipo dos dados disponíveis influencia de forma decisiva os resultados que podem ser encontrados (Santos & Ramos, 2009).

3.3.5. Interpretação de resultados

A interpretação de resultados consiste na análise dos resultados obtidos pela implementação dos algoritmos utilizados na fase anterior. A ocorrência de falhas ao longo do processo de descoberta de conhecimento, originadas por decisões que se revelam, nesta fase, inapropriadas, é normalmente traduzida na obtenção de modelos que não satisfazem o interesse do utilizador. Nestes casos, existe a possibilidade de retrocesso a fases anteriores para alterar as decisões tomadas ou para incluir novos dados na análise. O processo é posteriormente retomado, permitindo identificar novos modelos que resultam das alterações efetuadas (Santos & Ramos, 2009).

3.4. Categorias de Data Mining

As atividades associadas ao *Data Mining* podem ser divididas em dois grupos: descrição ou previsão. No *Data Mining* descritivo, identificam-se regras que caracterizam os dados analisados. Por outro lado, no *Data Mining* preditivo utilizam-se determinados atributos da base de dados ou conjunto de dados para prever o valor desconhecido ou futuro de uma outra variável de interesse.

Esta distinção está associada ao objetivo da atividade de *Data Mining*, que pode permitir aumentar o conhecimento acerca dos dados, descrição, ou suportar o processo de tomada de decisão, previsão, através de modelos capazes de prever o valor de uma variável (Berry & Linoff, 2000).

Um exemplo de *data mining* na categoria descritiva é a determinação dos perfis de compras dos clientes de uma organização para criação de campanhas de marketing direcionadas, com base na análise da base de dados das transações dos clientes. Um exemplo de *data mining* na categoria de previsão é a previsão de vendas de um produto em termos futuros, para uma melhor gestão de *stocks*, com base na análise do histórico de vendas desse produto.

Relativamente à previsão, o melhor modelo é aquele que apresenta a precisão mais elevada, permitindo uma percentagem de acerto superior à

percentagem de acerto conseguida por outros modelos, ainda que estes tenham sido mais fáceis de obter e de perceber. Por outro lado, o melhor modelo em descrição pode não ser aquele que obtém resultados mais precisos em termos de confiança do modelo, mas sim o que permite um conhecimento mais alargado dos dados analisados.

Cada uma das categorias de *data mining* identificadas (descrição e previsão) inclui um conjunto de métodos que deverão ser utilizados tendo em conta a natureza do problema a resolver. A organização dos métodos de *data mining* por categorias é apresentada na Tabela 3.

Tabela 3 – Categorias, métodos e casos de utilização de *Data Mining*

	Categoria	Técnica	Variável alvo	Tipo de Aprendizagem	Métodos	Casos de utilização
Data mining	**Preditivo**	**Regressão** *(Regression)*	Contínua	Supervisionada	• Regressão linear simples • Regressão linear múltipla	Previsão do valor de uma habitação Previsão de vendas com base numa campanha promocional
		Séries *temporais* *(Time series)*	Contínua	Supervisionada	• Suavização exponencial • Modelo auto-regressivo	Previsão de vendas mensais de produtos Previsão de cotações de ações
		Classificação *(Classification*)	Discreta	Supervisionada	• Árvores de classificação • Métodos bayesianos • Regressão logística • Redes neuronais • Máquina de vetores de suporte	Avaliação de clientes para atribuição de crédito Deteção de fraudes de clientes Previsão do modelo de resposta a campanhas de marketing (resposta/não resposta) Previsão de fuga (abandono) de clientes
	Descritivo	**Regras de associação** *(Association rules)*	Não existe	Não supervisionada	• Regras de associação simples • Algoritmo apriori	Análise do cesto de compras Recomendação de produtos Identificação de oportunidades de venda cruzada (*cross-selling*)
		Segmentação *(Clustering)*	Não existe	Não supervisionada	• Métodos de particionamento • Métodos hierárquicos	Segmentação de clientes Determinação do perfil de clientes de alto valor

3.5. Técnicas e Métodos de Data Mining

Nesta secção descrevem-se de uma forma mais detalhada os principais métodos de *data mining* apresentados na Tabela 3.

3.5.1. Regressão (*Regression*)

São modelos de aprendizagem supervisionada (*supervised learning models*) que tratam conjuntos de dados de observações históricas, para as quais o valor dos atributos explanatórios assim como o valor numérico contínuo da variável alvo é conhecido (Vercellis C, 2009).

O objetivo dos modelos de regressão, também conhecidos como *explanatory models,* é identificar a relação entre a variável alvo e os atributos contidos no conjunto de dados. Estes modelos são também usados para prever o valor futuro da variável alvo, com base na relação desse valor com os atributos.

Os modelos de regressão são, por exemplo, usados para interpretar as vendas de um produto relacionando-as com os investimentos feitos em diferentes meios de comunicação social. O modelo pode ainda ser usado para prever o efeito das diferentes políticas de marketing nas vendas do produto.

Regressão linear simples (*Simple line regression*)

A regressão é utilizada sempre que se pretende prever uma variável com valores contínuos. Na regressão linear, os dados são modelados aproximando-os a uma linha reta. Constitui a forma mais simples de regressão, e é representada através de uma equação (cf. Equação 10) com duas variáveis, X e Y, tal que:

$$Y = \alpha + \beta X$$

Equação 10 – Regressão linear simples

em que X representa a variável independente, Y a variável dependente calculada a partir de X, sendo α e β os coeficientes de regressão.

Estes coeficientes de regressão podem ser determinados a partir do método dos mínimos quadrados, que procura minimizar o erro existente entre os valores reais dos dados e o valor estimado para a reta.

Utilizando este modelo, é por exemplo, possível verificar a relação entre o investimento feito em publicidade num jornal e as vendas do produto (X: investimento em publicidade no jornal; Y: vendas do produto).

Regressão linear múltipla (*Multiple line regression*)

É a tentativa de regressão linear para modelar a relação entre duas ou mais variáveis explicativas e uma variável resposta ajustando uma equação linear aos dados observados. Cada valor da variável independente X está associado com um valor da variável dependente Y, conforme apresentado na Equação 11:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_n X_n + E$$

Equação 11 – Regressão linear múltipla

X – Variável explicativa ou independente medida sem erro, não aleatória

E – Variável aleatória residual na qual se procuram incluir todas as influências no comportamento da variável Y que não podem ser explicadas linearmente pelo comportamento da variável X

β_0, β_1 – Parâmetros desconhecidos do modelo (a estimar)

Y – Variável explicada ou dependente (aleatória)

Utilizando este modelo, é por exemplo, possível verificar a relação entre o investimento feito em publicidade num jornal (X_1), a descida de preço do produto (X_2) e as vendas do produto (Y).

3.5.2. Séries temporais (*Time series*)

Existem conjuntos de dados em que o atributo alvo é dependente do tempo, ou seja, está associado a uma sequência consecutiva de períodos, e o interesse é conhecer essa dependência. O objetivo dos modelos de séries temporais é identificar padrões regulares de observações históricas com o objetivo de fazer previsões para o futuro. Estes modelos são usados para a previsão de vendas de produtos, determinação de tendências económicas, etc. (Vercellis C, 2009).

Suavização exponencial (*Exponential smoothing*)

A suavização exponencial é um método muito utilizado na produção de uma série temporal. O método considera pesos exponenciais, que diminuem conforme a antiguidade das observações. Ou seja, as observações recentes têm mais peso que as antigas para a previsão.

Na suavização exponencial, há um ou mais parâmetros de suavização a serem determinados (ou estimados) e essas escolhas determinam os pesos atribuídos às observações.

O rótulo Holt-Winters (HW) é frequentemente atribuído a um conjunto de procedimentos que formam o núcleo da família de métodos de previsão de suavização exponencial. As estruturas básicas foram fornecidas por C.C. Holt em 1957 e pelo seu aluno P. Winters em 1960 (Winters, 1960).

A suavização exponencial simples (*Simple exponential smoothing*) não funciona bem quando há uma tendência nos dados. As tendências de dados permitem avaliar como os dados de resposta mudam ao longo do tempo; por exemplo, pode-se criar um inquérito de satisfação de clientes durante um ano, e verificar se o número de clientes satisfeitos aumentou ou diminuiu ao longo desse ano.

Para essas situações, vários métodos foram criados sob o nome de "Suavização exponencial de segunda ordem" (*Double exponential smoothing (Holt))*, em que existe a aplicação recursiva de um filtro exponencial duas vezes, justificando, por isso, a designação "suavização exponencial dupla" (cf. Equação

12). A ideia de base inerente à suavização exponencial dupla consiste na introdução de um prazo para ter em conta a possibilidade de uma série que apresenta algum tipo de tendência. Este componente de inclinação é em si atualizado via suavização exponencial.

$$s_t = \alpha y_t + (1-\alpha)(s_{t-1} + m_{t-1})$$

$$m_t = \beta(s_t - s_{t-1}) + (1-\beta)\, m_{t-1}$$

$$f_{t+1} = s_t + m_t$$

Equação 12 – Suavização Exponencial Dupla (aditivo)

Existe ainda a "Suavização exponencial tripla" (Triple *exponential smoothing (Holt-Winters)*, que considera ainda mais uma componente, a sazonalidade (Makridakis, Wheelwright, C., & Hyndman, 1998), conforme apresentado na Equação 13.

$$s_t = \alpha \frac{y_t}{q_{t-L}} + (1-\alpha)(s_{t-1} + m_{t-1})$$

$$m_t = \beta(s_t - s_{t-1}) + (1-\beta)\, m_{t-1}$$

$$q_t = \gamma \frac{y_t}{s_t} + (1-\gamma)\, q_{t-L}$$

$$f_{t+1} = (s_t + m_t)\, q_{t-L+1}$$

Equação 13 – Suavização Exponencial Tripla (aditivo)

em que:

t = Período de tempo corrente (current time period)

y_t = Valor atual observado no momento t

α = Constante do processo de suavização

β = Constante de tendência de suavização

s_t = Valor de suavização no período t

m_t = Valor de tendência no período t

$f_{t+1} = s_t + m_t$ = Valor da previsão para t+1

γ = Constante de sazonalidade de suavização

L = Número de ciclos sazonais

A robustez e precisão das previsões efetuadas por suavização exponencial levou à sua ampla utilização em aplicações em que um grande número de séries necessita de um procedimento automatizado, tal como no caso do controle de *stocks*.

Embora o método de Holt tenda a ser a abordagem mais popular para a tendência da série, a sua função de previsão linear tem sido criticada pela tendência a ultrapassar os dados para além do curto prazo.

Gardner e McKenzie (1985) resolveram este problema através da inclusão de um parâmetro extra no método de Holt para amortecer a tendência projetada. Apesar da sua popularidade, a evidência empírica tem demonstrado que a função linear de previsão Holt tende a superestimar (Gardner, E.S. & McKenzie, 1985). Consequentemente, Gardner e McKenzie propõem a utilização de um parâmetro de amortecimento Ø no método de Holt para controlar melhor a extrapolação de tendências. Assim, o método de Holt amortecido sugerido por Gardner e McKenzie (1985) é descrito pela Equação 14:

$$C_t = \alpha X_t + (1-\alpha)(C_{t-1} + \emptyset\, T_{t-1})$$

$$T_t = \beta(C_t - C_{t-1}) + (1-\beta)\,\emptyset\, T_{t-1}$$

$$X_t(m) = C_t + \sum_{i=1}^{m} \emptyset^i\ T_t$$

Equação 14 – Método de Holt amortecido

em que:

t = Período de tempo corrente

y_t = Valor atual observado no momento t

α = Constante do processo de suavização ($0 < \alpha < 1$)

β = Constante de tendência de suavização (0 < β < 1))

C_t = Valor de suavização no período t

T_t = Valor de tendência no período t

Ø = Constante de amortecimento (0 < Ø < 1)

$X_t(m)$ = Valor da previsão para t+m

Pegels (1969) sugere que o seu método multiplicativo da tendência possa ser mais útil do que o método de Holt que considera uma tendência aditiva, uma vez que a tendência multiplicativa tende a ser mais provável em aplicações da vida real.

Segundo Taylor (2003), pode haver vantagem em incluir um parâmetro extra na formulação de Pegels para amortecer a tendência extrapolada, de um modo análogo ao parâmetro de amortecimento utilizado no método Holt. Assim, no método de Pegels com tendência multiplicativa, Taylor (2003) sugere a inclusão de um parâmetro de amortecimento (cf. Equação 15).

$$C_t = \alpha X_t + (1-\alpha)\,(C_{t-1}\,T_{t-1}^{\emptyset})$$

$$T_t = \beta(C_t\,/C_{t-1}) + (1-\beta)\,T_{t-1}^{\emptyset}$$

$$X_t(m) = C_t + T_t^{\sum_{i=1}^{m}\emptyset^i}$$

Equação 15 – Método de Pegels amortecido

em que:

t = Período de tempo corrente

y_t = Valor atual observado no momento t

α = Constante do processo de suavização (0 < α < 1)

β = Constante de tendência de suavização (0 < β < 1)

C_t = Valor de suavização no período t

T_t = Valor de tendência no período t

Ø = Constante de amortecimento (0 < Ø < 1)

$X_t(m)$ = Valor da previsão para o período t+m

Para aferir a precisão do modelo a medida de precisão SMAPE (Symmetric Mean Absolute Percentage Error) pode ser utilizada, bem como para determinar os melhores valores dos parâmetros $(\alpha, \beta, \emptyset)$ a considerar (cf. Equação 16).

A utilização do SMAPE é vantajosa em relação à utilização do MAPE (Mean Absolute Percentage Error) mais tradicional, uma vez que evita grandes erros quando o x_i real está perto de zero, e grandes diferenças entre o erro percentual absoluto quando x_i é maior do que a previsão f_i, e quando f_i é maior do que x_i.

$$SMAPE = \sum_i \frac{|x_i - f_i|}{(x_i + f_i)/2}$$

Equação 16 – SMAPE (Symmetric Mean Absolute Percentage Error)

Modelo auto-regressivo (*Autoregressive model*)

Os métodos auto-regressivos são baseados na ideia de que é possível identificar a relação entre as observações e a série temporal, estudando a auto correlação entre observações separadas por um intervalo fixo de tempo.

Na análise estatística de séries temporais, os modelos *autoregressive–moving-average* (ARMA) fornecem uma pequena descrição de um processo estocástico estacionário em termos de dois polinómios, um para a auto regressão e outro para a média móvel. O modelo geral ARMA foi descrito em 1951 na tese de Peter Whittle e foi popularizado em 1976 no livro dos autores George Box EP e Gwilym Jenkins (Box & Jenkins, 1976).

Dada uma série temporal de dados X_t, o modelo ARMA é uma ferramenta para a compreensão e previsão de valores futuros da série. O modelo consiste em duas partes, uma parte auto-regressiva (AR) e uma parte média móvel (MA).

Assim sendo, o modelo é geralmente designado por modelo ARMA (p, q), onde p é a ordem da parte auto-regressiva e q é a ordem da parte média móvel.

Os modelos AR (p) podem ser definidos de uma forma geral pela Equação 17:

$$X_t = \emptyset_1 X_{t-1} + \emptyset_2 X_{t-2} + \cdots + \emptyset_p X_{t-p} + a_t$$

Equação 17 – Modelo AR (p)

Onde a_t é um termo de ruído branco (é uma variável aleatória independente e uniformemente distribuída) e p corresponde ao número de termos auto-regressivos. Nos modelos AR (1), o valor de X no período t depende do seu valor no período t-1 e um termo de erro aleatório, sendo os valores de X expressos como desvios da sua média, ou seja, o valor previsto para X_t é simplesmente uma proporção ($\emptyset_1$) do valor de X_{t-1} .

Outro modelo pertencente à família ARMA é o MA (q) que pode ser expresso de acordo com a Equação 18:

$$X_t = \mu + \beta_0\alpha_t + \beta_1\alpha_{t-1} + \beta_2\alpha_{t-2} + \cdots + +\beta_q\alpha_{t-q}$$

Equação 18 – Modelo MA (q)

em que μ é uma constante e q é o número de médias móveis existentes.

Nos modelos MA(q) o valor de X no período t é uma constante (μ) mais uma média móvel dos termos de erro presentes (β_0) e passados (β_q). O processo de MA (q) é simplesmente uma combinação linear de termos de erro de um ruído branco (a_t).

Um modelo ainda pode apresentar características de um processo AR (p) e de um processo MA (q). Os Modelos ARMA (p, q) podem ser representados pela Equação 19:

$$X_t = \emptyset + \emptyset_1 X_{t-1} + \beta_0\alpha_t + \beta_1\alpha_{t-1} + \emptyset_2 X_{t-2} + \beta_2\alpha_{t-2} + \cdots + \emptyset_p X_{t-p} + \beta_q\alpha_{t-q}$$

Equação 19 – Modelo ARMA (p,q)

onde $\emptyset$ é um termo constante. É interessante notar que os modelos AR (p), MA (q) e ARMA (p, q) pressupõem que a série analisada seja estacionária (média e variância constantes ao longo do tempo).

O ARMA é apropriado quando um sistema é uma função de uma série de eventos não observados (a parte MA), bem como o seu próprio comportamento.

Por exemplo, os valores das cotações das ações podem ser alterados com base em informações fundamentais sobre o mercado, bem como exibindo tendências e efeitos dos participantes do mercado. Este método pode ser utilizado para fazer a previsão das cotações de ações na Bolsa de Valores, com base no histórico de cotações.

3.5.3. Classificação (*Classification*)

A classificação permite o enquadramento de um conjunto de dados em classes predefinidas, identificando a classe a que cada elemento pertence, no atributo designado por atributo de saída. A classificação é utilizada para prever valores discretos ou nominais.

A classificação de dados é um processo que engloba duas etapas (Han et al., 2011). Na primeira etapa, é identificado um modelo que integra um conjunto predefinido de classes que dividem os dados. Este modelo é obtido a partir do conjunto de dados de treino, através da análise dos registos contidos no mesmo. Estes registos são atribuídos a classes predefinidas, que constituem o conjunto de valores possíveis para o atributo de saída (atributo em relação ao qual os registos são classificados).

Na segunda etapa do processo de classificação, o modelo obtido é usado para classificar. Este modelo é assim aplicado ao conjunto de dados de teste, permitindo verificar o seu desempenho na classificação de dados desconhecidos. Os resultados obtidos são analisados, a fim de verificar o desempenho do modelo. A precisão do modelo é determinada com base na quantidade de registos classificados corretamente, comparando o valor real disponível, armazenado no conjunto de dados de teste, com o valor previsto pelo modelo (classe identificada pelo modelo para o registo). Se a precisão do modelo é considerada aceitável, o que depende do domínio de aplicação em causa, então este pode ser utilizado em tarefas de previsão para identificar a classe a que cada registo pertence.

A classificação é considerada uma tarefa de aprendizagem supervisionada, uma vez que o atributo e as classes que vão conduzir o processo de classificação dos dados são conhecidos à partida.

Partindo de um histórico de observações em que a classe alvo é conhecida, são gerados um conjunto de regras que permitem a previsão e a classificação de exemplos futuros (novas instâncias de dados).

Um exemplo típico da aplicação do método de classificação consiste na classificação de clientes de uma instituição bancária em que estes são classificados como "cumpridores" ou "incumpridores" atendendo ao cumprimento ou não cumprimento do pagamento das prestações correspondentes aos empréstimos concedidos.

Existem aplicações do método de classificação em diversos domínios: seleção de clientes para uma campanha, deteção de fraudes, reconhecimento de imagem, diagnóstico antecipado de doenças, catalogação de texto, reconhecimento de *spam*, etc. (Vercellis C, 2009).

Árvores de Classificação (*Classification trees*)

As árvores de classificação ou de decisão, como o próprio nome indica, são constituídas por estruturas em árvore que representam um conjunto de decisões. Possuem uma representação simples, sendo facilmente interpretadas pelos utilizadores.

De acordo com Vercellis (2009), as árvores de classificação são talvez o método mais conhecido e usado em aplicações de *data mining*. As razões para esta popularidade residem no conceito simples e fácil de implementar, na velocidade de processamento, robustez em relação a falta de dados e aos outliers, e na interpretação das regras geradas (Vercellis C, 2009).

Os algoritmos de indução de árvores de decisão permitem gerar regras de classificação dos dados, baseados na informação armazenada na base de dados. Para além de poderem lidar com grandes quantidades de dados, permitem utilizar diretamente o resultado (informação gerada) nelas explícito (Santos & Ramos, 2009).

Uma árvore de decisão integra nós, ramos e folhas. Nos nós, encontram-se os atributos a classificar, enquanto os ramos descrevem os valores possíveis para esses atributos. As folhas da árvore indicam as diversas classes em que cada registo pode ser classificado.

A Figura 4 apresenta um exemplo de uma árvore de decisão que permite verificar a atribuição, ou não atribuição, de um determinado crédito, atendendo às características (atributos utilizados pelo modelo) dos clientes.

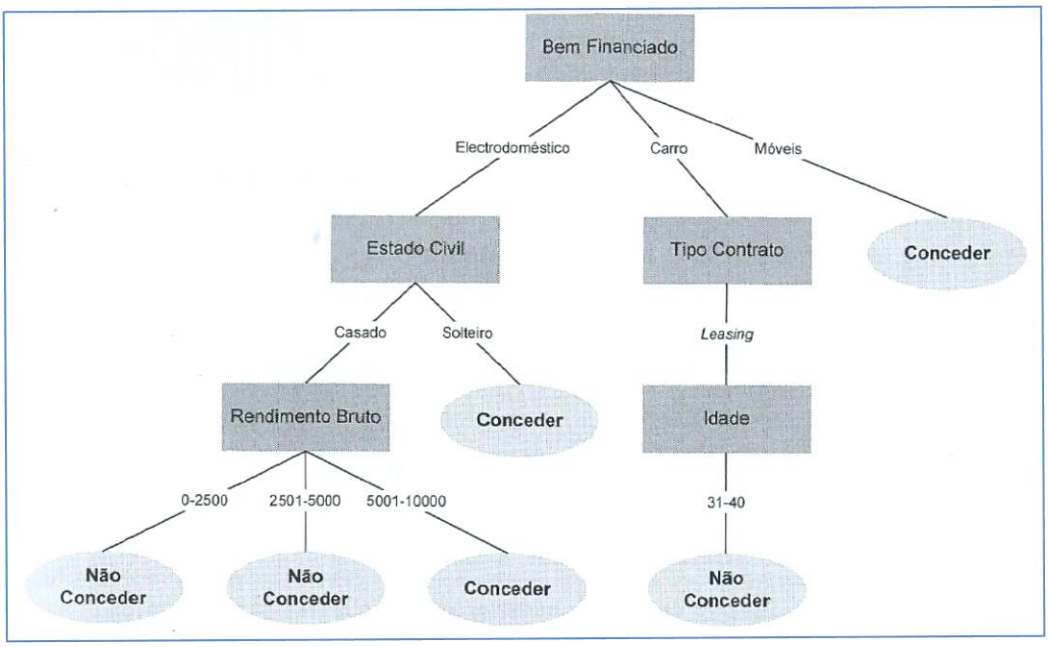


Figura 4 – Exemplo de Árvore de decisão. Fonte: Santos & Ramos, 2009

As árvores de decisão podem ainda ser representadas por conjuntos de regras. Cada folha da árvore dá origem a uma regra, sendo o seu conteúdo apresentado na parte consequente da regra. Na árvore de decisão apresentada na Figura 4, uma das regras associadas à concessão de crédito aos clientes é:

Se Bem Financiado = ”Electrodoméstico” e Estado Civil = ”Casado” e Rendimento Bruto = “5001-10000” Então “Conceder”

A Figura 5 apresenta um exemplo que retrata o processo de indução de regras, isto é, o modo de operação desta técnica. Como já referido anteriormente, na descrição da tarefa de classificação, na primeira etapa, o

conjunto de dados de treino é utilizado para identificar um modelo que classifica os dados atendendo à variável de saída (neste exemplo, Crédito).

Posteriormente, o modelo é utilizado no conjunto de dados de teste para verificar o seu desempenho e avaliar a sua utilidade em tarefas de previsão (Santos & Ramos, 2009) .

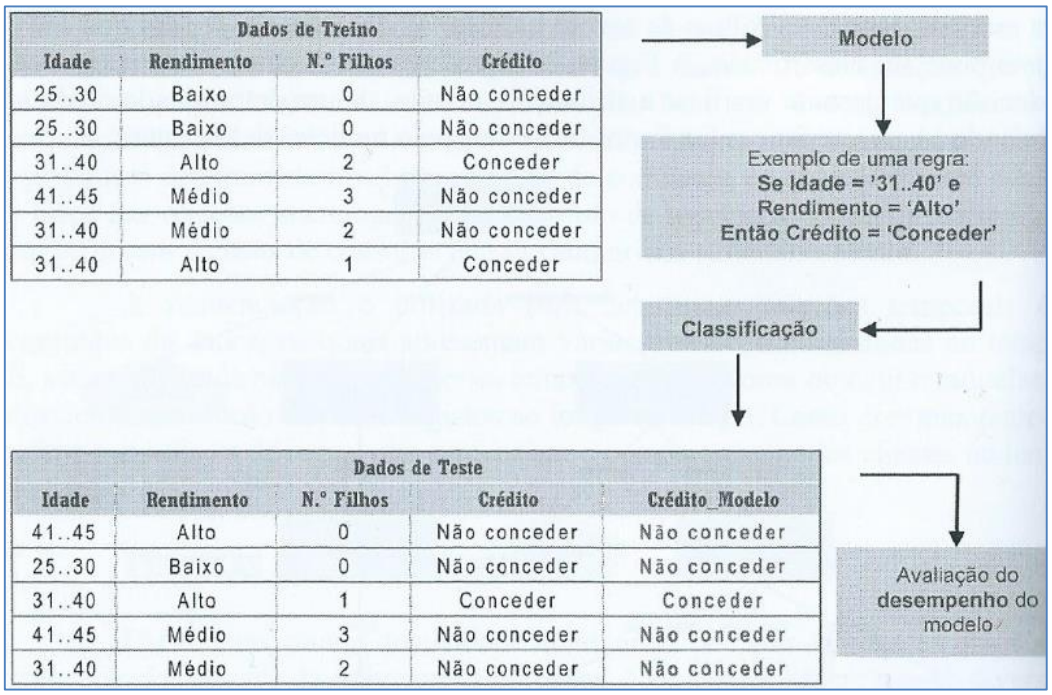


Figura 5 – Indução de regras e posterior avaliação do seu desempenho. Fonte: Santos & Ramos, 2009

Métodos Bayesianos (*Bayesian methods*)

Os métodos *Bayesianos* pertencem à família de modelos de classificação probabilística. Explicitamente calculam a probabilidade *a posteriori* $P(y|\boldsymbol{x})$ de que uma determinada observação pertença a uma determinada classe alvo usando o teorema de Bayes, sendo a probabilidade *a priori* $P(y)$ e as probabilidades condicionadas $P(\boldsymbol{x}|y)$ conhecidas.

Na aprendizagem de máquina, os classificadores *Naive Bayes* são uma família de classificadores probabilísticos simples com base na aplicação de teorema de *Bayes* com fortes pressupostos de independência entre as características.

O método *Naive Bayes* tem sido estudado extensivamente desde os anos 50. Continua a ser um método popular para a categorização de textos, problemas de classificação de documentos como pertencentes a uma ou outra categoria (como de spam ou legítimas, etc.), sendo a classificação feita com base na frequência de palavras. Com o pré-processamento adequado, o método é competitivo neste domínio com métodos mais avançados, incluindo máquinas de vetores de suporte. É também utilizado no diagnóstico médico automático.

Os classificadores *Naive Bayesian* têm por base o pressuposto que as variáveis explicativas são condicionalmente independentes de dada classe alvo. De acordo com Vercelis (2009), a hipótese permite expressar a probabilidade $P(x|y)$ segundo a Equação 20:

$$P(\boldsymbol{x}|y) = P(x_1|y) \times P(x_2|y) \times \ldots \times P(x_n|y) = \prod_{j=1}^{n} P\left(x_j|y\right)$$

Equação 20 – Probabilidade $P(\boldsymbol{x}|\boldsymbol{y})$

Por exemplo, para o caso de aplicação do método Bayesiano no despiste de uma doença ilustrado na Figura 6, considerando uma população de 10000 pessoas, 198+882 pessoas tiveram teste positivo; por isso, a probabilidade de alguém com teste positivo ter a doença é de 198/1080=0,1833 para $P(y|x)$:

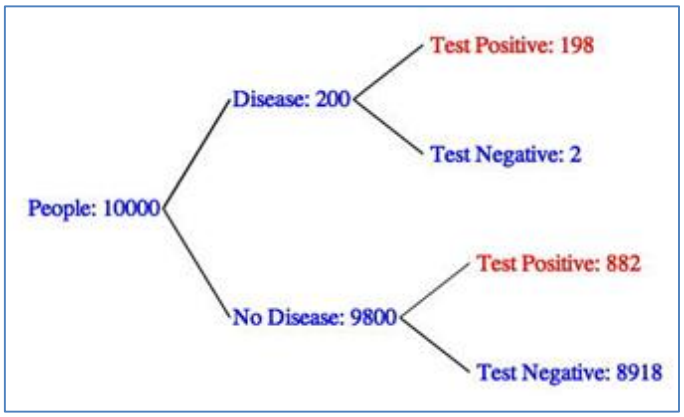


Figura 6 – Exemplo de utilização do método Bayesiano no despite de uma doença Fonte: Rice et al., 2010.

Regressão logística (*Logistic regression*)

A regressão logística foi desenvolvida em 1958 pelo estatístico David Cox e é uma técnica de conversão dos problemas de classificação binária em regressão linear, usando a transformação apropriada (Vercellis C, 2009). O modelo logístico binário é usado para prever uma resposta binária com base em uma ou mais variáveis preditores (características), tornando-se um modelo de classificação probabilística em *machine learning*.

A regressão logística (LR) é um modelo estatístico probabilístico padrão de classificação que tem sido amplamente utilizado em várias disciplinas, nomeadamente computação, marketing, ciências sociais.

Supondo que a variável de resposta y pode tomar os valores {0,1}, como num problema de classificação binário. O modelo de regressão logística postula que a probabilidade *a posteriori* $P(y|x)$ da variável resposta condicionada ao vetor x segue a função logística expressa na Equação 21:

$$P(y = 0|x) = \frac{1}{1 + e^{w'x\prime}}$$

$$P(y = 1|x) = \frac{e^{w'x\prime}}{1 + e^{w'x\prime}}$$

Equação 21 – Modelo de regressão logística - Função logística

Na Figura 7 é apresentado um exemplo de aplicação da função de regressão logística, ilustrando de uma forma gráfica a probabilidade de um aluno passar num exame *versus* o número de horas de estudo.

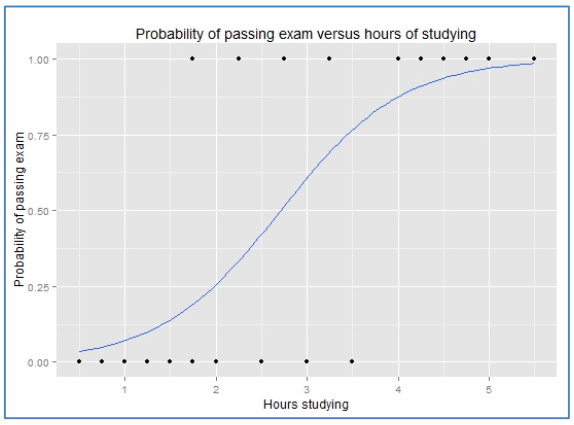


Figura 7 – Gráfico de uma curva de regressão logística apresentando a probabilidade de um aluno passar num exame *versus* o número de horas de estudo. Fonte: wikipedia, 2016.

Redes neuronais artificiais (*Neural networks*)

As redes neuronais artificiais são sistemas de classificação modelados segundo os princípios do sistema nervoso humano. São redes eletrónicas em bruto de "neurónios" com base na estrutura neuronal do cérebro.

Existem dois estágios distintos na utilização destas redes. O primeiro consiste na aprendizagem, no qual a rede é treinada para a execução de determinada tarefa. A segunda fase consiste na previsão, em que a rede é utilizada para classificar registos desconhecidos.

Os neurónios processam os registos um de cada vez, e "aprendem", comparando a sua classificação do registo (que, no início, é em grande parte arbitrária) com a classificação real conhecida do registo. Os erros de classificação inicial do primeiro registo são introduzidos novamente na rede, e utilizados para modificar o algoritmo na segunda vez, e assim por diante para muitas iterações.

De grosso modo, um neurónio de uma rede neural artificial é:

1. Um conjunto de valores de entrada (x_i) e pesos associados (w_i).

2. Uma função de (g) que soma os pesos e mapeia os resultados para uma saída $(f(x))$.

A Figura 8 ilustra o conceito de perceptrão (*perceptron)*, inventado em 1957 por Frank Rosenblatt no laboratório aeronáutico de Cornell. Um perceptrão é a rede neuronal mais simples possível: um modelo computacional de um único neurónio (Vercellis C, 2009). Um perceptrão consiste em uma ou mais entradas, um processador e uma única saída.

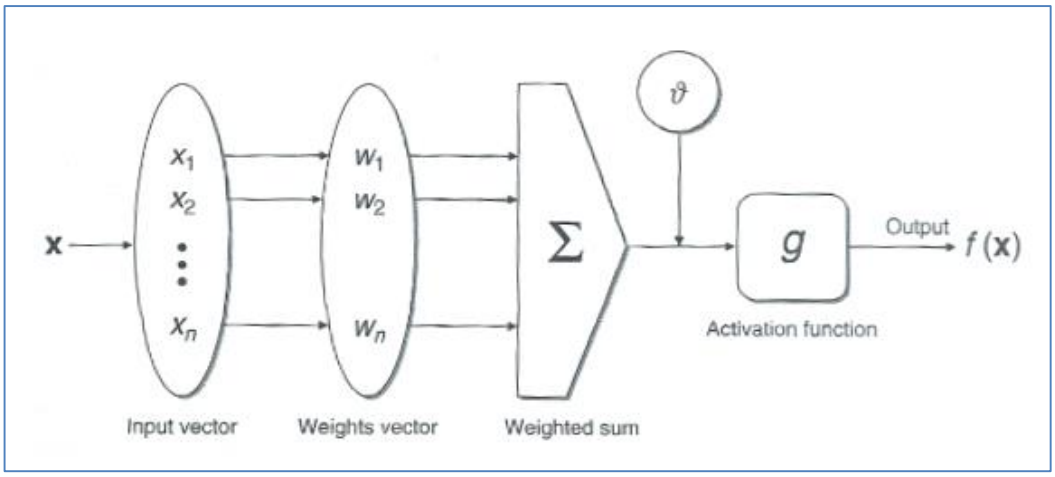


Figura 8 – Operação de uma unidade da rede neuronal. Fonte: Vercellis C, 2009

Um perceptrão segue o modelo "feed-forward", ou seja, as entradas são enviadas para o neurónio, são processadas, e resultam numa saída.

Os neurónios são organizados em camadas: entrada, escondida e de saída. A camada de entrada é composta não de neurónios completos, mas sim de valores do registo que são *inputs* para a próxima camada de neurónios. A camada seguinte é a camada escondida. Várias camadas escondidas podem existir numa rede neuronal. A camada final é a camada de saída, onde existe um nó em cada classe. Através dos resultados da rede na atribuição de um valor para cada nó de saída, o registo de saída é atribuído ao nó de classe com o valor mais elevado.

Nas redes de tipo perceptrão, não existe qualquer nível intermédio (apenas o de entrada e o de saída), o que torna o processo de aprendizagem mais simples, embora condicionando o tipo de tarefa em que estas redes podem ser utilizadas. A sua utilização está restrita a problemas aproximáveis através de funções lineares.

As redes neuronais que apresentam um ou mais níveis intermédios são designadas por perceptrão multinível (cf. Figura 9), permitindo aproximar qualquer função não linear. As redes neuronais de vários níveis constituem uma estrutura mais complexa do que o perceptrão, pois incluem os seguintes componentes (Vercellis C, 2009):

Input nodes: O objetivo dos *input* nodes é receber os valores explanatórios dos atributos para cada observação. Normalmente o número de *input nodes* é igual ao número de variáveis explanatórias.

Hidden nodes: Aplicam transformações aos valores de *input* dentro da rede.

Output nodes: Recebem conexões dos *hidden nodes* ou dos *input nodes* e devolvem o valor que corresponde à previsão da variável de resposta.

Basicamente cada nodo da rede neuronal opera como um perceptrão. O método que determina os pesos de todas as conexões e a distorção dos nodos, denomina-se por *backpropagation algorithm*.

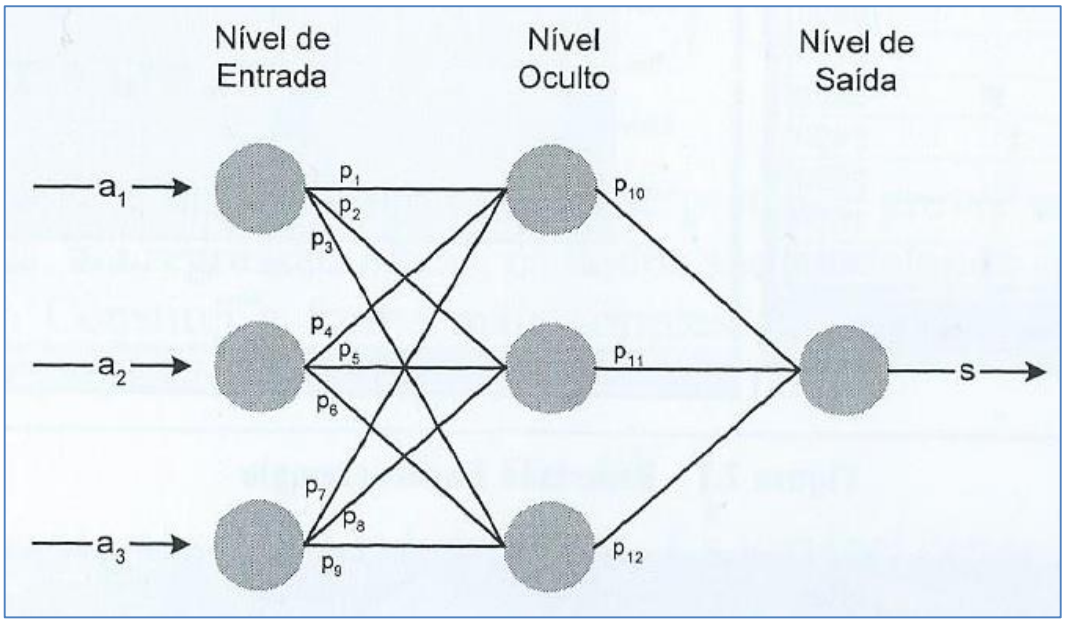


Figura 9 – Rede neuronal artificial. Fonte: Santos & Ramos, 2009

A aprendizagem de uma rede é iniciada com a atribuição de pesos semelhantes a todas as ligações da rede. A rede é então treinada com o conjunto de dados de treino. Em cada iteração do processo de aprendizagem, a saída da rede é comparada com a saída desejada (explícita nos casos conhecidos, armazenados no conjunto de dados de treino). O resultado desta comparação é propagado na rede, sendo os pesos das ligações gradualmente ajustados. À medida que a aprendizagem progride, a rede fica cada vez mais precisa na replicação dos resultados conhecidos (Santos e Ramos, 2009).

A Figura 10 ilustra um exemplo prático da utilização de uma rede neuronal para auxiliar a decisão de esperar ou não pelo atendimento num restaurante. O domínio do problema é formado por um conjunto de *perceptron* (P) constituído pelos seguintes atributos: Possui sala de espera? Estou com Fome? O preço é acessível?

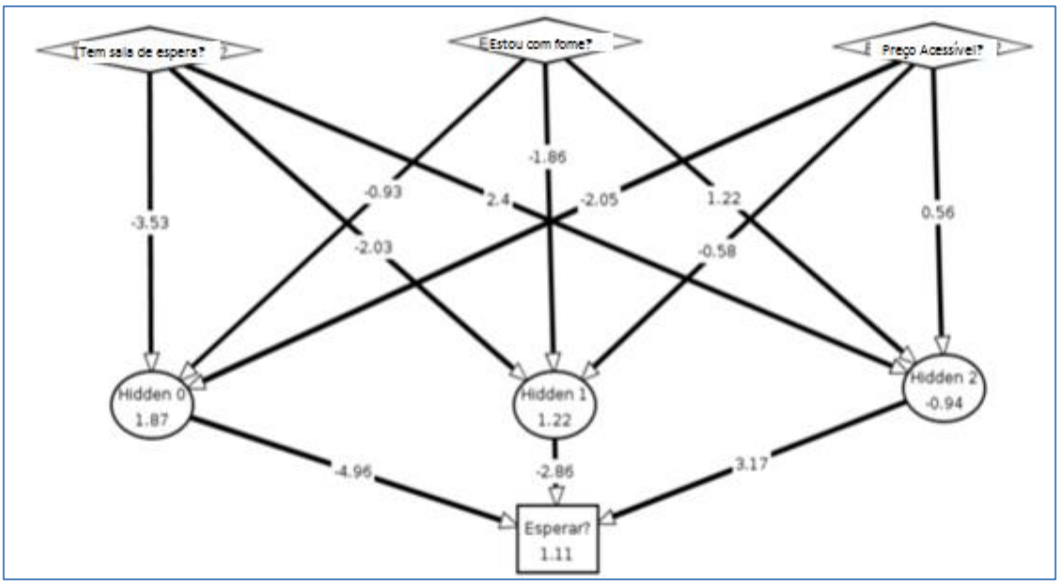


Figura 10 – Exemplo prático de uma rede neural. Fonte: Arnaldo-jr, 2015.

Máquina de vetores de suporte (*Support vector machines*)

Uma máquina de vetores de suporte é uma família de métodos de separação para classificação e regressão desenvolvidos no contexto da teoria de aprendizagem estatística. Verificou-se que conseguem melhor performance em termos de exatidão em relação a outros classificadores e são eficientemente escaláveis em grandes problemas. Identificam uma série de exemplos designados por vetores de suporte, que apresentam as observações mais representativas de cada classe (Vercellis C, 2009).

Em aprendizagem máquina, as máquinas de vetor de suporte são modelos de aprendizagem supervisionados com algoritmos de aprendizagem associados que analisam os dados e reconhecem padrões, utilizados para a classificação e análise de regressão.

Dado um conjunto de exemplos de treino, cada um marcado por pertencer a uma das duas categorias, um algoritmo de treino SVM (*Support Vector Machine*) constrói um modelo que atribui novos exemplos a uma ou outra categoria, tornando-se um classificador linear binário não-probabilístico. Um modelo SVM é uma representação dos pontos no espaço, mapeados de modo a que os exemplos das categorias separadas são divididas por um intervalo claro que é tão largo quanto possível. Novos exemplos são então mapeados no mesmo espaço e a previsão da categoria a que pertencem é feita com base no lado do intervalo em que caem.

Na escolha de uma função particular, deve-se considerar a função que produz o menor erro com o conjunto de treino. Seja o risco empírico de f (R_{emp}), definido pela Equação 22. Sendo $V(y_i, f(x_i))$ a função perda que mede a discrepância entre os valores devolvidos pela função $f(x)$ e os valores atuais da classe y. O processo de busca por uma função f que apresente menor R_{emp} é denominado Minimização do Risco Empírico:

$$R_{emp}(f) = \frac{1}{t} + \sum_{i=1}^{t} V(y_i, f(x_i))$$

Equação 22 – Função de Minimização do Risco Empírico.

A Figura 11 ilustra o resultado do processo de máquina de vetores de suporte na deteção de faces (Osuna, Freund, & Girosi, 1999). A imagem apresentada na Figura 11 foi analisada e foram reconhecidas as faces dos jogadores como pertencentes à classe faces.

Figura 11 – Exemplo prático do processo de máquina de vetores de suporte na deteção de faces. Fonte: Osuna, Freund, & Girosi, 1999

3.5.4. Regras de associação (*Association rules*)

São modelos de aprendizagem não supervisionada, que podem ser usados quando o conjunto de dados não inclui um atributo alvo. O objetivo destes métodos é identificar padrões regulares e recorrências em grandes conjuntos de transações. São frequentemente usados na análise de vendas de carrinho de compras e navegação em *websites* (Vercellis C, 2009).

Regras de associação simples (*Single Association rules*)

O método de Regras de Associação simples é um método popular para descobrir relações interessantes entre as variáveis em grandes bases de dados. Pretende-se identificar regras fortes em bases de dados usando diferentes medidas de interesse.

A descoberta de regras de associação em grandes bases de dados foi inicialmente analisada por Agrawal, Imieliński, & Swami (1993), com o objetivo de identificar uma conclusão (por exemplo, a compra de um produto) com um conjunto de condições (por exemplo, a compra de outros produtos). As regras

de associação permitem encontrar relacionamentos entre os atributos existentes numa base de dados, representando-os na forma de uma regra:

Se X Então Y ou “X => Y”

Por exemplo, a regra {cebolas, batatas} => {hambúrguer} encontrada nos dados das transacções de vendas de um supermercado indicaria que, se um cliente compra cebolas e batatas em conjunto, é suscetível também de comprar carne de hambúrguer. Esta informação pode ser usada como base para decisões sobre atividades de marketing, nomeadamente em colocações de preços ou no lançamento de produtos promocionais.

No exemplo apresentado na Figura 12, a regra “Pão & Manteiga >= Leite (2:50%, 1)” indica que os clientes que compram o produto “Pão” juntamente com o produto “Manteiga”, também compram o produto “Leite”. Esta regra apresenta um suporte de 50%, o que significa que metade dos registos analisados verificam a referida regra. A confiança da regra apresenta o valor de 1 (100%), uma vez que em todos os registos em que foi verificada a ocorrência das compras “Pão” e “Manteiga”, também foi confirmada a aquisição de “Leite” (Santos & Ramos, 2009).

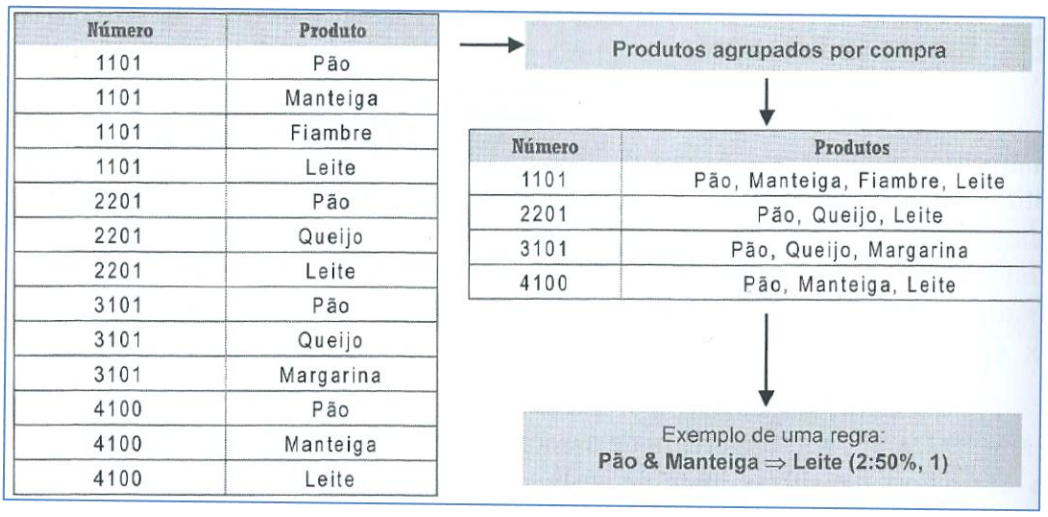
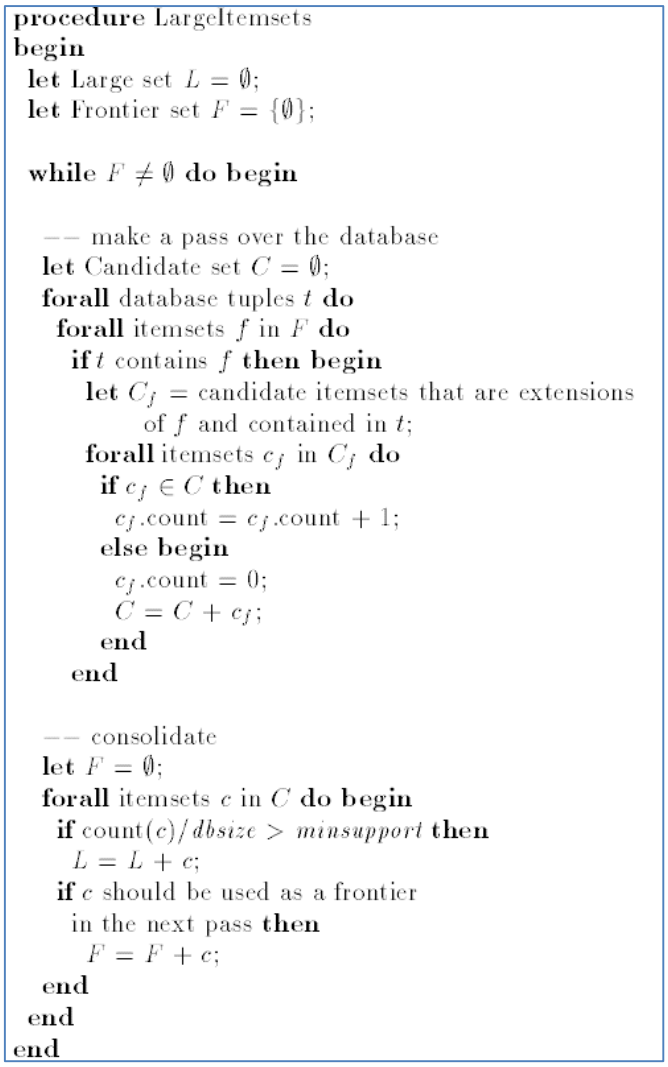


Figura 12 – Processo de indução de regras de associação. Fonte: Santos & Ramos, 2009

As regras de associação são também atualmente utilizadas em outras áreas de aplicação, nomeadamente na prospeção do uso da Web, deteção de intrusão, produção contínua, e bioinformática. Em contraste com a prospeção

sequencial, a aprendizagem de regras de associação normalmente não considera a ordem dos itens ou dentro de uma transação ou através de transações.

A Figura 13 apresenta o algoritmo do modelo para grandes conjuntos de dados. Dado um conjunto de itens L, um conjunto de itens X + Y de artigos de L é dito ser uma extensão do conjunto de itens X se $X \cap Y = 0$. O parâmetro *dbsize* é o número total de *tuples* (lista ordenada de elementos finitos) na base de dados. O algoritmo faz múltiplas passagens sobre a base dados. A *frontier set* para uma passagem é constituída pelos conjuntos de itens que são estendidos durante a passagem. Estes conjuntos de itens, designados por *itemsets* candidatos, são derivados dos *tuples* da base de dados e os conjuntos de itens contidos no conjunto de fronteira. Associado a cada conjunto de itens, existe um contador que armazena o número de transações em que o correspondente *itemset* apareceu.



```
procedure LargeItemsets
begin
 let Large set L = ∅;
 let Frontier set F = {∅};

 while F ≠ ∅ do begin

  -- make a pass over the database
  let Candidate set C = ∅;
  forall database tuples t do
   forall itemsets f in F do
    if t contains f then begin
     let C_f = candidate itemsets that are extensions
          of f and contained in t;
     forall itemsets c_f in C_f do
      if c_f ∈ C then
       c_f.count = c_f.count + 1;
      else begin
       c_f.count = 0;
       C = C + c_f;
      end
    end

  -- consolidate
  let F = ∅;
  forall itemsets c in C do begin
   if count(c)/dbsize > minsupport then
    L = L + c;
   if c should be used as a frontier
    in the next pass then
     F = F + c;
  end
 end
end
```

Figura 13 – Algoritmo Agrawal, Imieliński, & Swami, 1993. Fonte: Agrawal, Imieliński, & Swami, 1993

Algoritmo Apriori (*Apriori algorithm*)

O algoritmo Apriori, é um algoritmo para prospeção da frequência de itens e regras de associação e de aprendizagem sobre conjuntos de dados transacionais. Procede, identificando os itens individuais frequentes na base de dados e expande-os para conjuntos de itens maiores, desde que esses conjuntos de itens apareçam muitas vezes na base de dados. Os conjuntos de itens frequentes determinados pelo algoritmo podem ser usados para determinar regras de associação que destacam as tendências gerais da base de dados (Agrawal & Srikant, 1994).

O primeiro passo do algoritmo (cf. Figura 14) consiste em contar as ocorrências dos itens para determinar o *itemset*. As passagens subsequentes, vamos chamar-lhe k, consistem em 2 fases. Na primeira fase, o *itemset* $L_{(k-1)}$ é usado para gerar o *itemset* candidato C_k. No segundo passo, a base de dados é pesquisada e os candidatos C_k são contados.

$L_1 = \{\text{large 1-itemsets}\};$
for ($k = 2$; $L_{k-1} \neq \emptyset$; k++) **do begin**
  $C_k = \text{apriori-gen}(L_{k-1})$; // New candidates
  forall transactions $t \in \mathcal{D}$ **do begin**
    $C_t = \text{subset}(C_k, t)$; // Candidates contained in t
    forall candidates $c \in C_t$ **do**
      c.count++;
  end
  $L_k = \{c \in C_k \mid c.\text{count} \geq \text{minsup}\}$
end
Answer $= \bigcup_k L_k$;

Figura 14 – Algoritmo Apriori. Fonte: Agrawal & Srikant, 1994

3.5.5. Segmentação (*Clustering*)

A segmentação, também designada por *clustering,* permite identificar um conjunto de classes ou segmentos (*clusters*) que dividem os dados analisados.

É um método de aprendizagem não supervisionada, uma vez que o utilizador não tem qualquer influência na definição das classes, que são definidas a partir dos dados. Os segmentos surgem de agrupamentos que são detetados nos dados e que obedecem a métricas de similaridade.

Ao definir métricas apropriadas e induzir noções de distância e semelhança/verosimilhança num conjunto de observações, o propósito do *clustering* é agrupar registos nos segmentos ou *clusters.*

Os modelos de *clustering* são utilizados em vários domínios (Banca, Retalho, Seguros, etc), pois os *clusters* gerados fornecem informação útil para interpretação de fenómenos de interesse. Por exemplo, na área do retalho, a segmentação de clientes pode ser feita para agrupar clientes pelos comportamentos de compra; nos setores da Banca e Seguros, a segmentação de clientes pode ser feita para distinguir perfis de clientes (Vercellis C, 2009). Os

clusters podem também ser úteis para a descoberta de *outliers*, utilizando estimadores robustos (Pereira & Pires, 2002).

Diversos algoritmos podem ser utilizados na identificação de segmentos nos dados ou *clusters.* Nas seções seguintes descrevem-se com mais detalhe esses algoritmos.

Métodos de particionamento (*Partition methods*)

Os métodos de particionamento desenvolvem uma subdivisão do conjunto de dados num número determinado de subconjuntos. São adequados para conjuntos de dados de pequena ou média dimensão.

Os métodos de particionamento iniciam com a atribuição das M observações disponíveis a K *clusters*. Depois interactivamente realocam observações a outros *clusters*, para que a qualidade da subdivisão seja melhorada. Embora várias medidas de homogeneidade alternativas possam ser usadas, todos os critérios expressam a homogeneidade das observações do mesmo *cluster* e a sua heterogeneidade em relação a outros clusters.

Os métodos *K-means* e *K-medoids* constituem dois dos melhores algoritmos de particionamento (Vercellis C, 2009).

- *K-means* - O algoritmo utiliza uma técnica de refinamento iterativo. O algoritmo *k-means* é também referido como o algoritmo de Lloyd, particularmente na comunidade científica de Informática. A implementação desta técnica visa a construção de partições, dos objetos armazenados na base de dados, num conjunto de k classes (ou segmentos), sendo k um parâmetro de entrada. Cada classe é representada pelo seu centro de gravidade. Para determinar as classes, cada registo é transformado num ponto do espaço, apresentando este tantas dimensões quantos os atributos em análise. O valor de cada atributo é interpretado como a distância da origem até à sua localização num dado eixo (Santos & Ramos, 2009). No k-means, o processo de identificação das classes é iniciado com centróides (que representam a média de cada

segmento) em posições aleatórias, as quais são otimizadas iterativamente através da movimentação dos centros. Na primeira iteração do processo, selecionam-se *k* pontos para serem as sementes dos segmentos. Na segunda etapa, cada registo é atribuído ao segmento cujo centróide lhe é mais próximo. O próximo passo consiste em calcular os centróides dos novos segmentos, o que corresponde a calcular a média dos pontos que integram a classe. Uma vez que o *k-means* requer que os atributos integrem valores numéricos, a média do segmento é facilmente calculada através do valor médio dos seus pontos. Após identificação dos novos segmentos, com o respetivo centróide, é necessário voltar a verificar em que segmento cada registo se enquadra, atendendo à sua proximidade aos novos centróides. O processo de atribuição dos registos aos segmentos e o recalcular dos centróides prossegue até que não sejam verificadas quaisquer alterações na formação dos segmentos. No exemplo apresentado na Figura 15, verifica-se, que no conjunto inicial de dados, são selecionados três pontos aleatórios, como sendo os centróides dos segmentos, tendo-se definido K=3 (a). Para estes pontos, verifica-se quais os registos que integram cada um dos segmentos (b). No passo seguinte (c), são ajustados os centróides, através da verificação do ponto médio dos elementos de um dado segmento. Para as novas posições dos centróides é verificado a que segmento cada registo pertence, atendendo às alterações observadas na posição dos centróides (d). Este processo é sucessivamente repetido até que não sejam verificadas quaisquer mudanças na posição dos centróides e, como tal, nos registos que integram cada segmento (Santos & Ramos, 2009).

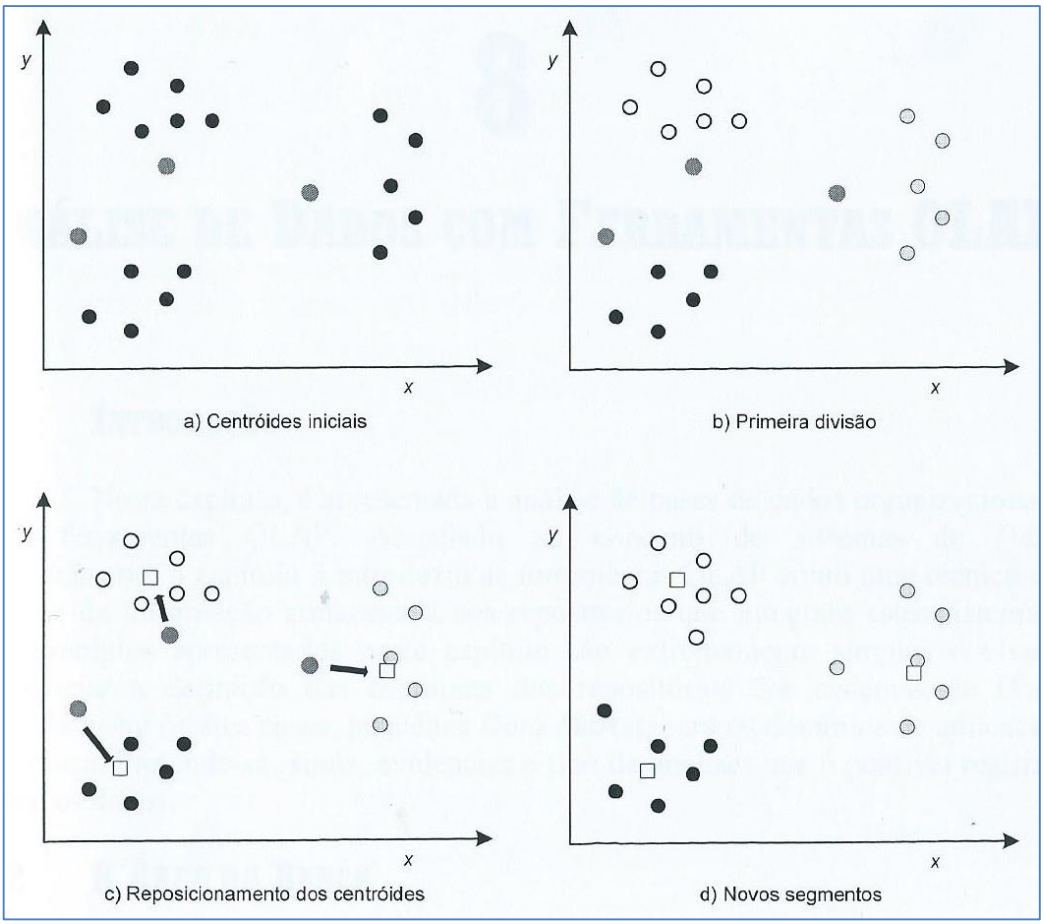


Figura 15 – Processo de identificação dos segmentos. Fonte: Santos & Ramos, 2009

•*K-medoids* - é um algoritmo de agrupamento relacionado com o algoritmo K-means e o algoritmo medoidshift. Ambos os algoritmos são particionais (dividindo o conjunto de dados em grupos) e ambos tentam minimizar o erro médio quadrado, da distância entre os pontos marcados para estar num *cluster* e um ponto designado como o centro desse cluster. K-medoids é também uma técnica de particionamento de *clustering* que agrupa o conjunto de dados de n objetos em k clusters com k conhecido a priori. Pode ser considerado mais robusto ao ruído e a valores discrepantes em relação ao K-means, pois minimiza a soma das diferenças entre pares gerais em vez de uma soma das distâncias euclidianas quadradas. A medoide de um conjunto de dados finito é um ponto de dados a partir desse conjunto, cuja dissimilaridade média de todos os pontos de dados é mínima ou seja, é o ponto mais central no conjunto. Medoids são

objetos representativos de um conjunto de dados ou um *cluster* com um conjunto de dados cuja dissimilaridade média de todos os objetos no cluster é mínima. O conceito de medoids é semelhante ao conceito de média ou centróides, mas os medoids são sempre os membros do conjunto de dados. O termo é utilizado na ciência da computação em algoritmos de agrupamento de dados.

A implementação mais comum de agrupamento *K-medoide* é a seguinte:

1- Inicialização: selecionar aleatoriamente k dos n pontos de dados como os medoides
2- Atribuição: associar cada ponto de dados para o medoide mais próximo.
3- Atualização: associar cada m medoide e cada ponto de dados e calcular o custo total da configuração (isto é, a dissimilaridade média para todos os pontos de dados associados aos m). Escolher o medoide com o menor custo da configuração.

Métodos hierárquicos (*Hierarchical methods*)

Os métodos hierárquicos são baseados numa estrutura em árvore e não necessitam que o número de *clusters* seja determinado à partida. De maneira a avaliar a distância entre dois *clusters*, recorrem a cinco medidas alternativas: distância mínima, distância máxima, distância média, distância entre os centróides e variância mínima de Ward (Vercellis C, 2009).

- *Agglomerative hierarchical methods* – são técnicas *bottom-up* em que cada observação inicial representa um *cluster* distinto. Estes *clusters* são agregados durante as interações seguintes.
  Algoritmo *Agglomerative*:
  1- Inicialização: todas as observações constituem um *cluster*, e a distância entre *clusters* é dada por uma matriz.
  2- A distância mínima entre *clusters* é calculada e são agregadas as observações.

3- A distância entre o novo *cluster* e os outros *clusters* é calculada.
4- São repetidos os passos até só haver 1 *cluster*, ou o número de *clusters* pré-definidos.

O exemplo ilustrado na Figura 16, corresponde à aplicação do algoritmo do método hierárquico aglomerativo para a segmentação de cidades Italianas tendo em conta a sua localização espacial no mapa.

O par mais próximo das cidades é MI e TO, a uma distância de 138. Estas cidades são agrupadas num *cluster* único designado por "MI/TO".

	BA	FI	MI	NA	RM	TO
BA	0	662	877	255	412	996
FI	662	0	295	468	268	400
MI	877	295	0	754	564	138
NA	255	468	754	0	219	869
RM	412	268	564	219	0	669
TO	996	400	138	869	669	0

Figura 16 – Exemplo do início do processo *Agglomerative hierarchical* para cidades Italianas. Fonte: Cristian Mihaescu, 2010.

Ao aplicar este processo do par mais próximo iterativamente, as cidades são agrupadas em dois *clusters* designados por "MI/TO" e “BA/FI/NA/RM”, conforme ilustrado na Figura 17.

	BA/FI/NA/RM	MI/TO
BA/FI/NA/RM	0	295
MI/TO	295	0

Figura 17 – Exemplo do fim do processo *Agglomerative hierarchical* para cidades Italianas. Fonte: Cristian Mihaescu, 2010.

- *Divisive hierarchical methods* – são técnicas *bottom-down* em que todas as observações começam num único *cluster*, e as divisões são realizadas de forma recursiva quando se desce na hierarquia. As divisões são feitas de maneira a que a distância seja máxima entre *clusters*. O único *cluster* que tem todas as observações no início é dividido em 2 *clusters*, de maneira a que a distância entre os dois seja máxima; à medida que se vai avançando neste método, a necessidade de computação vai aumentado exponencialmente.

4. Deteção de clientes e produtos *outliers* na empresa

4.1. Apresentação da Empresa

A empresa de distribuição farmacêutica onde foi desenvolvido o estudo, tem sede em Portugal, é uma das maiores empresas de comercialização e distribuição farmacêutica de âmbito nacional. A empresa garante diariamente às farmácias o acesso aos vários produtos por ela comercializados e distribuídos. Os produtos comercializados são: produtos farmacêuticos, cosmética, perfumaria, dietética, medicina natural, dispositivos médicos, acessórios e matérias-primas relacionadas.

De salientar ainda que a empresa está integrada, num grupo mundial, que é um dos líderes do mercado no comércio, logística e prestação de serviços farmacêuticos e cuidados de saúde a nível mundial.

É ainda importante referir que a empresa tem manifestado ao longo dos anos uma atitude positiva, proactiva e dinamizadora, tendo crescido de forma sustentada e ocupando atualmente uma posição de destaque no mercado. Presentemente existem 6 *players* principais a competir no mercado, que representam 92% das vendas do setor.

Atualmente a empresa possui vários armazéns que garantem uma cobertura eficaz de todo o território nacional, disponibilizando uma vasta gama de produtos que se encontram localizados em cada armazém de acordo com a procura local.

Por último, todas as instalações da empresa cumprem com as exigências legais, detêm as autorizações das Entidades Reguladoras competentes e possuem modernos sistemas de gestão logística, armazenamento, aviamentos automáticos e comunicação.

4.2. Comparação dos métodos Box-plot e Z-score modificado

Como apresentado no Capítulo 2, constata-se que existem vários métodos para a deteção de *outliers*, entre os quais, o método de *Box-plot* e o teste Z-score modificado. Nesta secção, com base no estudo apresentado nesse capítulo, realiza-se apenas a comparação entre os 2 métodos acima.

Para realizar a comparação entre os 2 métodos, começa-se por estratificar a totalidade dos clientes da empresa (1325 farmácias), segundo o seu potencial de vendas, isto é, os clientes são rotulados (por ordem decrescente de vendas) como: Clientes C1, Clientes C2, Clientes C3 e Clientes C4. De seguida, e por motivos de confidencialidade (dado o tipo de produto que está em análise), os três medicamentos MSRM são designados genericamente no estudo por: Produto A, Produto B e Produto C.

Por último, escolhe-se um estrato de clientes (C1, C2, C3 ou C4) e para todos os clientes pertencentes a esse estrato analisam-se os valores obtidos pelos 2 métodos (método de Box-plot e Z-score modificado) para cada um dos 3 produtos (A, B e C). Os resultados obtidos encontram-se na Tabela 4, 5 e 6.

Tabela 4 – Comparação dos métodos para o Produto A

Cliente	Quantidade encomenda	Z-scores Modificados	Box-plot
4572113414	125	outlier	outlier severo
10486	72	outlier	outlier moderado
4386678	71	outlier	outlier moderado
10888	66	outlier	outlier moderado
10651	50	outlier	-
7027118572	50	outlier	-
11924	39	outlier	-
4646706415	35	outlier	-
284867569	31	-	-
10675774	30	-	-
…	…	…	…

Tabela 5 – Comparação dos métodos para o Produto B

Cliente	Quantidade encomenda	Z-scores Modificado	Box-plot
839337868	610	outlier	outlier severo
7840084	413	outlier	outlier severo
4389540	162	outlier	outlier severo
270972	108	outlier	outlier severo
271521	91	outlier	outlier moderado
50310284	72	outlier	outlier moderado
6926712970	70	outlier	outlier moderado
13064039	70	outlier	outlier moderado
6798	46	outlier	-
3978999733	43	outlier	-
2310	38	outlier	-
1032896927	36	outlier	-
7646	35	outlier	-
1967	33	-	-
9348490916	33	-	-
…	…	…	…

Tabela 6 – Comparação dos métodos para o Produto C

Cliente	Quantidade encomenda	Z-scores Modificados	Box-plot
4697030932	886	outlier	outlier severo
839337868	610	outlier	outlier severo
4169086059	575	outlier	outlier severo
5217	378	outlier	outlier severo
36573719	309	outlier	outlier severo
8266	281	outlier	outlier severo
129284257	276	outlier	outlier severo
15260492	263	outlier	outlier severo
2115203298	237	outlier	outlier severo
10659	237	outlier	outlier severo
10765	234	outlier	outlier moderado
4958	207	outlier	outlier moderado
267988	185	outlier	outlier moderado
1503187383	180	outlier	outlier moderado
6987914	179	outlier	outlier moderado
170740439	174	outlier	outlier moderado
4842092193	155	outlier	outlier moderado
267640	146	outlier	outlier moderado
5834	143	outlier	outlier moderado

3277	**140**	**outlier**	**-**
8140568809	**132**	**-**	**-**
2838	**129**	**-**	**-**
...	**...**	**...**	**...**

4.2.1. Conclusão da comparação dos dois métodos

Sabe-se que, matematicamente o método Z-score modificado é mais robusto do que o método *Box-plot*, isto porque a distância interquartis do método *box-plot* é muito afetada no caso de haver valores infinitos.

Na prática, depois de verificar os valores das quantidades encomendadas, conclui-se que o método Z-score modificado identificou demasiados possíveis *outliers*. A razão para tal é simples, quando um cliente (farmácia) não recebe o produto (medicamento) encomendado, continua a pedi-lo, aumentado assim a quantidade encomendada.

Em consequência disto e com base no conhecimento que se tem do negócio, só devem ser considerados como *outliers* os *outliers* severos identificados pelo método *Box-plot*. Por esta razão nas secções seguintes, o estudo de detecção de *outliers* (clientes e produtos) é feito usando o método de *Box-plot*.

4.3. Deteção de *outliers* (clientes e produtos) – método *Box-plot*

Nesta secção, realiza-se a avaliação dos resultados obtidos para os dados das vendas em outubro de 2013 e outubro de 2014 (cf. Anexos 1, 2, 3, 4, 5 e 6), dos produtos (medicamentos) A, B e C, com o objetivo de verificar se existem clientes (farmácias) e/ou produtos (medicamentos) *outliers*.

Como referido atrás (Secção 4), o método selecionado para a identificação dos *outliers* é o método Box-plot. Utilizou-se como *software* estatístico o SPSS - *Statistical Package for the Social Sciences*, versão 20.

4.3.1. Clientes *outliers* por produto

Nesta secção, vai-se identificar quais são os clientes (farmácias) *outliers* para cada um dos produtos (medicamentos) A, B e C, com recurso ao uso do software SPSS.

4.3.1.1. Produto A

Para o produto A em outubro de 2013 foram detetados 79 clientes *outliers* (cf. Tabela 7) equivalendo a 7,96% dos 993 que encomendaram este produto (medicamento).

Tabela 7 – Nº clientes *outliers* para o Produto A (outubro 2013)

Nº clientes encomendou (out.2013)		
993		
Produto A	**Total**	**%Clientes**
Moderados	28	2,82
Severos	51	5,14

Relativamente, a outubro de 2014 foram detetados 64 clientes *outliers* (cf. Tabela 8) equivalendo a 7,41% dos 864 clientes que encomendaram este produto (medicamento).

Tabela 8 – Nº clientes *outliers* para o Produto A (outubro 2014)

Nº clientes encomendou (out.2014)		
864		
Produto A	**Total**	**%Clientes**
Moderados	35	4,05
Severos	29	3,36

4.3.1.2. Produto B

Para o produto B em outubro de 2013 foram detetados 62 clientes *outliers* (cf. Tabela 9) equivalendo a 7,46% dos 832 clientes que encomendaram este produto (medicamento).

Tabela 9 – Nº clientes *outliers* para o Produto B (outubro 2013)

Nº clientes encomendou (out.2013)		
832		
Produto B	**Total**	**%Clientes**
Moderados	26	3,13
Severos	36	4,33

Já em outubro de 2014 foram detetados 95 clientes *outliers* equivalendo (cf. Tabela 10) a 11,34% dos 838 clientes que encomendaram este produto (medicamento).

Tabela 10 – Nº clientes *outliers* para o Produto B (outubro 2014)

Nº clientes encomendou (out.2014)		
838		
Produto B	**Total**	**%Clientes**
Moderados	48	5,73
Severos	47	5,61

4.3.1.3. Produto C

Relativamente ao produto C, em outubro de 2013 foram detetados 87 clientes *outliers* (cf. Tabela 11) equivalendo a 9,79% dos 888 clientes que encomendaram este produto (medicamento).

Tabela 11 – Nº clientes *outliers* para o Produto C (outubro 2013)

Nº clientes encomendou (out.2013)		
888		
Produto C	**Total**	**%Clientes**
Moderados	32	3,60
Severos	55	6,19

Em outubro de 2014 foram detetados 83 clientes *outliers* (cf. Tabela 12) equivalendo a 8,71% dos 953 clientes que encomendaram este produto (medicamento).

Tabela 12 – Nº clientes *outliers* para o Produto C (outubro 2014)

Nº clientes encomendou (out.2014)		
953		
Produto C	**Total**	**%Clientes**
Moderados	43	4,51
Severos	40	4,20

4.3.2. Quantidades encomendadas pelos clientes *outliers*

Nesta secção, para os clientes *ouliers* detetados na secção 4.3.1, vai-se avaliar a quantidade movimentada por eles, relativamente a cada um dos produtos (medicamentos) A, B e C. O objetivo é agora, para esses clientes (farmácias) identificar quais os produtos *outliers* (isto é, quais as quantidades de medicamentos consideradas *outliers* pelas farmácias).

4.3.2.1. Produto A

Para o produto A em outubro de 2013, os 79 clientes considerados *outliers* encomendaram 62,07% do total da quantidade encomendada pelos 993 clientes que encomendaram o produto.

Tabela 13 - Quant. encom. pelos clientes *outliers* - Produto A (outubro 2013)

Total quantidade clientes encomendaram (out.2013)		
19117		
Produto A	**Total**	**%Quantidade**
Moderados	825	4,32
Severos	11040	57,75

Em outubro de 2014, para os 64 clientes *outliers* detetados relativos aos 864 que encomendaram o produto, a percentagem total da quantidade encomendada foi 40,82%.

Tabela 14 - Quant. encom. pelos clientes *outliers* - Produto A (outubro 2014)

Total quantidade clientes encomendaram (out.2014)		
21642		
Produto A	**Total**	**%Quantidade**
Moderados	2283	10,55
Severos	6551	30,27

4.3.2.2. Produto B

Para o produto B, em outubro de 2013, os 62 clientes considerados *outliers* encomendaram 55,95% do total de quantidade encomendada pelos 832 clientes que encomendaram o produto.

Tabela 15 - Quant. encom. pelos clientes *outliers* - Produto B (outubro 2013)

Total quantidade clientes encomendaram (out.2013)		
18362		
Produto B	**Total**	**%Quantidade**
Moderados	1307	7,12
Severos	8966	48,83

Relativamente ao mês de outubro de 2014, para os 95 clientes *outliers* dos 838 que encomendaram o produto, a percentagem da quantidade encomendada deste produto foi 61,68%.

Tabela 16 - Quant. encom. pelos clientes *outliers* - Produto B (outubro 2014)

Total quantidade clientes encomendaram (out.2014)

20275		
Produto B	**Total**	**%Quantidade**
Moderados	2364	11,66
Severos	10142	50,02

4.3.2.3. Produto C

Por último, para o produto C em outubro de 2013, os 87 clientes considerados *outliers* encomendaram 65,36% do total de quantidade encomendada pelos 888 clientes que encomendaram o produto.

Tabela 17 - Quant. encom. pelos clientes *outliers* - Produto C (outubro 2013)

Total quantidade clientes encomendaram (out.2013)		
18092		
Produto C	**Total**	**%Quantidade**
Moderados	1137	6,28
Severos	10688	59,08

E, em outubro de 2014, dos 83 clientes *outliers* relativos aos 953 que encomendaram o produto, a percentagem total da quantidade encomendada é 42,56%.

Tabela 18 - Quant. encom. pelos clientes *outliers* - Produto C (outubro 2014)

Total quantidade clientes encomendaram (out.2014)		
37930		
Produto C	**Total**	**%Quantidade**
Moderados	4321	11,39
Severos	11824	31,17

4.3.3. Principais resultados para clientes/quantidades *outliers*

As percentagens das quantidades encomendadas em outubro de 2013, pelos clientes detetados como *outliers* (que no máximo foram 9,79% do nº total clientes), foram sempre superiores a 55% do total das quantidades encomendadas (tendo em 2 dos produtos atingido mais de 62%).

Tabela 19 - Percentagens Clientes/Quantidades em outubro 2013

outubro 2013		
Produto	**% Clientes**	**% Quantidade Encomendada**
Produto A	7,96	62,07
Produto B	7,46	55,95

Produto C	9,79	65,36

Relativamente ao período de outubro de 2014, as percentagens das quantidades encomendadas pelos clientes detetados como *outliers* (que no máximo foram 11,34% do nº total clientes), foram sempre superiores a 40% do total das quantidades encomendadas dos produtos, atingindo o produto B mais de 61%.

Tabela 20 - Percentagens Clientes/Quantidades em outubro 2014

outubro 2014		
Produto	**% Clientes**	**% Quantidade Encomendada**
Produto A	7,41	40,82
Produto B	11,34	61,68
Produto C	8,71	42,56

Por último, no gráfico comparativo (cf. Figura 18), pode-se concluir que:

- a percentagem de clientes *outliers* situa-se em ambos os períodos abaixo dos 10% (com exceção do produto B no ano de 2014);

- a percentagem das quantidades encomendas detetadas como *outliers*, embora tenha baixado em outuro de 2014 (com exceção do produto B) continua a ser bastante elevada, podendo originar a rutura de *stock*.

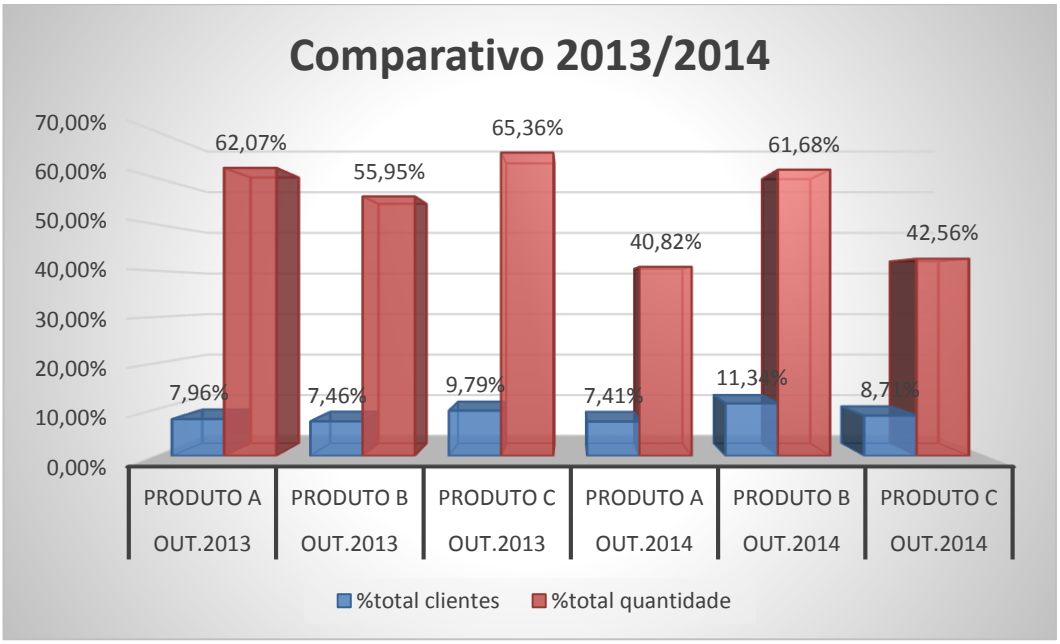


Figura 18 – Gráfico comparativo das percentagens das quant. encom. e clientes *outliers* (2013/14)

4.3.4. Principais resultados

Com base nos resultados obtidos usando o SPSS (versão 20), foi possível detetar para outubro de 2013 a existência, em todos os produtos analisados (A,B e C), de *outliers* superiores. Outro resultado de interesse, foi a verificação que o número de clientes *outliers* severos, é sempre maior que o número de *outliers* moderados na análise dos produtos (medicamentos). Neste período, relativamente aos clientes (farmácias) é importante salientar que, a soma das quantidades encomendadas pelos clientes detetados como *outliers*, foi sempre superior a 55% do total das quantidades encomendadas.

Já no período de outubro de 2014, ao contrário do anterior (outubro de 2013) o número de clientes *outliers* severos foi sempre inferior ao número de *outliers* moderados. A soma das quantidades encomendadas pelos clientes detetados como *outliers*, foi sempre superior a 40% do total das quantidades encomendadas.

Em suma, apesar do abaixamento verificado em outubro de 2014 na quantidade encomendada pelos clientes *outliers*, estas continuam a ser bastante elevadas com perigo de levar à rutura de *stock*.

4.4. Deteção e alerta de outliers na Empresa

A empresa de distribuição farmacêutica movimenta milhares de faturas por mês (aproximadamente 180.000 faturas), facto este que levou à necessidade de criação de mecanismos para, a detecção de *outliers* ser mais rápida e a avaliação dos *outliers* severos ser mais precisa.

Deste modo, no sentido de tornar o processo de classificação de *outliers* mais rápido foi criada uma *Materialized View* (cf. Apêndice 7) com os dados dos últimos 31 dias, que é atualizada aos fins-de-semana, antes do processo de deteção de *outliers* (cf. Apêndice 9) ser executado. Foi também criada uma tabela (cf. Apêndice 8) com os resultados do processo de classificação de deteção dos *outliers.* Esta tabela é usada para o processo de envio de correio eletrónico com a identificação dos produtos e clientes considerados *outliers*.

De seguida, com o intuito de obter uma avaliação mais precisa para os outliers severos, uma vez que, foram detetados imensos *outliers* severos, quer para clientes quer para quantidades encomendadas dos 3 produtos, criou-se uma regra que permitiu discriminar entre *outlier* “severo” e “muito severo”.

Para tal, em 1º lugar, começam por ser identificados os *outliers* (clientes e produtos) severos usando a regra de Tukey (método *Box-plot*). Seguidamente, é calculado para os outliers um valor designado por valor_outlier que é obtido da seguinte forma:

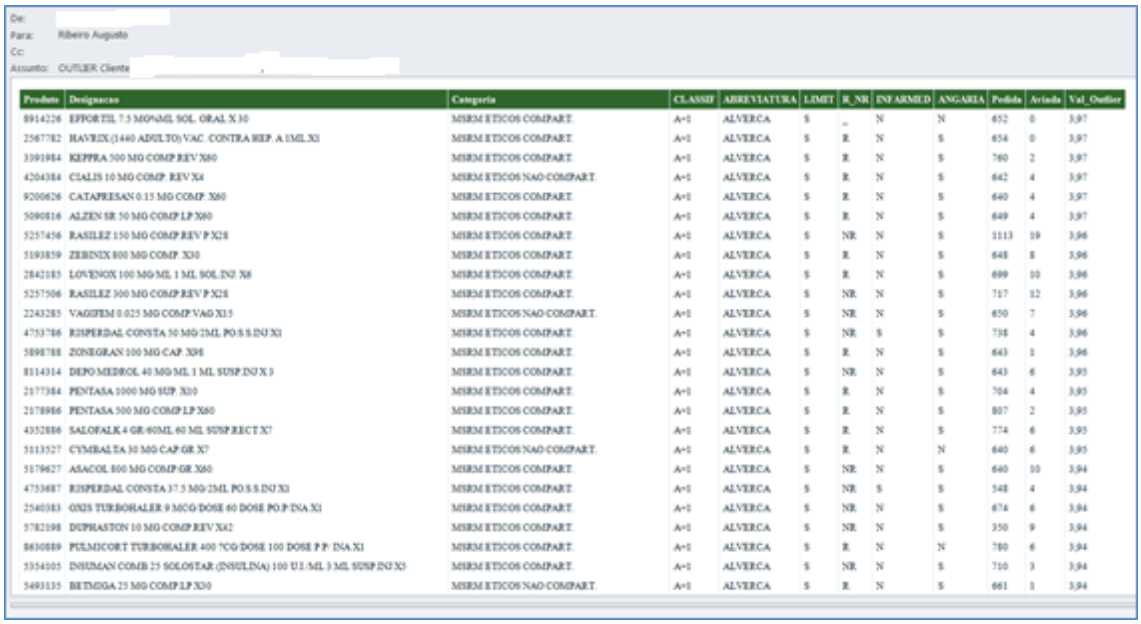
1) Se a quantidade pedida > quantidades pedidas por todos os clientes de igual classificação (3 vezes o IQR)
   Então soma ao valor_outlier +1 - (1 / numero clientes igual classificação)

2) Se a quantidade pedida > quantidades pedidas por todos os clientes do mesmo armazém (3 vezes o IQR)
   Então soma ao valor_outlier +1 - (1 / numero clientes mesmo armazém)

3) Se a quantidade pedida > consumo mensal do armazém
   Então soma ao valor_outlier +1 - (1 / numero clientes mesmo armazém)

4) Se a quantidade pedida > consumo mensal nacional (empresa)
   Então soma ao valor_outlier +1

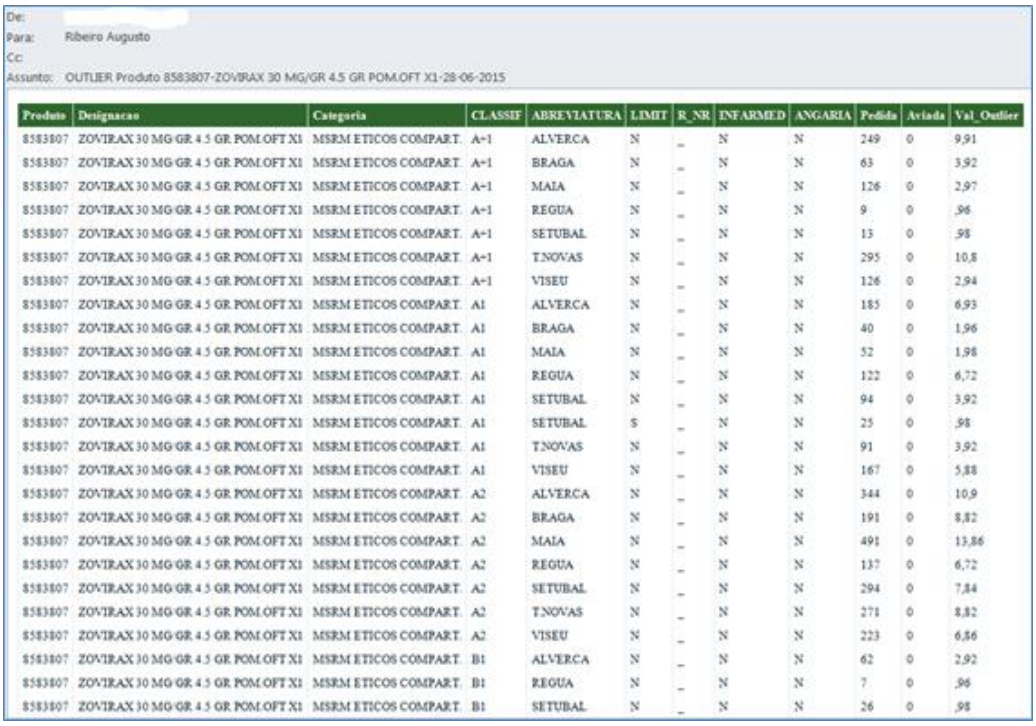
Esta regra obtida por experimentação prática atribui pesos (numéricos) a cada um dos *outliers* severos, o que permite classifica-los por ordem decrescente. Nas figuras 19 e 20, são apresentados exemplos dos *e-mails* produzidos.



De:
Para: Ribeiro Augusto
Cc:
Assunto: OUTLIER Cliente ,

Produto	Designacao	Categoria	CLASSIF	ABREVIATURA	LIMIT	R_NR	INFARMED	ANGARIA	Pedida	Avianda	Val_Outlier
8914226	EFFORTIL 7.5 MG/ML SOL. ORAL X 30	MSRM ETICOS COMPART.	A+1	ALVERCA	S	_	N	N	652	0	3,97
2567782	HAVRIX (1440 ADULTO) VAC. CONTRA HEP. A 1ML X1	MSRM ETICOS COMPART.	A+1	ALVERCA	S	R	N	S	654	0	3,97
3391984	KEPPRA 500 MG COMP REV X60	MSRM ETICOS COMPART.	A+1	ALVERCA	S	R	N	S	760	2	3,97
4204384	CIALIS 10 MG COMP. REV X4	MSRM ETICOS NAO COMPART.	A+1	ALVERCA	S	R	N	S	642	4	3,97
9200626	CATAPRESAN 0.15 MG COMP. X60	MSRM ETICOS COMPART.	A+1	ALVERCA	S	R	N	S	640	4	3,97
5090816	ALZEN SR 50 MG COMP LP X60	MSRM ETICOS COMPART.	A+1	ALVERCA	S	R	N	S	649	4	3,97
5257456	RASILEZ 150 MG COMP REV P X28	MSRM ETICOS COMPART.	A+1	ALVERCA	S	NR	N	S	1113	19	3,96
5193859	ZEBINIX 800 MG COMP. X30	MSRM ETICOS COMPART.	A+1	ALVERCA	S	R	N	S	648	8	3,96
2842183	LOVENOX 100 MG/ML 1 ML SOL INJ X6	MSRM ETICOS COMPART.	A+1	ALVERCA	S	R	N	S	699	10	3,96
5257506	RASILEZ 300 MG COMP REV P X28	MSRM ETICOS COMPART.	A+1	ALVERCA	S	NR	N	S	717	12	3,96
2243283	VAGIFEM 0.025 MG COMP VAG X15	MSRM ETICOS NAO COMPART.	A+1	ALVERCA	S	NR	N	S	650	7	3,96
4753786	RISPERDAL CONSTA 50 MG/2ML PO S S.INJ X1	MSRM ETICOS COMPART.	A+1	ALVERCA	S	NR	S	S	738	4	3,96
5098788	ZONEGRAN 100 MG CAP. X98	MSRM ETICOS COMPART.	A+1	ALVERCA	S	R	N	S	643	1	3,96
8114314	DEPO MEDROL 40 MG/ML 1 ML SUSP.INJ X 1	MSRM ETICOS COMPART.	A+1	ALVERCA	S	NR	N	S	643	6	3,95
2177384	PENTASA 1000 MG SUP. X20	MSRM ETICOS COMPART.	A+1	ALVERCA	S	R	N	S	704	4	3,95
2178986	PENTASA 500 MG COMP LP X60	MSRM ETICOS COMPART.	A+1	ALVERCA	S	R	N	S	807	2	3,95
4352886	SALOFALK 4 GR/60ML 60 ML SUSP.RECT X7	MSRM ETICOS COMPART.	A+1	ALVERCA	S	R	N	S	774	6	3,95
5113527	CYMBALTA 30 MG CAP GR X7	MSRM ETICOS NAO COMPART.	A+1	ALVERCA	S	R	N	N	640	6	3,95
5179627	ASACOL 800 MG COMP GR X60	MSRM ETICOS COMPART.	A+1	ALVERCA	S	NR	N	S	640	10	3,94
4753687	RISPERDAL CONSTA 37.5 MG/2ML PO S S.INJ X1	MSRM ETICOS COMPART.	A+1	ALVERCA	S	NR	S	S	548	4	3,94
2540383	OXIS TURBOHALER 9 MCG/DOSE 60 DOSE PO.P.INA X1	MSRM ETICOS COMPART.	A+1	ALVERCA	S	NR	N	S	674	6	3,94
5782198	DUPHASTON 10 MG COMP REV X42	MSRM ETICOS COMPART.	A+1	ALVERCA	S	NR	N	S	350	9	3,94
8630889	PULMICORT TURBOHALER 400 ?CG/DOSE 100 DOSE P.P. INA X1	MSRM ETICOS COMPART.	A+1	ALVERCA	S	R	N	N	780	6	3,94
5354105	INSUMAN COMB 25 SOLOSTAR (INSULINA) 100 UI/ML 3 ML SUSP.INJ X5	MSRM ETICOS COMPART.	A+1	ALVERCA	S	NR	N	S	710	3	3,94
5493133	BETMIGA 25 MG COMP LP X30	MSRM ETICOS NAO COMPART.	A+1	ALVERCA	S	R	N	S	661	1	3,94

Figura 19 – Exemplo de *e-mail* produzido com os produtos de um cliente detetado como *outlier*



De:
Para: Ribeiro Augusto
Cc:
Assunto: OUTLIER Produto 8583807-ZOVIRAX 30 MG/GR 4.5 GR POM.OFT X1-28-06-2015

Produto	Designacao	Categoria	CLASSIF	ABREVIATURA	LIMIT	R_NR	INFARMED	ANGARIA	Pedida	Aviada	Val_Outlier
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A+1	ALVERCA	N	_	N	N	249	0	9,91
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A+1	BRAGA	N	_	N	N	63	0	3,92
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A+1	MAIA	N	_	N	N	126	0	2,97
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A+1	REGUA	N	_	N	N	9	0	,96
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A+1	SETUBAL	N	_	N	N	13	0	,98
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A+1	T.NOVAS	N	_	N	N	295	0	10,8
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A+1	VISEU	N	_	N	N	126	0	2,94
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A1	ALVERCA	N	_	N	N	185	0	6,93
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A1	BRAGA	N	_	N	N	40	0	1,96
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A1	MAIA	N	_	N	N	52	0	1,98
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A1	REGUA	N	_	N	N	122	0	6,72
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A1	SETUBAL	N	_	N	N	94	0	3,92
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A1	SETUBAL	S	_	N	N	25	0	,98
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A1	T.NOVAS	N	_	N	N	91	0	3,92
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A1	VISEU	N	_	N	N	167	0	5,88
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A2	ALVERCA	N	_	N	N	344	0	10,9
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A2	BRAGA	N	_	N	N	191	0	8,82
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A2	MAIA	N	_	N	N	491	0	13,86
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A2	REGUA	N	_	N	N	137	0	6,72
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A2	SETUBAL	N	_	N	N	294	0	7,84
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A2	T.NOVAS	N	_	N	N	271	0	8,82
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	A2	VISEU	N	_	N	N	223	0	6,86
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	B1	ALVERCA	N	_	N	N	62	0	2,92
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	B1	REGUA	N	_	N	N	7	0	,96
8583807	ZOVIRAX 30 MG GR 4.5 GR POM.OFT X1	MSRM ETICOS COMPART.	B1	SETUBAL	N	_	N	N	26	0	,98

Figura 20 – Exemplo de *e-mail* produzido com os produtos detetados como *outliers*

5. Previsão de Vendas

5.1. Caracterização do problema

A determinação de encomendas de um produto (medicamento) para a empresa de distribuição farmacêutica é feita com base na análise do histórico das quantidades desse produto pedidas pelos clientes (farmácias) da empresa. Por outro lado, é sabido que existem períodos no ano em que a procura de determinados medicamentos é maior (por exemplo, os antigripais e antipiréticos no Inverno).

Pretende-se, então, identificar padrões regulares de observações históricas nas quantidades encomendadas pelas farmácias de um medicamento, com o objetivo da empresa de distribuição farmacêutica fazer previsões de vendas desses medicamentos para um período futuro.

Assim, no conjunto de dados a analisar (quantidades de um produto pedidas pelos clientes/farmácias), o atributo alvo (quantidade do produto a encomendar) é dependente do tempo, ou seja, está associado a uma sequência consecutiva de períodos, e o interesse é conhecer essa dependência.

O problema em causa, enquadra-se num caso de aplicação da categoria de data mining preditivo (pretende-se fazer uma previsão do valor futuro de um atributo de interesse) e do método de *data mining* de séries temporais descrito na Secção 3.5, uma vez que o atributo alvo (quantidade a encomendar de um medicamento por mês) é dependente do tempo, está associado a uma sequência consecutiva de períodos e se pretende prever o seu valor para um determinado período futuro (no caso concreto, o mês atual e os dois meses posteriores).

5.2. Volume de dados envolvidos

A empresa de distribuição farmacêutica movimenta cerca de 180.000 faturas por mês.

A tabela de Faturas incluí dados desde 2005 e tem 5,5GB de tamanho com 13.160.841 registos. A tabela das Linhas de Fatura possui 271.290.537 registos e 38,2 GB de tamanho, sendo adicionados por mês cerca de 3.700.000 novos registos (cf. Tabela 21).

Tabela 21 – Dados das tabelas de faturas da empresa, em 30 de Junho de 2014

	Cabeçalho Fatura	Linha Fatura
Nome tabela	arm_facturas_estab	arm_facturas_estab_produtos
Tamanho	5,5 GB.	38,2 GB.
Nº registos	13.160.841	271.290.537
Nº registos adicionados por mês	≈ 180.000	≈ 3.700.000

Foi criada uma nova tabela para guardar os valores das quantidades de produtos pedidas pelos clientes da empresa, com os dados agregados por produto/mês referentes a cada armazém.

REG_ID_PRODUTO	ANO_MES	ESTAB_ID	QT_PEDIDA
427795	201512	2	40881
427795	201511	2	51538
427795	201510	2	51538
427795	201509	2	55932
427795	201508	2	38908
427795	201507	2	57406
427795	201506	2	54386
427795	201505	2	47933

Figura 21 – Exemplo de registos da tabela com dados agregados por produto/mês

Assim, da análise da Figura 21, é possível verificar que o produto com código 427795 teve uma encomenda em dezembro de 2015 de 40881 unidades alocadas ao armazém da empresa com o código “2” e de 51538 unidades alocadas ao mesmo armazém em novembro de 2015.

Os dados desta tabela serviram de base ao processo de previsão de vendas de produtos efetuado e que, do ponto de vista da empresa farmacêutica, corresponderá às quantidades de produto a encomendar.

5.3. Aplicação do método de previsão de vendas utilizado

O método de Pegels amortecido (Taylor, 2003) e a medida de precisão SMAPE descritos na Secção 3.5 foram utilizados no cálculo da previsão.

Foi efetuada a previsão de vendas para 357 medicamentos comercializados pela empresa de distribuição farmacêutica. A seleção destes produtos (dos cerca de 20000 produtos comercializados pela empresa) foi feita tendo em conta a relevância das vendas dos mesmos em termos de negócio para a empresa.

A previsão de vendas foi feita para o mês atual à data da realização deste trabalho (janeiro de 2016) e para os dois meses posteriores (fevereiro e março de 2016).

A implementação deste método e do cálculo do erro associado foram feitos em SQL (cf. Apêndice 6), uma vez que a base de dados da empresa (produtos, vendas) é uma base de dados ORACLE versão 11.2.

Cálculo do erro associado e análise e intrepretação dos resultados obtidos

Conforme apresentado na Figura 22, os valores do método de Pegels amortecido α, β, $\emptyset$ gerados e o erro associado SMAPE são guardados na tabela; subsequentemente, é selecionada a combinação das constantes com o menor erro associado para cada produto/estabelecimento.

α	β	Ø	SMAPE
PROCESS_CONST	TREND_CONST	DAMPED_CONST	ERROR
0.1	0.9	0.9	1.039487615983068122725688296714487689 26
0.1	0.8	0.9	1.075087565568402913227338692197366235 47
0.1	0.7	0.9	1.148581425518697690200553038704730122 42
0.1	0.6	0.9	1.263081021314171157369271697604979123 38
0.2	0.5	0.9	1.276419513818803467945344304777201535 12
0.2	0.4	0.9	1.293067845199291080506455020098945860 9

Figura 22 – Apresentação do erro (SMAPE) ordenado por ordem ascendente e obtido pela combinação das 3 constantes do método

Depois de selecionada a combinação com o menor erro, esta é aplicada para o cálculo da previsão de vendas, sendo o resultado obtido guardado no campo “Quantidade_Prevista” da tabela (cf. Figura 23).

ANO_MES	ESTAB_ID	REG_ID_PRODUTO	QUANTIDADE_VENDIDA	QUANTIDADE_PREVISTA
201603	2	427795		49081
201602	2	427795		50074
201601	2	427795	51153	51191
201512	2	427795	40881	53372
201511	2	427795	51538	53502
201510	2	427795	51538	53443
201509	2	427795	55932	52668
201508	2	427795	38908	53975
201507	2	427795	57406	51841
201506	2	427795	54386	50230

Figura 23 – Exemplo da aplicação do método de previsão de vendas

Para uma visualização dos resultados da aplicação do método de previsão de vendas mais “user-friendly”, foi desenvolvida uma aplicação visual em ORACLE APEX versão 4.2 (cf. Figura 24).

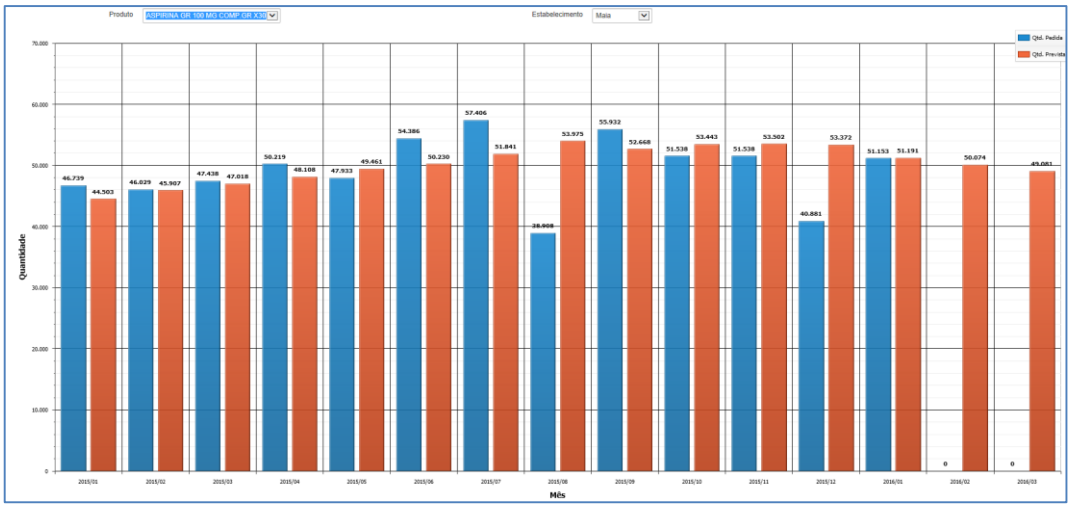


Figura 24 – Ecrã do CRM da empresa, com a representação gráfica das quantidades pedidas pelos clientes (a cor azul – Qtd. Pedida) de um produto selecionado e correspondente previsão de vendas (a cor vermelha – Qtd. Prevista)

A Figura 24 apresenta os resultados da aplicação do método de Pegels amortecido para, com base nos dados históricos das vendas de um produto, calcular as vendas previstas desse produto para o mês atual e os dois meses seguintes.

Assim, para o produto selecionado (ASPIRINA GR 100 MG COMP GR X30) e para o armazém de venda “Maia”, tendo em conta a análise do histórico das quantidades desse medicamento pedidas pelos clientes entre janeiro de 2014 e dezembro de 2015 (24 meses), as quantidades previstas de venda do medicamento para dezembro de 2015, o mês atual (janeiro de 2016) e para fevereiro e março de 2016 serão de 53372, 51191, 50074 e 49081 unidades respetivamente.

À data da realização deste trabalho, o processo utilizado pela empresa de distribuição farmacêutica para determinar a quantidade a encomendar de um dado produto para o mês seguinte era baseado no cálculo da média aritmética das quantidades vendidas desse produto nos últimos 12 meses (cf. Figura 25)

Para o caso do produto já referido (ASPIRINA GR 100 MG COMP GR X30) e para o memo armazém de venda, a tabela da Figura 25 apresenta os valores das quantidades previstas de vendas determinadas através do método de

Pegels amortecido e com o processo de cálculo da média de vendas utilizado pela empresa.

ANO MES	ESTAB ID	REG ID PRODUTO	QUANTIDADE VENDIDA	QUANTIDADE PREVISTA PEGELS	QUANTIDADE PREVISTA (processo utilizado empresa)
201603	2	427795		49.081	
201602	2	427795		50.074	
201601	2	427795	51.153	51.191	49.079
201512	2	427795	40.881	53.372	48.317
201511	2	427795	51.538	53.502	47.398
201510	2	427795	51.538	53.443	46.786
201509	2	427795	55.932	52.668	45.417
201508	2	427795	38.908	53.975	44.671
201507	2	427795	57.406	51.841	43.304
201506	2	427795	54.386	50.230	41.901
201505	2	427795	47.933	49.461	41.235
201504	2	427795	50.219	48.108	40.171
201503	2	427795	47.438	47.018	39.028
201502	2	427795	46.029	45.907	37.447
201501	2	427795	46.739	44.503	36.011

Figura 25 – Comparação da aplicação do método de previsão de vendas com o processo utilizado pela empresa

Da análise da tabela, constata-se que os valores obtidos pelo método de Pegels estão na generalidade mais próximos dos valores reais do que os obtidos com o processo atualmente em vigor na empresa. Para além disso, é importante referir que os valores obtidos pelo método de Pegels satifazem na maioria dos casos a procura, o que não acontece no cálculo feito pelo outro processo.

Conclui-se, assim, que o desempenho do método de Pegels amortecido foi favorável na previsão de vendas de produtos ao nível individual, obtendo-se resultados mais aproximados aos reais e mais confiáveis, em comparação com o processo anteriormente utilizado.

6. Conclusão

6.1. Resposta às questões de investigação

Revisitando as questões de investigação colocadas no início deste trabalho e apresentadas no Capítulo 1, é possível chegar às seguintes conclusões:

(Q1) Será possível usando a deteção de *outliers*, controlar o *stock* mínimo de medicamentos, por parte das empresas de distribuição farmacêutica, de forma a impedir ruturas no abastecimento normal do mercado (farmácias)?

Sim. A partir do estudo desenvolvido e aplicação prática à empresa de distribuição farmacêutica, conclui-se que usando o método de Box-plot para a detecção de *outliers* (clientes e produtos) e a regra criada (na Secção 4.4) é possível identificar de uma forma rápida e precisa quais são esses *outliers* com vista a tomar medidas preventivas. Isto é, aplicando este processo à empresa é possível sempre que se detecta *outliers*, enviar um alerta (*e-mail*) permitindo que o *stock* desse produto seja rateado de forma a impedir a rutura do mesmo.

Obteve-se assim um processo simples, rápido e económico de detecção de *outliers* para controle do *stock* de medicamentos na empresa.

(Q2) Será possível utilizando um método de *data mining* para previsão de vendas, prever, de uma forma fiável, os valores a encomendar de cada produto, pelas empresas de distribuição farmacêutica, de forma a obter um bom controlo dos níveis de *stock* e respetivos custos?

Do estudo e aplicação efetuados na empresa de distribuição farmacêutica retratada neste trabalho, é possível concluir que o desempenho do método de séries temporais Pegels amortecido foi favorável na previsão de vendas de produtos ao nível individual, obtendo-se resultados mais aproximados aos reais e mais confiáveis, em comparação com o processo anteriormente utilizado pela empresa. Verificou-se que utilizando os valores obtidos pelo método de Pegels amortecido para a previsão de vendas, a empresa poderia melhor determinar os

valores a encomendar dos produtos em causa, garantido na maioria dos casos níveis de *stock* que efetivamente pudessem satisfazer os valores de procura real, sem considerar níveis de stocks excessivos e com forte impacto nos custos.

6.2. Trabalho Futuro

Concluído o tempo disponível para o desenvolvimento deste estudo, foram identificados alguns pontos que, poderiam ser melhorados.

Relativamente à detecção de *outliers* (clientes e produtos) seria importante realizar o estudo para os 3256 Medicamentos Sujeitos a Receita Médica (MSRM) comercializados pela empresa.

Outro ponto de interesse a desenvolver futuramente, seria a realização de deteção de *outliers* diariamente, de modo a detetar *outliers* atempadamente, antes da rutura de *stock*.

Adicionalmente, apesar do método Box-plot ter sido considerado o mais adequado para o estudo presente, poderia ser feita a comparação da detecção de outliers diária com os outros métodos para os 3256 MSRM.

Quanto ao método de previsão de vendas utilizado e apesar dos resultados obtidos terem sido satisfatórios, considera-se que ainda há espaço para introduzir melhorias na técnica de previsão. Por um lado, seria interessante verificar se, com a utilização de um horizonte de previsão mais amplo do que o que foi utilizado neste trabalho, os resultados obtidos poderiam ter uma precisão que pudesse ser considerada aceitável. A inclusão de um horizonte de previsão mais alargado (por exemplo, 12 meses) na previsão de vendas, poderá ter interesse para a empresa de distribuição farmacêutica ter uma maior margem de manobra na negociação de preços com os fornecedores.

Por outro lado, aspetos como dados da realização de ações publicitárias por parte dos fornecedores e dados de promoções efetuadas pela concorrência (outras empresas de distribuição farmacêutica) poderão ter impacto nas vendas reais dos produtos. Assim, poderia ser equacionada a possibilidade de estes

aspetos serem considerados como variáveis adicionais na determinação da previsão de vendas.

Adicionalmente, a comparação do desempenho do método de Pegels amortecido com outros métodos de previsão de séries temporais poderia ser objeto de estudo.

Por último, a possibilidade de utilização, por parte da empresa, do *software* comercial X-13 ARIMA-SEATS (que se integra com o package estatístico R Oracle) (Kowarik, Meraner, Templ, & Schopfhauser, 2014) para previsão de vendas poderia ainda constituir um trabalho de exploração a efetuar, no sentido de verificar se o o método de previsão implementado pelo software produz resultados mais precisos.

Referências Bibliográficas

Agrawal, R., Imieliński, T., & Swami, A. (1993). Mining association rules between sets of items in large databases. In *ACM SIGMOD international conference on Management of data* (Vol. 22, pp. 207–216).

Agrawal, R., & Srikant, R. (1994). Fast algorithms for mining association rules. *Proceeding VLDB '94 Proceedings of the 20th International Conference on Very Large Data Bases*, *1215*, 487–499.

Arnaldo-jr. (2015). Exemplo prático de uma rede neural. Retrieved from http://docslide.com.br/documents/inteligencia-artificial-redes-neurais-exemplo-pratico.html

Barnett, V., & Lewis, T. (1994). *Outliers in Statistical Data*. John Wiley & Sons.

Berry, M. J. A., & Linoff, G. S. (2000). *Mastering Data Mining: The Art and Science of Customer Relationship Management*. Wiley.

Box, G. E. P., & Jenkins, G. M. (1976). *Time Series Analysis: Forecasting and Control* (2nd ed.). San Francisco: Holden-Day.

Brys, G., Hubert, M., & Rousseeuw, P. J. (2005). A robustification of independent component analysis. *Journal of Chemometrics*.

Chen, C., & Lon-Mu, L. (1993). Forecasting time series with outliers. *Journal of Forecasting*, *12*, 13–35.

Cristian Mihaescu. (2010). Hierarchical Clustering. Craiova: Universitatea din Craiova.

Dean, R. B., & Dixon, W. J. (1951). Simplified Statistics for Small Numbers of Observations. *Analytical Chemistry*, *23*(May), 636–638.

Doganis, P., Alexandridis, A., Patrinos, P., & Sarimveis, H. (2006). Time series sales forecasting for short shelf-life food products based on artificial neural networks and evolutionary computing. *Journal of Food Engineering*, *75*, 196–204.

Duncan, G., Gorr, W., & Szczypula, J. (1998). *Forecasting analogous time series*. Pittsburgh.

Gardner, E.S., J., & McKenzie, E. (1985). Forecasting trends in time series. *Management Science*, *31*, 1237–1246.

Grubbs, F. E. (1969). Procedures for Detecting Outlying Observations in Samples. In *Technometrics* (Vol. 11, pp. 1–21).

Gupta, A., Maranas, C. D., & McDonald, C. M. (2000). Mid-term supply chain planning under demand uncertainty: customer demand satisfaction and inventory management. *Computers & Chemical Engineering*, *24*, 2613–2621.

Han, J., Kamber, M., & Pei, J. (2011). *Data Mining: Concepts and Techniques*. (Elsevier, Ed.). Elsevier.

Hawkins, D. (1980). *Identification of Outliers*. London: Chapman and Hall.

Hoaglin, D. C., & Iglewicz, B. (1987). Fine Tuning Some Resistant Rules for Outlier Labeling. *Journal of American Statistical Association*, *82*, 1147–1149.

Hoaglin, D. C., Iglewicz, B., & Tukey, J. W. (1986). Performance of Some Resistant Rules for Outlier Labeling. *Journal of American Statistical Association*, *81*, 991–999.

Hodge, V., & Austin, J. (2004). A survey of outlier detection methodologies. *Artificial Intelligence Review*, 1–43.

Iglewicz, B., & Hoaglin, D. (1993). *How to Detect and Handle Outliers*. ASQC Quality Press.

Infarmed, O. (2012). Rupturas de Stock de Medicamentos. Retrieved from https://www.infarmed.pt/portal/page/portal/INFARMED/PUBLICACOES/TEMATICOS/SAIBA_MAIS_SOBRE/SAIBA_MAIS_ARQUIVO/43_Rupturas_Stock.pdf

John, G. H. (1995). Robust Decision Trees: Removing Outliers from Databases. In *Proceedings of the First International Conference on Knowledge Discovery and Data Mining* (pp. 174–179). Menlo Park, CA: Press, AAAI.

Jornal Público. (2013). Farmácias que exportem medicamentos que fazem falta em Portugal vão pagar coima quatro vezes superior. *Público*. Retrieved from https://www.publico.pt/sociedade/noticia/farmacias-que-exportem-medicamentos-que-fazem-falta-em-portugal-vao-pagar-coima-quatro-vezes-superior-1614144

Kanji, G. K. (1993). *100 STATISTICAL TESTS* (3rd ed.). London: SAGE Publication Ltd.

Kowarik, A., Meraner, A., Templ, M., & Schopfhauser, D. (2014). Seasonal Adjustment with the R Packages x12 and x12GUI. *Journal Of Statistical Software*, *62*(2), 1–21.

Kriegel, H. P., Kröger, P., & Zimek, A. (2009). Outlier detection techniques. In *Tutorial at the 13th Pacific-Asia …* (p. 6).

Leys, C., Ley, C., Klein, O., Bernard, P., & Licata, L. (2013). Detecting outliers : Do not use standard deviation around the mean , use absolute deviation around the median. *Journal of Experimental Social Psychology*, (outliers), 4–6.

Makridakis, S., Wheelwright, C., S., & Hyndman, R. J. (1998). *Forecasting : methods and aplications*. John Wiley & Sons, cop.

Oliveira, E. C. De. (2008). Comparação das diferentes técnicas para a exclusão de “outliers.” *Metrologia*.

Osuna, E. E., Freund, R., & Girosi, F. (1999). Support Vector Machines : Training and Applications, *9217041*(1602).

Pereira, C. M. S., & Pires, A. M. (2002). Detection of Outliers in Multivariate Data: A Method Based on Clustering and Robust Estimators. In W. Hordle & B. Ronz (Eds.), . Berlin: Proceedings in Computational Statistics COMSTAT.

Rice, U. of, Houston, U. of, & Tufts, U. of. (2010). Exemplo de utilização do método Bayesiano no despite de uma doença. Retrieved from http://onlinestatbook.com/2/probability/bayes_demo.html

Santos, M., & Ramos, I. (2009). *Business Intelligence: Tecnologias da Informação na Gestão de Conhecimento*. *FCA - Editora de Informática*. FCA.

Shiffler, R. E. (1988). Maximum Z Scores and Outliers. *The American Statistician*, *42*, 79–80.

STAT4U. (2008). Data Warehouse. Retrieved from http://www.datawarehouse4u.info/

Tolvi, J. (n.d.). *Outliers in time series : A review **.

Tukey, J. W. (1977). *Exploratory Data Analysis*. Reading, MA: Addison-Wesley.

Vanderviere, E., & Huber, M. (2004). AN ADJUSTED BOXPLOT FOR SKEWED DISTRIBUTIONS Ellen Vanderviere and Mia Huber. *COMPSTAT 2004: Proceedings in Computational Statistics. 2004*, 1933–1940.

Vercellis C. (2009). *Business intelligence data mining and organization for decision making*. *Wiley* (2nd ed.). Chichester: John Wiley & Sons.

wikipedia. (2016). Regressão logística. Retrieved from https://en.wikipedia.org/wiki/Logistic_regression

Winters, P. R. (1960). Forecasting Sales by Exponentially Weighted Moving Averages. *Management Science*, *6*, 324–342.

Zadeh, N. K., Sepehri, M. M., & Farvaresh, H. (2014). Intelligent Sales Prediction for Pharmaceutical Distribution Companies : A Data Mining Based Approach, *2014*.

Anexos

Todas as tabelas e gráficos constantes destes anexos, foram construídos usando o *software* SPSS (versão 20).

Anexo 1. PRODUTO A – outubro 2013

Tabela 22 – Quant. encomendada por classificação do cliente - produto A (out. 2013)

Extreme Values					
CLASSIF				Case Number	Value
QT_ENCOMENDA	C11	Highest	1	1	102
			2	2	61
			3	3	39
			4	4	31
			5	5	30
		Lowest	1	33	1
			2	32	1
			3	31	2
			4	30	2
			5	29	2[a]
	C12	Highest	1	34	31
			2	35	27
			3	36	26
			4	37	22
			5	38	20
		Lowest	1	123	1
			2	122	1
			3	121	1
			4	120	1
			5	119	1[b]
	C13	Highest	1	124	98
			2	125	44
			3	126	39
			4	127	24
			5	128	16
		Lowest	1	169	1
			2	168	1
			3	167	1
			4	166	1
			5	165	1[b]
	C21	Highest	1	170	68

			2	171	39
			3	172	33
			4	173	20
			5	174	16
		Lowest	1	204	1
			2	203	1
			3	202	1
			4	201	1
			5	200	1[b]
	C22	Highest	1	205	51
			2	206	49
			3	207	41
			4	208	32
			5	209	30
		Lowest	1	280	1
			2	279	1
			3	278	1
			4	277	1
			5	276	1[b]
	C23	Highest	1	281	31
			2	282	29
			3	283	20
			4	284	14
			5	285	11
		Lowest	1	302	1
			2	301	1
			3	300	1
			4	299	1
			5	298	1[b]
	C31	Highest	1	303	1697
			2	304	1360
			3	305	990
			4	306	880
			5	307	221
		Lowest	1	378	2
			2	377	3
			3	376	5
			4	375	5
			5	374	6[c]
	C32	Highest	1	379	490
			2	380	109
			3	381	84
			4	382	62
			5	383	52

		Lowest	1	510	2
			2	509	2
			3	508	2
			4	507	2
			5	506	2
	C33	Highest	1	511	1550
			2	512	36
			3	513	26
			4	514	23
			5	515	23
		Lowest	1	671	1
			2	670	1
			3	669	2
			4	668	2
			5	667	2[a]
	C41	Highest	1	672	65
			2	673	59
			3	674	45
			4	675	29
			5	676	26
		Lowest	1	718	2
			2	717	2
			3	716	2
			4	715	2
			5	714	2[a]
	C42	Highest	1	719	261
			2	720	243
			3	721	175
			4	722	90
			5	723	87
		Lowest	1	931	1
			2	930	1
			3	929	1
			4	928	1
			5	927	1[b]
	C43	Highest	1	932	734
			2	933	395
			3	934	65
			4	935	29
			5	936	19
		Lowest	1	993	1
			2	992	1
			3	991	1
			4	990	1

			5	989	1

a. Only a partial list of cases with the value 2 are shown in the table of lower extremes.

b. Only a partial list of cases with the value 1 are shown in the table of lower extremes.

c. Only a partial list of cases with the value 6 are shown in the table of lower extremes.

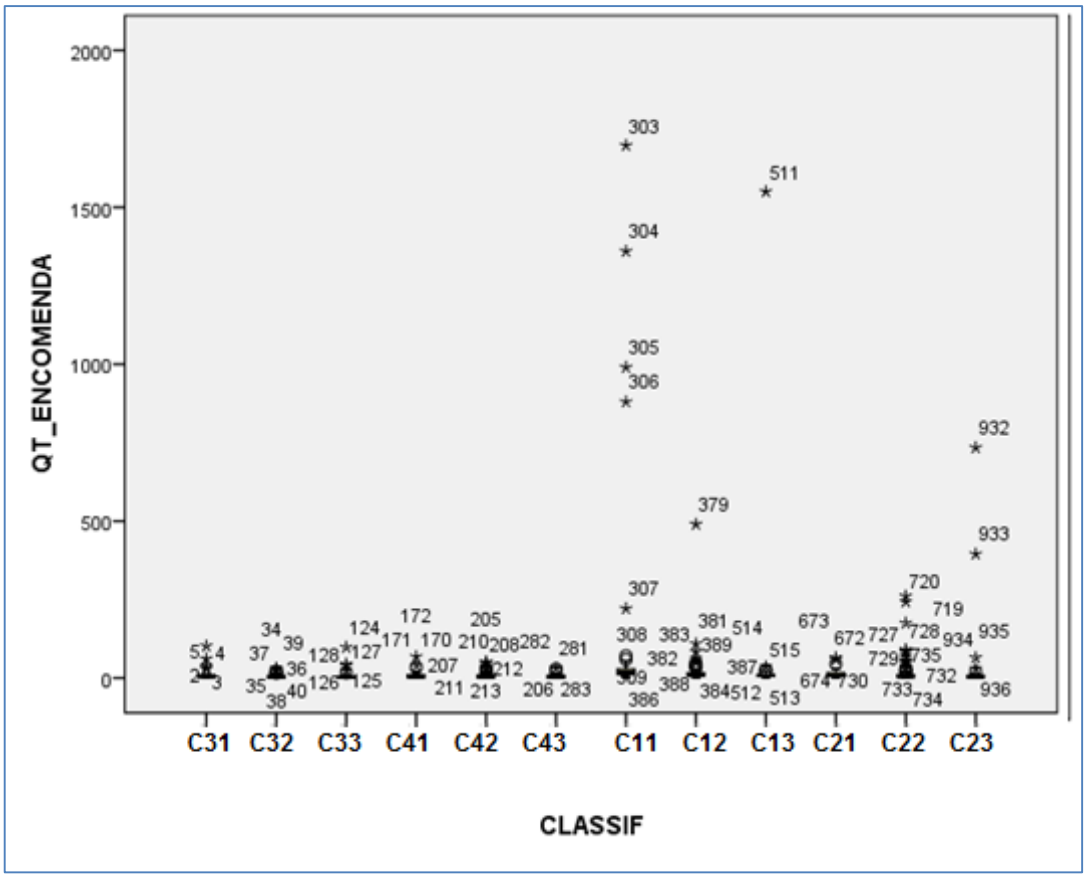


Figura 26 – Diagrama caixa de bigodes da quant. encomendada do produto A (outubro 2013) por classificação do cliente

Anexo 2. PRODUTO A – outubro 2014

Tabela 23 – Quant. encomendada por classificação do cliente - produto A (out. 2014)

Extreme Values					
CLASSIF				Case Number	Value
QT_ENCOMENDA	C11	Highest	1	22	72
			2	21	43
			3	20	27
			4	19	24
			5	17	22[a]
		Lowest	1	2	2
			2	1	2
			3	3	3
			4	5	4
			5	4	4
	C12	Highest	1	98	530
			2	97	238
			3	96	198
			4	95	68
			5	93	34[b]
		Lowest	1	37	1
			2	36	1
			3	35	1
			4	34	1
			5	33	1[c]
	C13	Highest	1	133	89
			2	132	46
			3	131	41
			4	130	28
			5	129	24
		Lowest	1	103	1
			2	102	1
			3	101	1
			4	100	1
			5	99	1
	C21	Highest	1	152	18
			2	151	16
			3	150	15
			4	148	14
			5	149	14
		Lowest	1	137	1
			2	136	1
			3	135	1
			4	134	1
			5	138	2
	C22	Highest	1	214	65
			2	212	38
			3	213	38
			4	211	31
			5	210	22
		Lowest	1	164	1
			2	163	1
			3	162	1
			4	161	1
			5	160	1[c]
	C23	Highest	1	238	21
			2	237	15
			3	236	11
			4	235	9
			5	232	8[d]
		Lowest	1	220	1

			2	219	1
			3	218	1
			4	217	1
			5	216	1[c]
	C31	Highest	1	308	1297
			2	307	786
			3	306	193
			4	305	190
			5	304	163
		Lowest	1	239	1
			2	242	2
			3	241	2
			4	240	2
			5	243	3
	C32	Highest	1	415	794
			2	414	199
			3	413	116
			4	412	103
			5	411	100
		Lowest	1	309	1
			2	312	2
			3	311	2
			4	310	2
			5	315	3[e]
	C33	Highest	1	610	159
			2	609	106
			3	608	80
			4	607	78
			5	606	76
		Lowest	1	419	1
			2	418	1
			3	417	1
			4	416	1
			5	433	2[f]
	C41	Highest	1	643	276
			2	642	80
			3	641	70
			4	640	69
			5	639	48
		Lowest	1	613	2
			2	612	2
			3	611	2
			4	615	4
			5	614	4
	C42	Highest	1	802	970
			2	801	441
			3	800	221
			4	799	143
			5	798	133
		Lowest	1	652	1
			2	651	1
			3	650	1
			4	649	1
			5	648	1[c]
	C43	Highest	1	864	125
			2	863	72
			3	862	71
			4	861	66
			5	859	50[g]
		Lowest	1	807	1
			2	806	1
			3	805	1
			4	804	1
			5	803	1

a. Only a partial list of cases with the value 22 are shown in the table of upper extremes.

b. Only a partial list of cases with the value 34 are shown in the table of upper extremes.

c. Only a partial list of cases with the value 1 are shown in the table of lower extremes.
d. Only a partial list of cases with the value 8 are shown in the table of upper extremes.
e. Only a partial list of cases with the value 3 are shown in the table of lower extremes.
f. Only a partial list of cases with the value 2 are shown in the table of lower extremes.
g. Only a partial list of cases with the value 50 are shown in the table of upper extremes.

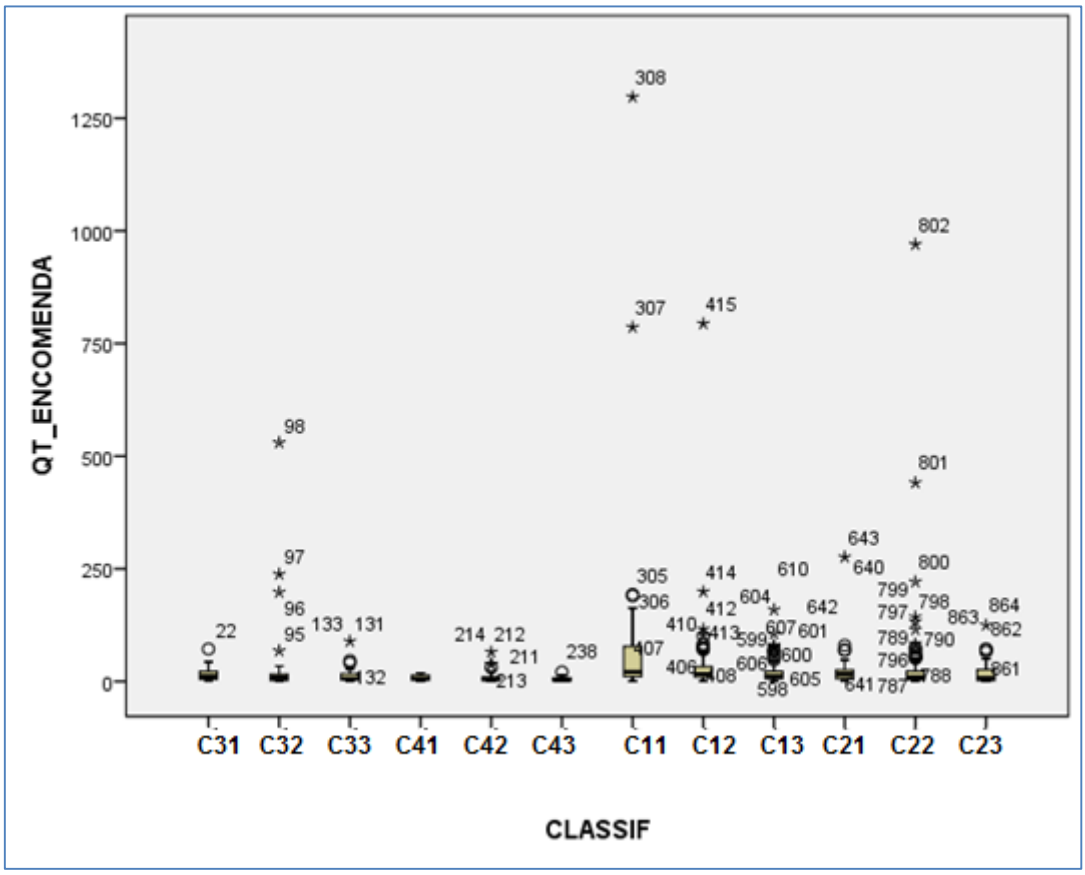


Figura 27 – Diagrama caixa de bigodes da quant. encomendada do produto A (outubro 2014) por classificação do cliente

Anexo 3. PRODUTO B – outubro 2013

Tabela 24 – Quant. encomendada por classificação do cliente - produto B (out. 2013)

Extreme Values					
CLASSIF				Case Number	Value
QT_ENCOMENDA	C11	Highest	1	1	46
			2	2	42
			3	3	35
			4	4	34
			5	5	24
		Lowest	1	25	1
			2	24	1
			3	23	1
			4	22	1
			5	21	1
	C12	Highest	1	26	111
			2	27	60
			3	28	59
			4	29	51
			5	30	45[a]
		Lowest	1	103	1
			2	102	1
			3	101	1
			4	100	1
			5	99	1[b]
	C13	Highest	1	104	36
			2	105	33
			3	106	24
			4	107	21
			5	108	20
		Lowest	1	140	1
			2	139	1
			3	138	1
			4	137	2
			5	136	2[c]
	C21	Highest	1	141	20
			2	142	20
			3	143	15
			4	144	11
			5	145	10[d]
		Lowest	1	163	1
			2	162	1
			3	161	2

			4	160	2
			5	159	2[c]
	C22	Highest	1	164	63
			2	165	24
			3	166	18
			4	167	12
			5	168	12
		Lowest	1	214	1
			2	213	1
			3	212	1
			4	211	1
			5	210	1[b]
	C23	Highest	1	215	23
			2	216	21
			3	217	17
			4	218	17
			5	219	12[e]
		Lowest	1	230	1
			2	229	1
			3	228	2
			4	227	2
			5	226	2
	C31	Highest	1	231	1310
			2	232	1000
			3	233	931
			4	234	266
			5	235	178
		Lowest	1	303	2
			2	302	2
			3	301	2
			4	300	2
			5	299	2[c]
	C32	Highest	1	304	122
			2	305	112
			3	306	105
			4	307	104
			5	308	86
		Lowest	1	421	1
			2	420	1
			3	419	1
			4	418	1
			5	417	1[b]
	C33	Highest	1	422	1571
			2	423	270

			3	424	122
			4	425	111
			5	426	86
		Lowest	1	573	1
			2	572	1
			3	571	1
			4	570	1
			5	569	1[b]
	C41	Highest	1	574	192
			2	575	94
			3	576	58
			4	577	55
			5	578	54
		Lowest	1	608	1
			2	607	1
			3	606	1
			4	605	1
			5	604	2[c]
	C42	Highest	1	609	200
			2	610	80
			3	611	80
			4	612	75
			5	613	61
		Lowest	1	777	1
			2	776	1
			3	775	1
			4	774	1
			5	773	1[b]
	C43	Highest	1	778	730
			2	779	458
			3	780	52
			4	781	40
			5	782	40
		Lowest	1	832	1
			2	831	1
			3	830	1
			4	829	1
			5	828	1

a. Only a partial list of cases with the value 45 are shown in the table of upper extremes.

b. Only a partial list of cases with the value 1 are shown in the table of lower extremes.

c. Only a partial list of cases with the value 2 are shown in the table of lower extremes.

d. Only a partial list of cases with the value 10 are shown in the table of upper extremes.

e. Only a partial list of cases with the value 12 are shown in the table of upper extremes.

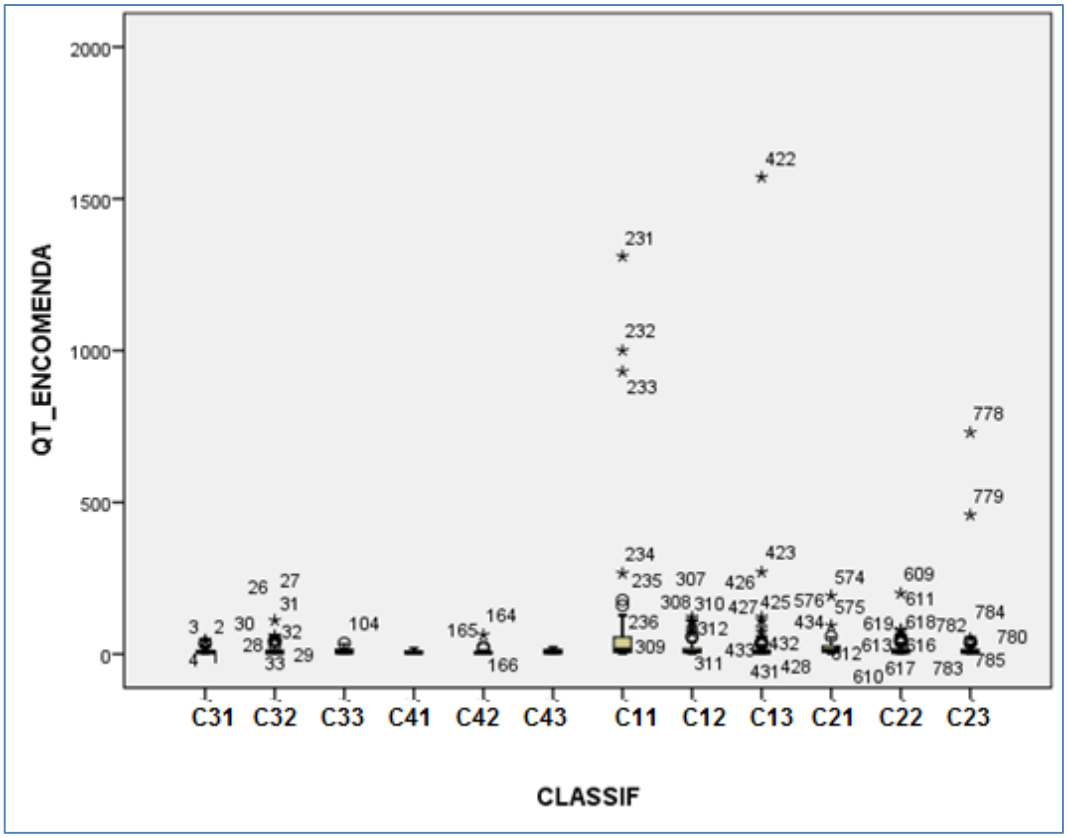


Figura 28 – Diagrama caixa de bigodes da quant. encomendada do produto B (outubro 2013) por classificação do cliente

Anexo 4. PRODUTO B – outubro 2014

Tabela 25 – Quant. encomendada por classificação do cliente - produto B (out. 2014)

Extreme Values					
CLASSIF				Case Number	Value
QT_ENCOMENDA	C11	Highest	1	9	64
			2	18	62
			3	1	53
			4	15	35
			5	12	30
		Lowest	1	14	4
			2	16	5
			3	17	6
			4	6	6
			5	7	8[a]
	C12	Highest	1	25	610
			2	31	413
			3	64	162
			4	78	108
			5	69	91
		Lowest	1	71	1
			2	58	1
			3	55	1
			4	53	1
			5	45	1[b]
	C13	Highest	1	104	132
			2	122	90
			3	115	40
			4	137	38
			5	120	36
		Lowest	1	131	1
			2	121	1
			3	135	2
			4	134	2
			5	126	2[c]
	C21	Highest	1	157	26
			2	149	24
			3	152	15
			4	155	15
			5	164	15
		Lowest	1	163	1
			2	154	1
			3	151	1
			4	146	1
			5	142	1
	C22	Highest	1	173	30
			2	180	30
			3	179	27
			4	210	25
			5	202	20[d]
		Lowest	1	218	1
			2	217	1
			3	213	1
			4	211	1
			5	198	1[b]
	C23	Highest	1	245	75
			2	237	43
			3	252	30

			4	251	17
			5	239	16
		Lowest	1	250	1
			2	238	1
			3	229	1
			4	247	2
			5	236	2[c]
	C31	Highest	1	298	1230
			2	266	697
			3	259	188
			4	283	172
			5	260	162
		Lowest	1	273	1
			2	318	2
			3	299	4
			4	288	4
			5	285	4[e]
	C32	Highest	1	410	690
			2	377	239
			3	413	207
			4	404	103
			5	328	96
		Lowest	1	396	1
			2	399	2
			3	397	2
			4	389	2
			5	378	2[c]
	C33	Highest	1	432	880
			2	516	354
			3	538	116
			4	473	114
			5	499	103
		Lowest	1	600	1
			2	599	1
			3	596	1
			4	593	1
			5	576	1[b]
	C41	Highest	1	624	113
			2	612	78
			3	628	68
			4	633	59
			5	610	51
		Lowest	1	602	2
			2	621	3
			3	630	4
			4	623	4
			5	632	5[f]
	C42	Highest	1	737	999
			2	683	181
			3	735	171
			4	751	144
			5	639	133
		Lowest	1	772	1
			2	754	1
			3	744	1
			4	710	1
			5	681	1[b]
	C43	Highest	1	821	74
			2	806	70
			3	782	69
			4	783	61
			5	807	40
		Lowest	1	837	1

			2	836	1
			3	834	1
			4	828	1
			5	825	1[b]

a. Only a partial list of cases with the value 8 are shown in the table of lower extremes.
b. Only a partial list of cases with the value 1 are shown in the table of lower extremes.
c. Only a partial list of cases with the value 2 are shown in the table of lower extremes.
d. Only a partial list of cases with the value 20 are shown in the table of upper extremes.
e. Only a partial list of cases with the value 4 are shown in the table of lower extremes.
f. Only a partial list of cases with the value 5 are shown in the table of lower extremes.

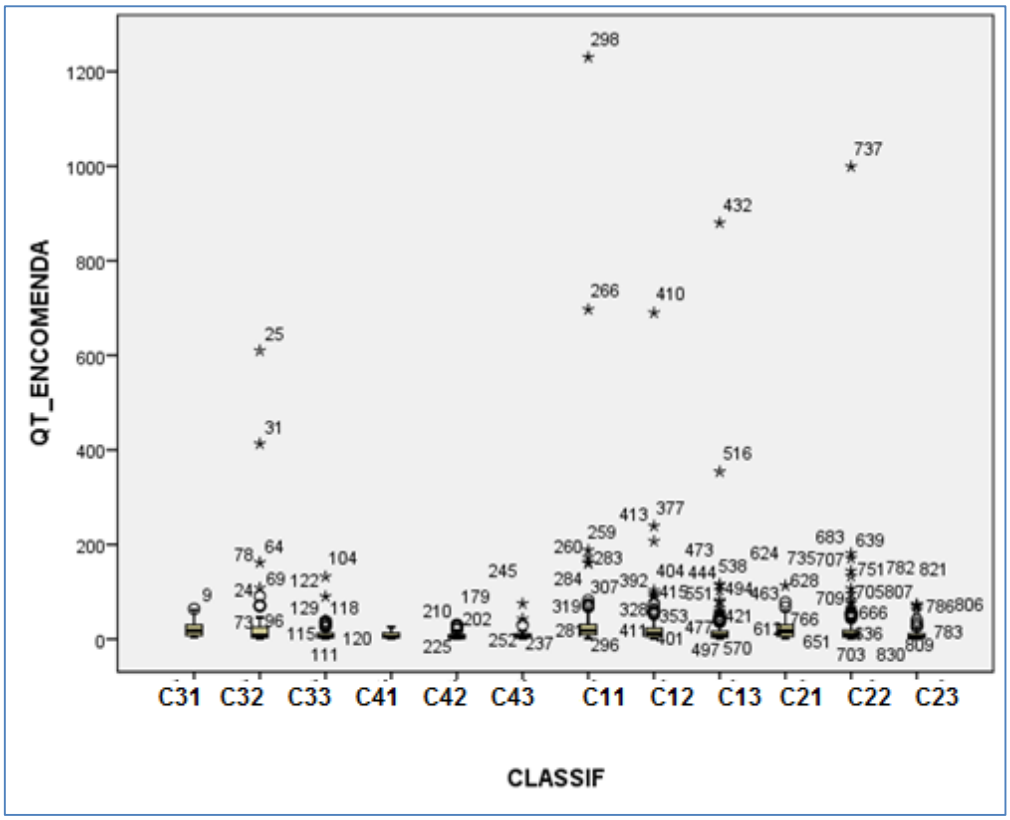


Figura 29 – Diagrama caixa de bigodes da quant. encomendada do produto B (outubro 2014) por classificação do cliente

Anexo 5. PRODUTO C – outubro 2013

Tabela 26 – Quant. encomendada por classificação do cliente - produto C (out. 2013)

Extreme Values					
CLASSIF				Case Number	Value
QT_ENCOMENDA	C11	Highest	1	1	38
			2	2	34
			3	3	20
			4	4	19
			5	5	17
		Lowest	1	27	1
			2	26	2
			3	25	2
			4	24	2
			5	23	2[a]
	C12	Highest	1	28	80
			2	29	65
			3	30	54
			4	31	47
			5	32	47
		Lowest	1	110	1
			2	109	1
			3	108	1
			4	107	1
			5	106	1[b]
	C13	Highest	1	111	56
			2	112	49
			3	113	28
			4	114	24
			5	115	21
		Lowest	1	148	1
			2	147	1
			3	146	1
			4	145	2
			5	144	2[a]
	C21	Highest	1	149	41
			2	150	11
			3	151	8
			4	152	6
			5	153	6
		Lowest	1	167	1
			2	166	1
			3	165	1

			4	164	2
			5	163	2[a]
	C22	Highest	1	168	60
			2	169	27
			3	170	24
			4	171	22
			5	172	21
		Lowest	1	215	1
			2	214	1
			3	213	1
			4	212	1
			5	211	1[b]
	C23	Highest	1	216	35
			2	217	27
			3	218	16
			4	219	10
			5	220	8
		Lowest	1	237	1
			2	236	1
			3	235	1
			4	234	1
			5	233	1[b]
	C31	Highest	1	238	1360
			2	239	990
			3	240	880
			4	241	260
			5	242	213
		Lowest	1	310	1
			2	309	1
			3	308	1
			4	307	2
			5	306	2[a]
	C32	Highest	1	311	490
			2	312	203
			3	313	182
			4	314	76
			5	315	55
		Lowest	1	431	1
			2	430	2
			3	429	2
			4	428	2
			5	427	2[a]
	C33	Highest	1	432	1575
			2	433	179

			3	434	103
			4	435	71
			5	436	60
		Lowest	1	581	1
			2	580	1
			3	579	1
			4	578	1
			5	577	1[b]
	C41	Highest	1	582	129
			2	583	83
			3	584	49
			4	585	48
			5	586	43
		Lowest	1	626	1
			2	625	1
			3	624	1
			4	623	2
			5	622	2
	C42	Highest	1	627	275
			2	628	198
			3	629	175
			4	630	165
			5	631	128
		Lowest	1	822	1
			2	821	1
			3	820	1
			4	819	1
			5	818	1[b]
	C43	Highest	1	823	730
			2	824	394
			3	825	68
			4	826	68
			5	827	29
		Lowest	1	888	1
			2	887	1
			3	886	1
			4	885	1
			5	884	1[b]

a. Only a partial list of cases with the value 2 are shown in the table of lower extremes.

b. Only a partial list of cases with the value 1 are shown in the table of lower extremes.

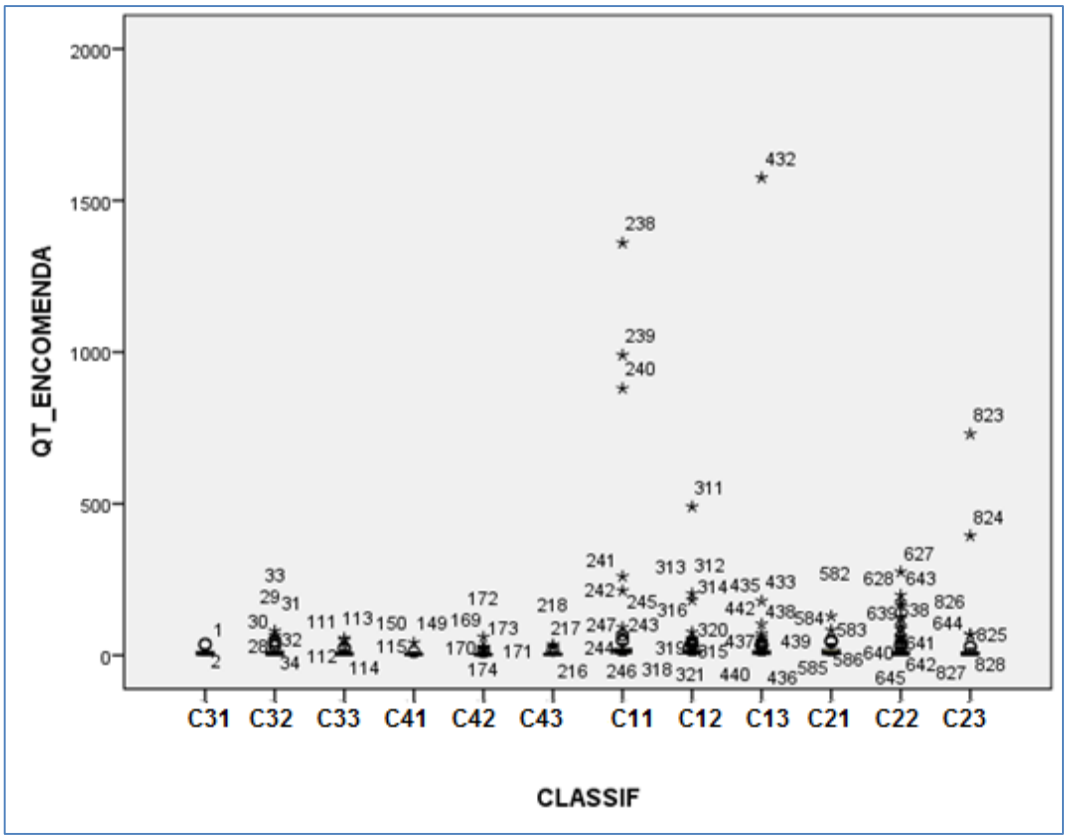


Figura 30 – Diagrama caixa de bigodes da quant. encomendada do produto C (outubro 2013) por classificação do cliente

Anexo 6. PRODUTO C – outubro 2014

Tabela 27 – Quant. encomendada por classificação do cliente - produto C (out. 2014)

Extreme Values					
CLASSIF				Case Number	Value
QT_ENCOMENDA	C11	Highest	1	355	163
			2	579	123
			3	819	64
			4	353	61
			5	663	45
		Lowest	1	373	2
			2	150	3
			3	330	5
			4	77	5
			5	796	8[a]
	C12	Highest	1	229	492
			2	726	395
			3	550	327
			4	331	236
			5	366	127
		Lowest	1	920	1
			2	736	1
			3	681	1
			4	319	1
			5	318	1[b]
	C13	Highest	1	525	218
			2	365	132
			3	116	114
			4	400	45
			5	327	41
		Lowest	1	879	1
			2	593	1
			3	557	1
			4	450	1
			5	214	1[b]
	C21	Highest	1	87	56
			2	865	47
			3	578	23
			4	518	22
			5	206	20
		Lowest	1	890	1
			2	861	1
			3	213	1
			4	185	1
			5	51	1
	C22	Highest	1	587	74
			2	435	56
			3	798	42
			4	38	36
			5	588	36
		Lowest	1	921	1
			2	892	1
			3	834	1
			4	811	1
			5	806	1[b]
	C23	Highest	1	428	46
			2	906	43
			3	583	41
			4	838	38
			5	833	36

		Lowest	1	871	1
			2	714	1
			3	589	1
			4	342	1
			5	313	1[b]
	C31	Highest	1	416	1267
			2	805	640
			3	842	357
			4	103	250
			5	19	232
		Lowest	1	457	1
			2	406	5
			3	199	6
			4	117	9
			5	323	10[c]
	C32	Highest	1	517	690
			2	167	311
			3	731	180
			4	758	140
			5	789	123
		Lowest	1	205	1
			2	121	1
			3	795	2
			4	654	2
			5	462	2
	C33	Highest	1	752	886
			2	926	575
			3	602	378
			4	403	309
			5	179	281
		Lowest	1	952	1
			2	918	1
			3	862	1
			4	824	1
			5	701	1[b]
	C41	Highest	1	797	344
			2	737	265
			3	651	148
			4	510	108
			5	430	90
		Lowest	1	628	1
			2	212	1
			3	415	2
			4	873	5
			5	586	5
	C42	Highest	1	902	970
			2	599	191
			3	889	165
			4	78	132
			5	872	128
		Lowest	1	827	1
			2	715	1
			3	704	1
			4	391	1
			5	329	1[b]
	C43	Highest	1	300	90
			2	597	72
			3	710	66
			4	868	66
			5	894	66
		Lowest	1	942	1
			2	883	1
			3	836	1
			4	622	1
			5	431	1[b]

a. Only a partial list of cases with the value 8 are shown in the table of lower extremes.

b. Only a partial list of cases with the value 1 are shown in the table of lower extremes.
c. Only a partial list of cases with the value 10 are shown in the table of lower extremes.

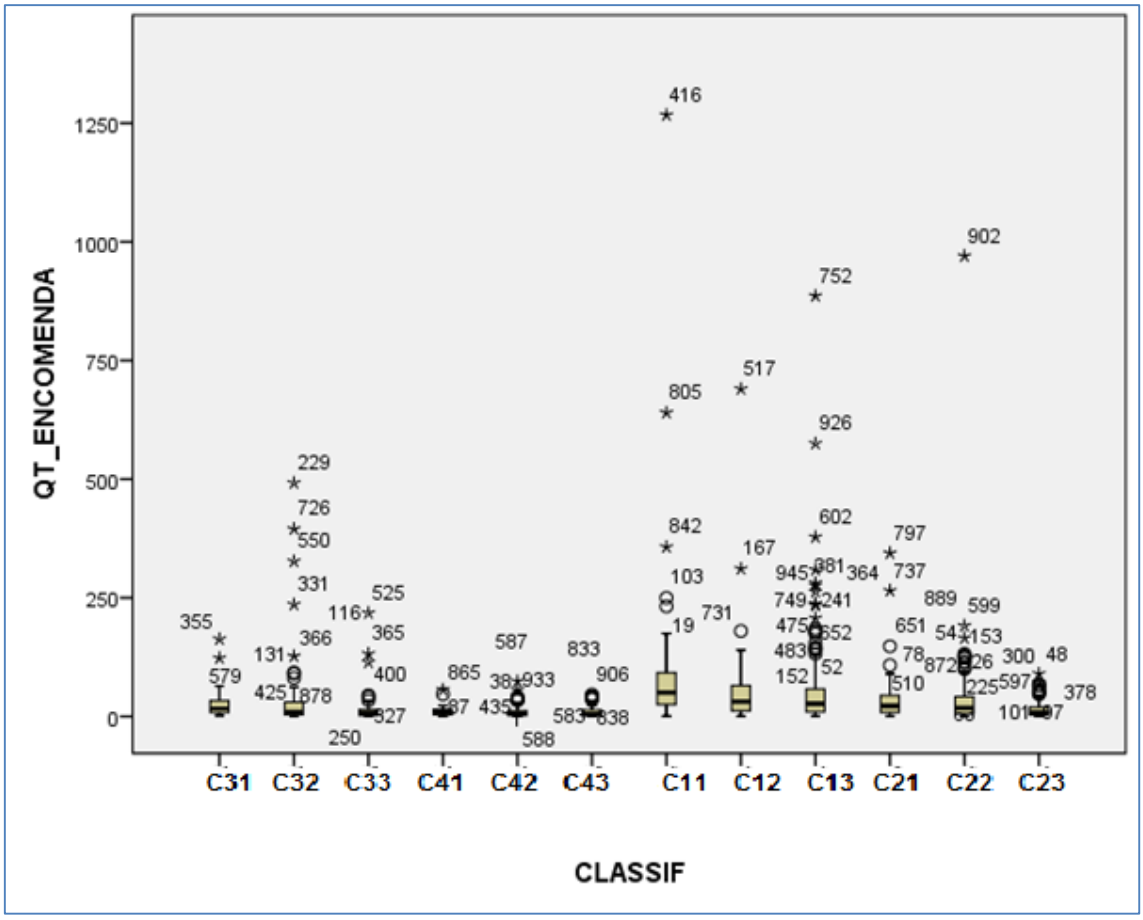


Figura 31 – Diagrama caixa de bigodes da quant. encomendada do produto C (outubro 2014) por classificação do cliente

Apêndices

Apêndice 1. Nº Clientes *outliers* por estratos – elaboração própria com base nos *outputs* do SPSS, (ver. 20)

	Produto A	C11	C12	C13	C21	C22	C23	C31	C32	C33	C41	C42	C43	Total	%Clientes	Nº clientes encomendou
out. 2013	Moderados	0	4	1	1	4	2	2	5	2	1	5	1	28	2,82	993
	Severos	5	3	4	2	4	1	5	6	2	2	12	5	51	5,14	
	Produto A	C11	C12	C13	C21	C22	C23	C31	C32	C33	C41	C42	C43	Total	%Clientes	Nº clientes encomendou
out. 2014	Moderados	1	0	2	0	2	1	2	6	6	3	9	3	35	4,05	864
	Severos	0	4	1	0	2	0	2	4	7	1	7	1	29	3,36	

	Produto B	C11	C12	C13	C21	C22	C23	C31	C32	C33	C41	C42	C43	Total	%Clientes	Nº clientes encomendou
out. 2013	Moderados	2	4	1	0	2	0	2	2	5	1	4	3	26	3,13	832
	Severos	2	4	0	0	1	0	4	7	8	2	5	3	36	4,33	
	Produto B	C11	C12	C13	C21	C22	C23	C31	C32	C33	C41	C42	C43	Total	%Clientes	Nº clientes encomendou
out. 2014	Moderados	1	4	4	6	0	1	5	6	7	2	8	4	48	5,73	838
	Severos	0	4	4	0	0	2	6	6	10	1	10	4	47	5,61	

	Produto C	C11	C12	C13	C21	C22	C23	C31	C32	C33	C41	C42	C43	Total	%Clientes	Nº clientes encomendou
out. 2013	Moderados	2	4	2	1	2	1	3	4	4	3	4	2	32	3,6	888
	Severos	0	3	3	1	5	2	7	6	7	2	16	3	55	6,19	
	Produto C	C11	C12	C13	C21	C22	C23	C31	C32	C33	C41	C42	C43	Total	%Clientes	Nº clientes encomendou
out. 2014	Moderados	0	3	3	1	2	5	2	1	9	2	8	7	43	4,51	953
	Severos	2	5	3	1	5	0	3	2	11	2	3	3	40	4,20	

Apêndice 2. Quantidades encomendadas por clientes *outliers* – elaboração própria com base nos *outputs* do SPSS (ver. 20)

	Produto A	C11	C12	C13	C21	C22	C23	C31	C32	C33	C41	C42	C43	Total	%Quantidade total	Total quantidade clientes encomendaram
out. 2013	Moderados	0	75	16	33	103	49	130	194	49	45	118	13	825	4,32	19117
	Severos	263	84	205	107	173	31	5148	846	1586	124	1231	1242	11040	57,75	
	Produto A	C11	C12	C13	C21	C22	C23	C31	C32	C33	C41	C42	C43	Total	%Quantidade total	Total quantidade clientes encomendaram
out. 2014	Moderados	72	0	174	0	69	21	383	481	362	219	539	209	2283	10,55	21642
	Severos	0	1034	89	0	103	0	2083	1212	646	276	2106	125	6551	30,27	

	Produto B	C11	C12	C13	C21	C22	C23	C31	C32	C33	C41	C42	C43	Total	%Quantidade total	Total quantidade clientes encomendaram
out. 2013	Moderados	69	161	36	0	42	0	338	114	175	58	195	119	1307	7,12	18362
	Severos	88	281	0	0	63	0	3507	675	2330	286	496	1240	8966	48,83	
	Produto B	C11	C12	C13	C21	C22	C23	C31	C32	C33	C41	C42	C43	Total	%Quantidade total	Total quantidade clientes encomendaram
out. 2014	Moderados	64	303	122	152	0	30	366	362	282	146	410	127	2364	11,66	20275
	Severos	0	1293	300	0	0	118	2609	1424	1957	113	2054	274	10142	50,02	

	Produto C	C11	C12	C13	C21	C22	C23	C31	C32	C33	C41	C42	C43	Total	%Quantidade total	Total quantidade clientes encomendaram
out. 2013	Moderados	72	160	45	11	24	16	164	157	135	140	116	97	1137	6,28	18092
	Severos	0	199	133	41	154	62	3864	1060	2084	212	1687	1192	10688	59,08	
	Produto C	C11	C12	C13	C21	C22	C23	C31	C32	C33	C41	C42	C43	Total	%Quantidade total	Total quantidade clientes encomendaram
out. 2014	Moderados	0	216	120	47	183	204	482	180	1302	256	940	391	4321	11,39	37930
	Severos	286	1577	464	56	130	0	2264	1001	3883	609	1326	228	11824	31,17	

Apêndice 3. Valores observados Z-score modificado para o produto A

cliente	classif	qt_encom	mediana	NUM	DA	MAD	MI
4572113414	C43	125	7	79,59	118	5	15,9
10486	C43	72	7	43,84	65	5	8,77
4386678	C43	71	7	43,17	64	5	8,63
10888	C43	66	7	39,8	59	5	7,96
10651	C43	50	7	29	43	5	5,8
7027118572	C43	50	7	29	43	5	5,8
11924	C43	39	7	21,58	32	5	4,32
4646706415	C43	35	7	18,89	28	5	3,78
284867569	C43	31	7	16,19	24	5	3,24
10675774	C43	30	7	15,51	23	5	3,1
2007	C43	30	7	15,51	23	5	3,1
271712	C43	28	7	14,16	21	5	2,83
2868261627	C43	28	7	14,16	21	5	2,83
2244	C43	28	7	14,16	21	5	2,83
10986	C43	27	7	13,49	20	5	2,7
2333673140	C43	26	7	12,82	19	5	2,56
1026413997	C43	25	7	12,14	18	5	2,43
8138919619	C43	20	7	8,77	13	5	1,75
3555	C43	20	7	8,77	13	5	1,75
495998086	C43	18	7	7,42	11	5	1,48
9658	C43	15	7	5,4	8	5	1,08
2480848471	C43	15	7	5,4	8	5	1,08
2130	C43	15	7	5,4	8	5	1,08
11140	C43	14	7	4,72	7	5	0,94
271580	C43	11	7	2,7	4	5	0,54
267484	C43	10	7	2,02	3	5	0,4
64677316	C43	9	7	1,35	2	5	0,27
500128310	C43	9	7	1,35	2	5	0,27
2882	C43	8	7	0,67	1	5	0,13
4255363012	C43	7	7	0	0	5	0
10243	C43	7	7	0	0	5	0
9144	C43	7	7	0	0	5	0
8739838576	C43	6	7	-0,67	1	5	-0,13
9741470	C43	6	7	-0,67	1	5	-0,13
8578	C43	6	7	-0,67	1	5	-0,13
11352	C43	6	7	-0,67	1	5	-0,13
267187	C43	5	7	-1,35	2	5	-0,27
4373219639	C43	5	7	-1,35	2	5	-0,27
515173	C43	4	7	-2,02	3	5	-0,4

9088	C43	4	7	-2,02	3	5	-0,4
5986	C43	4	7	-2,02	3	5	-0,4
3919	C43	4	7	-2,02	3	5	-0,4
7315	C43	4	7	-2,02	3	5	-0,4
5007211	C43	4	7	-2,02	3	5	-0,4
8010367	C43	3	7	-2,7	4	5	-0,54
8206	C43	3	7	-2,7	4	5	-0,54
7468	C43	2	7	-3,37	5	5	-0,67
47977723	C43	2	7	-3,37	5	5	-0,67
10554	C43	2	7	-3,37	5	5	-0,67
11018	C43	2	7	-3,37	5	5	-0,67
8070	C43	2	7	-3,37	5	5	-0,67
2337034912	C43	2	7	-3,37	5	5	-0,67
4239681214	C43	2	7	-3,37	5	5	-0,67
5425743121	C43	2	7	-3,37	5	5	-0,67
1958347103	C43	2	7	-3,37	5	5	-0,67
268185	C43	2	7	-3,37	5	5	-0,67
10526	C43	2	7	-3,37	5	5	-0,67
17989934	C43	1	7	-4,05	6	5	-0,81
435181440	C43	1	7	-4,05	6	5	-0,81
6774	C43	1	7	-4,05	6	5	-0,81
1393042	C43	1	7	-4,05	6	5	-0,81
267560	C43	1	7	-4,05	6	5	-0,81

Apêndice 4. Valores observados Z-score modificado para o produto B

cliente	classif	qt_encom	mediana	NUM	DA	MAD	MI
839337868	C12	610	8	406,1	602	5	81,2
7840084	C12	413	8	273,2	405	5	54,6
4389540	C12	162	8	103,9	154	5	20,8
270972	C12	108	8	67,45	100	5	13,5
271521	C12	91	8	55,98	83	5	11,2
50310284	C12	72	8	43,17	64	5	8,63
6926712970	C12	70	8	41,82	62	5	8,36
13064039	C12	70	8	41,82	62	5	8,36
6798	C12	46	8	25,63	38	5	5,13
3978999733	C12	43	8	23,61	35	5	4,72
2310	C12	38	8	20,24	30	5	4,05
1032896927	C12	36	8	18,89	28	5	3,78
7646	C12	35	8	18,21	27	5	3,64
1967	C12	33	8	16,86	25	5	3,37
9348490916	C12	33	8	16,86	25	5	3,37
271509	C12	32	8	16,19	24	5	3,24
6422	C12	31	8	15,51	23	5	3,1
8098	C12	30	8	14,84	22	5	2,97
8028879544	C12	29	8	14,16	21	5	2,83
2192	C12	29	8	14,16	21	5	2,83
11872	C12	22	8	9,44	14	5	1,89
6634868	C12	21	8	8,77	13	5	1,75
64158268	C12	19	8	7,42	11	5	1,48
9024	C12	17	8	6,07	9	5	1,21
379766181	C12	17	8	6,07	9	5	1,21
658929607	C12	1	8	-4,72	7	5	-0,94
267881	C12	1	8	-4,72	7	5	-0,94
9227	C12	1	8	-4,72	7	5	-0,94
25288858	C12	1	8	-4,72	7	5	-0,94
271200	C12	1	8	-4,72	7	5	-0,94
6255	C12	1	8	-4,72	7	5	-0,94
8242	C12	1	8	-4,72	7	5	-0,94
5488882	C12	14	8	4,05	6	5	0,81
2524704015	C12	2	8	-4,05	6	5	-0,81
6603	C12	2	8	-4,05	6	5	-0,81
2350	C12	2	8	-4,05	6	5	-0,81
3860	C12	2	8	-4,05	6	5	-0,81

2525	C12	2	8	-4,05	6	5	-0,81
4904	C12	13	8	3,37	5	5	0,67
1342186118	C12	3	8	-3,37	5	5	-0,67
11540	C12	3	8	-3,37	5	5	-0,67
2366	C12	3	8	-3,37	5	5	-0,67
10694	C12	13	8	3,37	5	5	0,67
9100	C12	13	8	3,37	5	5	0,67
11163	C12	3	8	-3,37	5	5	-0,67
4381	C12	13	8	3,37	5	5	0,67
7916626454	C12	3	8	-3,37	5	5	-0,67
9351	C12	3	8	-3,37	5	5	-0,67
3671	C12	3	8	-3,37	5	5	-0,67
10519936	C12	3	8	-3,37	5	5	-0,67
7511	C12	12	8	2,7	4	5	0,54
3835566493	C12	4	8	-2,7	4	5	-0,54
3379	C12	12	8	2,7	4	5	0,54
2543098772	C12	12	8	2,7	4	5	0,54
1613193135	C12	4	8	-2,7	4	5	-0,54
9104	C12	4	8	-2,7	4	5	-0,54
271320	C12	4	8	-2,7	4	5	-0,54
267433	C12	4	8	-2,7	4	5	-0,54
270999	C12	4	8	-2,7	4	5	-0,54
8497291899	C12	5	8	-2,02	3	5	-0,4
271294	C12	5	8	-2,02	3	5	-0,4
5280189005	C12	5	8	-2,02	3	5	-0,4
10086	C12	5	8	-2,02	3	5	-0,4
4376383282	C12	5	8	-2,02	3	5	-0,4
12100151	C12	5	8	-2,02	3	5	-0,4
4211628	C12	11	8	2,02	3	5	0,4
2134	C12	11	8	2,02	3	5	0,4
2090919508	C12	6	8	-1,35	2	5	-0,27
3842825000	C12	6	8	-1,35	2	5	-0,27
2698	C12	6	8	-1,35	2	5	-0,27
268182	C12	6	8	-1,35	2	5	-0,27
4456	C12	10	8	1,35	2	5	0,27
267656	C12	6	8	-1,35	2	5	-0,27
7780	C12	6	8	-1,35	2	5	-0,27
154791735	C12	10	8	1,35	2	5	0,27
9866	C12	7	8	-0,67	1	5	-0,13
4829349	C12	9	8	0,67	1	5	0,13
3324	C12	8	8	0	0	5	0
7081	C12	8	8	0	0	5	0
267027	C12	8	8	0	0	5	0

Apêndice 5. Valores observados Z-score modificado para o produto C

cliente	classif	qt_encom	mediana	NUM	DA	MAD	MI
4697030932	C33	886	27	579,4	859	21	27,6
4169086059	C33	575	27	369,6	548	21	17,6
5217	C33	378	27	236,8	351	21	11,3
36573719	C33	309	27	190,2	282	21	9,06
8266	C33	281	27	171,3	254	21	8,16
129284257	C33	276	27	168	249	21	8
15260492	C33	263	27	159,2	236	21	7,58
2115203298	C33	237	27	141,7	210	21	6,75
10659	C33	237	27	141,7	210	21	6,75
10765	C33	234	27	139,6	207	21	6,65
4958	C33	207	27	121,4	180	21	5,78
267988	C33	185	27	106,6	158	21	5,07
1503187383	C33	180	27	103,2	153	21	4,91
6987914	C33	179	27	102,5	152	21	4,88
170740439	C33	174	27	99,15	147	21	4,72
4842092193	C33	155	27	86,34	128	21	4,11
267640	C33	146	27	80,27	119	21	3,82
5834	C33	143	27	78,24	116	21	3,73
3277	C33	140	27	76,22	113	21	3,63
8140568809	C33	132	27	70,82	105	21	3,37
2838	C33	129	27	68,8	102	21	3,28
132537952	C33	107	27	53,96	80	21	2,57
8246	C33	107	27	53,96	80	21	2,57
6480160112	C33	106	27	53,29	79	21	2,54
2607900017	C33	105	27	52,61	78	21	2,51
4386671	C33	101	27	49,91	74	21	2,38
4377116677	C33	101	27	49,91	74	21	2,38
7638	C33	98	27	47,89	71	21	2,28
3293	C33	98	27	47,89	71	21	2,28
4240979789	C33	95	27	45,87	68	21	2,18
3312	C33	94	27	45,19	67	21	2,15
4854949439	C33	94	27	45,19	67	21	2,15
4829383	C33	93	27	44,52	66	21	2,12
1513806641	C33	93	27	44,52	66	21	2,12
267652	C33	92	27	43,84	65	21	2,09
7574	C33	90	27	42,49	63	21	2,02
3117	C33	85	27	39,12	58	21	1,86
9563	C33	81	27	36,42	54	21	1,73
3900	C33	81	27	36,42	54	21	1,73

846085610	C33	78	27	34,4	51	21	1,64
9612	C33	77	27	33,73	50	21	1,61
14182427	C33	75	27	32,38	48	21	1,54
9769	C33	73	27	31,03	46	21	1,48
10518	C33	73	27	31,03	46	21	1,48
7082826108	C33	70	27	29	43	21	1,38
11518481	C33	69	27	28,33	42	21	1,35
4162775693	C33	67	27	26,98	40	21	1,28
11638	C33	67	27	26,98	40	21	1,28
9698	C33	59	27	21,58	32	21	1,03
11045	C33	57	27	20,24	30	21	0,96
10470	C33	57	27	20,24	30	21	0,96
9446	C33	56	27	19,56	29	21	0,93
4211279	C33	54	27	18,21	27	21	0,87
10498	C33	54	27	18,21	27	21	0,87
5525	C33	54	27	18,21	27	21	0,87
11380	C33	52	27	16,86	25	21	0,8
6480829552	C33	50	27	15,51	23	21	0,74
11507955	C33	50	27	15,51	23	21	0,74
2637	C33	49	27	14,84	22	21	0,71
8190	C33	49	27	14,84	22	21	0,71
2090779582	C33	49	27	14,84	22	21	0,71
3951373636	C33	48	27	14,16	21	21	0,67
9798	C33	48	27	14,16	21	21	0,67
6024429	C33	47	27	13,49	20	21	0,64
267672	C33	46	27	12,82	19	21	0,61
9932	C33	45	27	12,14	18	21	0,58
3438993280	C33	45	27	12,14	18	21	0,58
271603	C33	45	27	12,14	18	21	0,58
7447185	C33	45	27	12,14	18	21	0,58
267150	C33	45	27	12,14	18	21	0,58
5606564124	C33	44	27	11,47	17	21	0,55
271236	C33	43	27	10,79	16	21	0,51
4372648	C33	41	27	9,44	14	21	0,45
3257	C33	41	27	9,44	14	21	0,45
23815143	C33	41	27	9,44	14	21	0,45
10924	C33	39	27	8,09	12	21	0,39
3607	C33	39	27	8,09	12	21	0,39
12795545	C33	37	27	6,75	10	21	0,32
1520579926	C33	37	27	6,75	10	21	0,32
10290	C33	36	27	6,07	9	21	0,29
8603656789	C33	36	27	6,07	9	21	0,29
4232928147	C33	35	27	5,4	8	21	0,26
271007	C33	35	27	5,4	8	21	0,26

3808	C33	34	27	4,72	7	21	0,22
2613	C33	33	27	4,05	6	21	0,19
7197	C33	33	27	4,05	6	21	0,19
5952794	C33	33	27	4,05	6	21	0,19
5780142	C33	33	27	4,05	6	21	0,19
8550	C33	32	27	3,37	5	21	0,16
9250452230	C33	32	27	3,37	5	21	0,16
10922498	C33	32	27	3,37	5	21	0,16
698413524	C33	31	27	2,7	4	21	0,13
123265603	C33	31	27	2,7	4	21	0,13
5696194841	C33	30	27	2,02	3	21	0,1
81998868	C33	29	27	1,35	2	21	0,06
4677	C33	29	27	1,35	2	21	0,06
6723	C33	28	27	0,67	1	21	0,03
6997	C33	27	27	0	0	21	0
4211670	C33	27	27	0	0	21	0
11502	C33	26	27	-0,67	1	21	-0,03
267660	C33	25	27	-1,35	2	21	-0,06
3707	C33	25	27	-1,35	2	21	-0,06
4998162232	C33	24	27	-2,02	3	21	-0,1
8888	C33	24	27	-2,02	3	21	-0,1
267312	C33	24	27	-2,02	3	21	-0,1
182908029	C33	23	27	-2,7	4	21	-0,13
64161847	C33	23	27	-2,7	4	21	-0,13
2118	C33	22	27	-3,37	5	21	-0,16
6582836984	C33	22	27	-3,37	5	21	-0,16
1983	C33	22	27	-3,37	5	21	-0,16
9191	C33	21	27	-4,05	6	21	-0,19
9211	C33	21	27	-4,05	6	21	-0,19
8015	C33	21	27	-4,05	6	21	-0,19
4210	C33	21	27	-4,05	6	21	-0,19
267931	C33	21	27	-4,05	6	21	-0,19
29931160	C33	20	27	-4,72	7	21	-0,22
4831422	C33	20	27	-4,72	7	21	-0,22
708539331	C33	20	27	-4,72	7	21	-0,22
2301556722	C33	20	27	-4,72	7	21	-0,22
2517331944	C33	19	27	-5,4	8	21	-0,26
3027562671	C33	19	27	-5,4	8	21	-0,26
10519264	C33	18	27	-6,07	9	21	-0,29
268067	C33	18	27	-6,07	9	21	-0,29
11561	C33	17	27	-6,75	10	21	-0,32
4075	C33	17	27	-6,75	10	21	-0,32
10816	C33	16	27	-7,42	11	21	-0,35
9870	C33	16	27	-7,42	11	21	-0,35

29935782	C33	16	27	-7,42	11	21	-0,35
7772	C33	15	27	-8,09	12	21	-0,39
6175	C33	15	27	-8,09	12	21	-0,39
158681834	C33	15	27	-8,09	12	21	-0,39
5568	C33	15	27	-8,09	12	21	-0,39
1909	C33	14	27	-8,77	13	21	-0,42
802842	C33	12	27	-10,12	15	21	-0,48
9737	C33	12	27	-10,12	15	21	-0,48
4049	C33	12	27	-10,12	15	21	-0,48
5107358	C33	12	27	-10,12	15	21	-0,48
2934	C33	12	27	-10,12	15	21	-0,48
11697410	C33	12	27	-10,12	15	21	-0,48
3755	C33	11	27	-10,79	16	21	-0,51
4010	C33	11	27	-10,79	16	21	-0,51
159401718	C33	11	27	-10,79	16	21	-0,51
2593	C33	11	27	-10,79	16	21	-0,51
9791	C33	11	27	-10,79	16	21	-0,51
2400386941	C33	11	27	-10,79	16	21	-0,51
344562598	C33	10	27	-11,47	17	21	-0,55
6634	C33	10	27	-11,47	17	21	-0,55
8948	C33	10	27	-11,47	17	21	-0,55
2749917642	C33	9	27	-12,14	18	21	-0,58
271290	C33	9	27	-12,14	18	21	-0,58
1342794506	C33	8	27	-12,82	19	21	-0,61
9897	C33	8	27	-12,82	19	21	-0,61
9051032159	C33	7	27	-13,49	20	21	-0,64
2293	C33	7	27	-13,49	20	21	-0,64
6794	C33	7	27	-13,49	20	21	-0,64
829242	C33	7	27	-13,49	20	21	-0,64
8944	C33	6	27	-14,16	21	21	-0,67
271646	C33	6	27	-14,16	21	21	-0,67
267425	C33	6	27	-14,16	21	21	-0,67
2090741690	C33	6	27	-14,16	21	21	-0,67
183472643	C33	6	27	-14,16	21	21	-0,67
33601915	C33	6	27	-14,16	21	21	-0,67
13006750	C33	5	27	-14,84	22	21	-0,71
8110	C33	5	27	-14,84	22	21	-0,71
964876138	C33	5	27	-14,84	22	21	-0,71
8685	C33	5	27	-14,84	22	21	-0,71
3328	C33	5	27	-14,84	22	21	-0,71
2970	C33	4	27	-15,51	23	21	-0,74
8606	C33	4	27	-15,51	23	21	-0,74
64153332	C33	4	27	-15,51	23	21	-0,74
7888	C33	4	27	-15,51	23	21	-0,74

16109617	C33	4	27	-15,51	23	21	-0,74
46813975	C33	4	27	-15,51	23	21	-0,74
8222	C33	4	27	-15,51	23	21	-0,74
1618	C33	3	27	-16,19	24	21	-0,77
4226	C33	3	27	-16,19	24	21	-0,77
3559511	C33	3	27	-16,19	24	21	-0,77
6241963998	C33	3	27	-16,19	24	21	-0,77
6155	C33	3	27	-16,19	24	21	-0,77
44658343	C33	3	27	-16,19	24	21	-0,77
4053	C33	3	27	-16,19	24	21	-0,77
9215	C33	3	27	-16,19	24	21	-0,77
267754	C33	3	27	-16,19	24	21	-0,77
3351	C33	2	27	-16,86	25	21	-0,8
8019	C33	2	27	-16,86	25	21	-0,8
25288387	C33	2	27	-16,86	25	21	-0,8
13417948	C33	1	27	-17,54	26	21	-0,84
7442586065	C33	1	27	-17,54	26	21	-0,84
267648	C33	1	27	-17,54	26	21	-0,84
17619759	C33	1	27	-17,54	26	21	-0,84
505138325	C33	1	27	-17,54	26	21	-0,84
4773	C33	1	27	-17,54	26	21	-0,84
6521457059	C33	1	27	-17,54	26	21	-0,84
4082505893	C33	1	27	-17,54	26	21	-0,84
5426017038	C33	1	27	-17,54	26	21	-0,84
4408819742	C33	1	27	-17,54	26	21	-0,84

Apêndice 6. Algoritmo do processo de previsão de vendas (PR_CALC_FOR_HW)

```
CREATE OR REPLACE PACKAGE BODY EVOL.OCP_MAYA_FOR_PKG IS
/******************************************************************************/
/******************************************************************************/
PROCEDURE DDL_TABELA ( cComando VARCHAR2 ,cNomeTabela VARCHAR2 ,cParametros VARCHAR2 ) IS
 nTratamento INTEGER;
 cLinhaComando VARCHAR2(4000);
 nResultado BINARY_INTEGER;
 wk_erro varchar2(4000):=NULL;
BEGIN
 IF cComando = 'TRUNCATE' THEN
   cLinhaComando := 'TRUNCATE TABLE '||cNomeTabela||cParametros;
 ELSIF cComando = 'CREATE' THEN
   cLinhaComando := ' CREATE TABLE '||substr(cNomeTabela,1,20)||'_'||TO_CHAR(SYSDATE,'YYMMDD')
            ||' AS SELECT * FROM '||cNomeTabela;
 ELSE
  NULL;
 END IF;
  ---
 nTratamento := DBMS_SQL.OPEN_CURSOR;
  ---
 DBMS_SQL.PARSE( nTratamento, cLinhaComando, DBMS_SQL.NATIVE );
  ---
 nResultado := DBMS_SQL.EXECUTE( nTratamento );
  ---
 DBMS_SQL.CLOSE_CURSOR( nTratamento );
  ---
 EXCEPTION
  WHEN OTHERS THEN
 wk_erro := SUBSTR(SQLERRM,1,222);
 arm_mail2_pkg.mail (sender => 'OCP_MAYA_FOR_PKG',
            recipients => 'augusto.ribeiro@ocp.pt',
            subject => 'OCP_MAYA_FOR_PKG',
            MESSAGE => wk_erro);
END;
/******************************************************************************/
/******************************************************************************/
procedure pr_calc_for_hw is
  ---
  CURSOR C_FOR  IS
    select p.reg_id reg_id_produto, f.ano_mes, f.REG_ID_ESTABELECIMENTO estab_id,  sum(f.QUANT_PEDIDA) d
    from OCP_FACTURACAO_CONNECT_est F, arm.arm_produtos@EVOL_PORTALOCP.CELESIOGROUP.COM P
    where f.produto_id = p.produto_id
```

```
      group by p.reg_id, f.ano_mes, f.REG_ID_ESTABELECIMENTO;
   ---
   cursor c_dados_error  is
      select REG_ID_PRODUTO, estab_id, ANO_MES, dados, hw_for1, ft_mes_ant
      from OCP_maya_vendas_for
      where nvl(dados,0)>0
      --and nvl(ft_mes_ant,0)>0
      and ano_mes < to_char(sysdate,'yyyymm')
      order by REG_ID_PRODUTO, estab_id, ANO_MES;
   ---
   /* process smoothing constant */
   process_const number;
   ---
   /* trend smoothing constant */
   damped_const number;
   ---
   trend_const number;
   ---
   process_const_1 number;
   trend_const_1 number;
   ---
   wk_erro varchar2(4000):=NULL;
   wk1 number;
   wk2 number;
   wk_operadores varchar2(1000);
   wk_reg_id_produto number;
   wk_estab_id number;
begin
   ---
   ---
   OCP_MAYA_FOR_PKG.DDL_TABELA( 'TRUNCATE', 'OCP_MAYA_VENDAS_FOR_ERROR', NULL );
   commit;
   delete OCP_MAYA_VENDAS_FOR_ERROR;
   commit;
   ---
   ---
   FOR pr_c in 1..9
   LOOP
      FOR tr_c in 1..9
      LOOP
         FOR damped in 1..9
         LOOP
            ---
            process_const := pr_c/10;
            trend_const := tr_c/10;
            damped_const := damped/10;
            ---
            OCP_MAYA_FOR_PKG.DDL_TABELA( 'TRUNCATE', 'OCP_MAYA_VENDAS_FOR', NULL );
            commit;
```

```
                delete ocp_maya_vendas_for;
                commit;
                ---
                for f in c_for
                loop
                   insert into ocp_maya_vendas_for
                      (ANO_MES,ESTAB_ID,TIPODOC_FACT,TIPODOC_AVI,
                      REG_ID_PRODUTO,REG_ID_PRODUTO_ARMAZEM,PRODUTO_ID,
                      DADOS,HW_FOR1,HW_TREND1,HW_FOR2,HW_TREND2,tfo,bonus, dados_hw)
                   values
                      (f.ANO_MES,f.estab_id,null,null,
                      f.REG_ID_PRODUTO,null,null,
                      f.D,null,null,null,null,null,null, f.D);
                   ---
                   commit;
                end loop;
                ---
                ---
                OCP_MAYA_FOR_PKG.pr_calc_for_const_hw (process_const, trend_const, damped_const);
                ---
                ---
                --- SQUARE  ERROR
                ---
                FOR E1 in c_dados_error
                LOOP
                   ---
                   wk_reg_id_produto := E1.REG_ID_PRODUTO;
                   wk_estab_id := E1.ESTAB_ID;
                   ---
                   wk1 := ABS( nvl(E1.dados,0) - nvl(E1.ft_mes_ant,0) );
                   wk2 :=  nvl(E1.dados,0) + nvl(E1.ft_mes_ant,0)  / 2;
                   ---
                   if nvl(wk2,0) = 0 then
                      wk2:=1;
                   end if;
                   ---
                   wk_operadores := nvl(E1.dados,0)||' '||nvl(E1.ft_mes_ant,0)||' '||process_const||' '||trend_const||'
'||damped_const||E1.REG_ID_PRODUTO||' '||E1.ANO_MES||' '||E1.ESTAB_ID;
                   ---
                   insert into EVOL.OCP_MAYA_VENDAS_FOR_ERROR
                      (ANO_MES,ESTAB_ID,REG_ID_PRODUTO,process_const,trend_const, damped_const ,square_error,
operadores)
                   values
                      --(E1.ANO_MES,E1.ESTAB_ID,E1.REG_ID_PRODUTO,process_const,trend_const, trunc(nvl(E1.dados,0)-
nvl(E1.ft_mes_ant,0)) );
                      --(E1.ANO_MES,E1.ESTAB_ID,E1.REG_ID_PRODUTO,process_const,trend_const,
trunc(power(nvl(E1.dados,0)-nvl(E1.ft_mes_ant,0),2)) );
                      (E1.ANO_MES,E1.ESTAB_ID,E1.REG_ID_PRODUTO,process_const,trend_const, damped_const, wk1/wk2,
wk_operadores );
                   ---
                   commit;
```

```
            END LOOP;
            ---
            ---
          END LOOP;
        END LOOP;
    END LOOP;
    ---
    ---
    OCP_MAYA_FOR_PKG.DDL_TABELA( 'TRUNCATE', 'OCP_MAYA_VENDAS_FOR', NULL );
    commit;
    delete ocp_maya_vendas_for;
    commit;
    ---
    for f in c_for
    loop
      insert into ocp_maya_vendas_for
        (ANO_MES,ESTAB_ID,TIPODOC_FACT,TIPODOC_AVI,
        REG_ID_PRODUTO,REG_ID_PRODUTO_ARMAZEM,PRODUTO_ID,
        DADOS,HW_FOR1,HW_TREND1,HW_FOR2,HW_TREND2,tfo,bonus, dados_hw)
      values
        (f.ANO_MES,f.estab_id,null,null,
        f.REG_ID_PRODUTO,null,null,
        f.D,null,null,null,null,null,null, f.D);
      ---
      commit;
    end loop;
    --- chama processo novamente com 0, 0  => processo verifica o menor erro e aplica o process e trend constants
    OCP_MAYA_FOR_PKG.pr_calc_for_const_hw (0, 0, 0);
    ---
    ---
EXCEPTION WHEN others THEN
   wk_erro := SUBSTR(SQLERRM,1,222);
   arm_mail2_pkg.mail (sender => 'olga',
                       recipients => 'augusto.ribeiro@ocp.pt',
                       subject => 'pr_calc_for_hw',
                       MESSAGE => wk_operadores||'   '||wk_erro);
END;



/*****************************************************************************/
/*****************************************************************************/
procedure pr_calc_for_const_hw (i_process_const in number , i_trend_const in number, i_damped in number) is
    ---
    CURSOR C_FOR  IS
      select p.reg_id reg_id_produto, f.ano_mes, f.REG_ID_ESTABELECIMENTO estab_id,  sum(f.QUANT_PEDIDA) d
      from OCP_FACTURACAO_CONNECT_est F, arm.arm_produtos@EVOL_PORTALOCP.CELESIOGROUP.COM P
      where f.produto_id = p.produto_id
      group by p.reg_id, f.ano_mes, f.REG_ID_ESTABELECIMENTO;
```

```
   ---
   cursor c_estab is
      select 2 reg_id_estab from dual
      union all
      select 3 reg_id_estab from dual
      union all
      select 4 reg_id_estab from dual
      union all
      select 5 reg_id_estab from dual
      union all
      select 6 reg_id_estab from dual
      union all
      select 7 reg_id_estab from dual
      union all
      select 802344 reg_id_estab from dual
      ;
   ---
   cursor c_produtos (i_REG_ID_estab in number) is
      select distinct p.reg_id reg_id_produto
      from OCP_FACTURACAO_CONNECT_est F, arm.arm_produtos@EVOL_PORTALOCP.CELESIOGROUP.COM P
      where f.produto_id = p.produto_id
      and F.REG_ID_ESTABELECIMENTO = i_REG_ID_estab
      --and p.reg_id in ( 427795 , 4676070)
      ;
/*
ASPIRINA GR 100 MG COMP.GR X30
BISOLVON LINCTUS ADULTO 1.6 MG/ML 200 ML XAR. X1
*/
   ---
   cursor c_dados (i_REG_ID_PRODUTO1 in number, i_REG_ID_estab1 in number) is
      select REG_ID_PRODUTO, estab_id, ANO_MES, dados
      from OCP_maya_vendas_for
      where reg_id_produto=i_REG_ID_PRODUTO1
      and ESTAB_ID = i_REG_ID_estab1
      order by REG_ID_PRODUTO, estab_id, ANO_MES;
   ---
   cursor c_dados_mes_ant(i_ANO_MES in varchar2, i_estab_id in varchar2, i_REG_ID_PRODUTO in number) is
      select HW_FOR1,HW_TREND1, dados, st, mt, ft
      from OCP_maya_vendas_for
      where ANO_MES = i_ANO_MES
          and estab_id = i_estab_id
          and REG_ID_PRODUTO = i_REG_ID_PRODUTO;
   ---
   --- calulo forecast   for
   cursor c_dados_for (i_REG_ID_PRODUTO in number, i_reg_id_estabelecimento in number, i_ANO_MES in varchar2) is
      select hw_for1
      from OCP_maya_vendas_for
      where REG_ID_PRODUTO = i_REG_ID_PRODUTO
      and estab_id = i_reg_id_estabelecimento
```

```
   and ano_mes = i_ANO_MES;
---
wk_HW_FOR1_ant number;
--wk_HW_FOR2_ant number;
wk_HW_FOR3_ant number;
---
cursor c_dados_error  is
   select REG_ID_PRODUTO, estab_id, ANO_MES, dados, FT_MES_ANT
   from OCP_maya_vendas_for
   where nvl(dados,0)>0
   --and nvl(FT_MES_ANT,0)>0
   and ano_mes < to_char(sysdate,'yyyymm')
   order by REG_ID_PRODUTO, estab_id, ANO_MES;
---
wk_dados_hw number;
wk_HW_FOR1 number;
xwk_dados_hw number;
xwk_HW_FOR1 number;
---
wk_cont number;
wk_dados_reg1 number;
wk_tfo_ant number;
---
/* process smoothing constant */
process_const number ;
---
/* trend smoothing constant */
trend_const number ;
---
damped_const number;
---
process_const_1 number;
trend_const_1 number;
---
---
wk_ano number;
wk_mes varchar(2);
wk_anomes varchar2(6);
wk_erro varchar(4000);
wk_HW_TREND1_ant number;
wk_dados_ant number;
wk_dados_hw_ant number;
---
st_ant number;
mt_ant number;
ft_ant number;
---
wk_HW_FOR2_ant number;
wk_HW_TREND2_ant number;
```

```
    ---
    wk_for1 number;
    wk_for2 number;
    wk_step varchar(10);
    ---
    cursor c_minimo_square_error (i_reg_id_estab1 number, i_reg_id_produto1 number) is
        select PROCESS_CONST, TREND_CONST, damped_const, sum(ABS(square_error)) err
        from OCP_MAYA_VENDAS_FOR_ERROR
        where estab_id=i_reg_id_estab1
        and reg_id_produto = i_reg_id_produto1
        and ano_mes between to_char(add_months(sysdate,-5),'yyyymm') and to_char(add_months(sysdate,-1),'yyyymm')
        group by process_const, trend_const, damped_const
        order by sum(ABS(square_error)) asc;
    wk_square_error number;
    wk_reg_id_produto number;
    wk_estab number;
    ---
    wk_st number;
    wk_mt number;
    wk_ft number;
    ---
    ---
    cursor c_dados_mes_ant2(i_ANO_MES in varchar2, i_estab_id in varchar2, i_REG_ID_PRODUTO in number) is
        select HW_FOR1,HW_TREND1, dados, st, mt, ft, FT_MES_ANT
        from OCP_maya_vendas_for
        where ANO_MES = i_ANO_MES
            and estab_id = i_estab_id
            and REG_ID_PRODUTO = i_REG_ID_PRODUTO;
    st_ant2 number;
    mt_ant2 number;
    ft2 number;
    ft_ant2 number;
    wk_erroforecast number;
begin
    ---
        ---
        process_const := i_process_const;
        trend_const := i_trend_const;
        damped_const := i_damped;
        ---
        process_const_1  :=   1 - process_const;
        trend_const_1  :=   1 - trend_const;
        ---
        ---
        ---
        FOR R0 IN C_ESTAB
        LOOP
        begin
            ---
```

```
FOR P in c_produtos (R0.reg_id_estab)
LOOP
   ---
   if nvl(i_process_const,0)=0 and nvl(i_trend_const,0)=0 then
      ---
      wk_reg_id_produto := p.reg_id_produto;
      wk_estab := R0.reg_id_estab;
      process_const := NULL;
      trend_const := NULL;
      damped_const := NULL;
      ---
      open c_minimo_square_error (R0.reg_id_estab, p.reg_id_produto);
      fetch c_minimo_square_error into process_const, trend_const, damped_const, wk_square_error;
      close c_minimo_square_error;
      ---
      process_const := nvl(process_const, 0.3);
      trend_const := nvl(trend_const, 0.3);
      damped_const := nvl(damped_const, 0.3);
      ---
      process_const_1  :=   1 - process_const;
      trend_const_1  :=   1 - trend_const;
   end if;
   ---
   --- 1º
   wk_cont:=0;
   for rf1 in c_dados(p.reg_id_produto, R0.reg_id_estab)
   loop
      ---
      wk_cont := wk_cont+1;
      if wk_cont = 1 then
         wk_dados_reg1 := rf1.dados;
      end if;
      ---
      if wk_cont = 2 then
         update OCP_maya_vendas_for
         set HW_FOR1 = round(nvl(rf1.dados,0),6),
            HW_TREND1 = round(nvl(rf1.dados,0) - nvl(wk_dados_reg1,0),6),
            process_const_ESCOLHA = process_const,
            trend_const_ESCOLHA = trend_const,
            damped = damped_const,
            st = round(nvl(rf1.dados,0),6),
            mt = round(nvl(rf1.dados,0) / greatest(nvl(wk_dados_reg1,0),1),6),
            ft = round(nvl(rf1.dados,0),6),
            ft_mes_ant = round(nvl(rf1.dados,0),6)
         where ANO_MES = rf1.ANO_MES
            and estab_id = rf1.estab_id
            and REG_ID_PRODUTO = rf1.REG_ID_PRODUTO;
         ---
         commit;
```

```
            end if;
            ---
        end loop;
        ---
        ---
        --- 2º
        wk_cont:=0;
        for rf2 in c_dados(p.reg_id_produto, R0.reg_id_estab)
        loop
            ---
            wk_cont := wk_cont+1;
            if wk_cont >= 3 then
                wk_mes:=substr(rf2.ano_mes,5,2);
                wk_ano:=substr(rf2.ano_mes,1,4);
                if to_number(wk_mes) = 1 then
                    wk_mes := '12';
                    wk_ano:= to_number(wk_ano)-1;
                else
                    wk_mes:= to_number(wk_mes)-1;
                end if;
                wk_anomes := to_char(wk_ano)||lpad(wk_mes,2,'0');
                ---
                wk_HW_FOR1_ant:=0;
                wk_HW_TREND1_ant:=0;
                wk_dados_ant:=0;
                ---
                st_ant:=0;
                mt_ant:=0;
                ft_ant:=0;
                ---
                open c_dados_mes_ant(wk_anomes, rf2.estab_id, rf2.REG_ID_PRODUTO);
                fetch c_dados_mes_ant into wk_HW_FOR1_ant,wk_HW_TREND1_ant, wk_dados_ant, st_ant, mt_ant, ft_ant;
                close c_dados_mes_ant;
                ---
                wk_for1:=
(process_const*(wk_HW_FOR1_ant+wk_HW_TREND1_ant))+(process_const_1*nvl(wk_dados_ant,0)) ;
                ---
                if nvl(wk_for1,0) <= 0 then
                    wk_for1 := wk_HW_FOR1_ant;
                end if;
                ---
                wk_st := (process_const*nvl(rf2.dados,0)) + (process_const_1*(nvl(st_ant,0)*
POWER(nvl(mt_ant,0),damped_const) ));
                wk_st := round(wk_st,6);
                ---
                wk_mt := (trend_const*(nvl(wk_st,0)/greatest(nvl(st_ant,0),1))) + (trend_const_1*
POWER(nvl(mt_ant,0),damped_const) );
                wk_mt := round(wk_mt,6);
                ---
                wk_ft := round(wk_st * POWER(wk_mt,damped_const) , 0);
```

```
                    ---
                    if nvl(wk_ft,0) <= 0 then
                       wk_st := 1;
                       wk_mt := 1;
                       wk_ft := 1;
                    end if;
                    ---
                    update OCP_maya_vendas_for
                    set HW_FOR1 = round(wk_for1,6),
                       HW_TREND1 = round((trend_const*wk_HW_TREND1_ant)+(trend_const_1 *(wk_for1-
wk_HW_FOR1_ant)),6),
                       process_const_ESCOLHA = process_const,
                       trend_const_ESCOLHA = trend_const,
                       damped = damped_const,
                       st = wk_st,
                       mt = wk_mt,
                       ft = wk_ft,
                       ft_mes_ant = ft_ant
                    where ANO_MES = rf2.ANO_MES
                       and estab_id = rf2.estab_id
                       and REG_ID_PRODUTO = rf2.REG_ID_PRODUTO;
                    ---
                    commit;
                 end if;
                 ---
              end loop;
              ---
              commit;
              ---
           END LOOP;
           ---
           commit;
           ---
           ---
           ---
           ---
           if nvl(i_process_const,0)=0 and nvl(i_trend_const,0)=0 then
              --- criar 2 meses forecast   dados_hw  fase fianl depois de calcular os erros por cosntante
              ---
              FOR P in c_produtos (R0.reg_id_estab)
              LOOP
                 wk_cont:=0;
                 for rf1 in c_dados(p.reg_id_produto, R0.reg_id_estab)
                 loop
                    ---
                    st_ant:=0;
                    mt_ant:=0;
                    ft_ant:=0;
                    ---
```

```
open c_dados_mes_ant(to_char(add_months(sysdate,-1),'yyyymm'), R0.reg_id_estab, p.reg_id_produto);
fetch c_dados_mes_ant into wk_HW_FOR1_ant,wk_HW_TREND1_ant, wk_dados_ant, st_ant, mt_ant, ft_ant;
close c_dados_mes_ant;
---
wk_ft := round(st_ant * POWER(mt_ant,2) , 0); /* 2 meses a frente do último mês completo */
---
if nvl(wk_ft,0) <= 0 then
   wk_st := 1;
   wk_mt := 1;
   wk_ft := 1;
end if;
---
---
begin
   insert into OCP_maya_vendas_for
      (ANO_MES,ESTAB_ID,REG_ID_PRODUTO,HW_FOR1,
      st, mt, ft, ft_mes_ant)
   values
      (to_char(add_months(sysdate,1),'yyyymm'), rf1.estab_id, rf1.REG_ID_PRODUTO, null,
       wk_st, wk_mt, ft_ant, wk_ft );
   commit;
exception when others then
   update OCP_maya_vendas_for
   set HW_FOR1 = null,
      st = wk_st,
      mt = wk_mt,
      ft = ft_ant,
      ft_mes_ant = wk_ft
   where ANO_MES = to_char(add_months(sysdate,1),'yyyymm')
   and ESTAB_ID = rf1.estab_id
   and REG_ID_PRODUTO = rf1.REG_ID_PRODUTO;
   commit;
end;
---
---
---
---
---
st_ant:=0;
mt_ant:=0;
ft_ant:=0;
---
open c_dados_mes_ant(to_char(add_months(sysdate,-1),'yyyymm'), R0.reg_id_estab, p.reg_id_produto);
fetch c_dados_mes_ant into wk_HW_FOR1_ant,wk_HW_TREND1_ant, wk_dados_ant, st_ant, mt_ant, ft_ant;
close c_dados_mes_ant;
---
wk_ft := round(st_ant * POWER(mt_ant,3) , 0);  /* 3 meses a frente do último mês completo */
---
if nvl(wk_ft,0) <= 0 then
```

```
                wk_st := 1;
                wk_mt := 1;
                wk_ft := 1;
            end if;
            ---
            ---
            begin
                insert into OCP_maya_vendas_for
                    (ANO_MES,ESTAB_ID,REG_ID_PRODUTO,HW_FOR1,
                    st, mt, ft, ft_mes_ant)
                values
                    (to_char(add_months(sysdate,2),'yyyymm'), rf1.estab_id, rf1.REG_ID_PRODUTO, null,
                     wk_st, wk_mt, ft_ant2, wk_ft );
                commit;
            exception when others then
                update OCP_maya_vendas_for
                set HW_FOR1 = null,
                    st = wk_st,
                    mt = wk_mt,
                    ft = ft_ant,
                    ft_mes_ant = wk_ft
                where ANO_MES = to_char(add_months(sysdate,2),'yyyymm')
                and ESTAB_ID = rf1.estab_id
                and REG_ID_PRODUTO = rf1.REG_ID_PRODUTO;
                commit;
            end;
            ---
        end loop;
        commit;
        ---
    END LOOP;
    ---
    commit;
    ---
end if;
---
---
commit;
---
---
exception when others then
    wk_erro := SUBSTR(SQLERRM,1,222);
    arm_mail2_pkg.mail (sender => 'olga',
                    recipients => 'augusto.ribeiro@ocp.pt',
                    subject => 'pr_calc_for_const_hw',
                    MESSAGE => wk_reg_id_produto||'-'||wk_estab||'-'||wk_erro);
end;
END LOOP;
---
```

```
   ---
   ---
EXCEPTION WHEN others THEN
  wk_erro := SUBSTR(SQLERRM,1,222);
  arm_mail2_pkg.mail (sender => 'olga',
                      recipients => 'augusto.ribeiro@ocp.pt',
                      subject => 'pr_calc_for_const_hw',
                      MESSAGE => wk_reg_id_produto||'-'||wk_estab||'-'||wk_erro);
END;


/*****************************************************************************/
/*****************************************************************************/

/*****************************************************************************/
/*****************************************************************************/
  PROCEDURE pr_calc_for_rl IS
   ---
   wk_erro VARCHAR2(4000);
   wk_ano VARCHAR2(4);
   wk_mes NUMBER;
   wk_conta NUMBER:=0;
   wk_num_meses_forecast NUMBER;
   wk_meses_vendas NUMBER;
   wk_data_inicio DATE;
   wk_reg_id_estabelecimento number;
   wk_reg_id_produto number;
   ---
   CURSOR c_estab IS
      SELECT distinct REG_ID_ESTABELECIMENTO
      FROM   OCP_FACTURACAO_CONNECT_est;
   ---
   CURSOR c_prod is
      SELECT distinct reg_id_produto
      FROM ocp_maya_vendas_for;
   ---
   -- Regressão linear. Aplicar a funcao y = m*x + b, em que:
   -- y -> valor a prever
   -- x -> periodo a prever
   -- m -> declive da reta ( slope )
   -- b -> ponto de origem da reta ( intersection )
   -- A query à tabela ocp_campanha_prods_sol apenas existe para multiplicar a query de baixo x vezes.

   CURSOR c_fore (p_reg_id_estabelecimento in number,
                  p_reg_id_produto in number ) IS
      SELECT cat.*, fore.*, CEIL((slope * ( num_lin_max + num_lin ) + intercept) / factor_sazonal ) AS qtd_forecast
      FROM (
          SELECT ADD_MONTHS(TRUNC(SYSDATE,'MM'), ROWNUM-21)  DATA
              ,TO_CHAR(ADD_MONTHS(TRUNC(SYSDATE,'MM'), ROWNUM-21) ,'Q') trimestre_fore
```

```
                    ,ROWNUM num_lin
                FROM ocp_campanha_prods_sol
               WHERE ROWNUM <= 23 -- numero de meses que se quer prever
             ) cat
           INNER JOIN (

            SELECT  DISTINCT reg_id_produto, reg_id_estabelecimento, trimestre, factor_sazonal,
                MAX(num_lin) OVER( PARTITION BY reg_id_produto ) AS num_lin_max,
                REGR_SLOPE(qtd_dessazonalizada, num_lin) OVER (PARTITION BY reg_id_produto ) slope,
                REGR_INTERCEPT(qtd_dessazonalizada, num_lin)  OVER (PARTITION BY reg_id_produto ) intercept
              FROM (
                SELECT reg_id_produto, reg_id_estabelecimento, num_lin, mes, trimestre, quantidade,media_trimestre,
media_geral, media_trimestre / media_geral AS factor_sazonal,
                    quantidade / ( media_trimestre / media_geral) qtd_dessazonalizada
                FROM (
                    SELECT dados.reg_id_produto, reg_id_estabelecimento, num_lin, mes, trimestre, quantidade,
                        AVG(quantidade) OVER (PARTITION BY dados.reg_id_produto, trimestre ) media_trimestre,
                        AVG(quantidade) OVER (PARTITION BY dados.reg_id_produto ) media_geral
                    FROM (
                        SELECT
                            p.reg_id reg_id_produto,
                            f.ANO_MES mes,
                            f.reg_id_estabelecimento,
                            TO_CHAR(to_date(f.ANO_MES,'yyyymm'),'Q') trimestre,
                            ROW_NUMBER()
                               OVER(PARTITION BY p.reg_id ORDER BY f.ANO_MES NULLS LAST) AS num_lin,
                            SUM(NVL(f.quant_pedida, 0)) quantidade
                        FROM
                            OCP_FACTURACAO_CONNECT_est f,
arm.arm_produtos@EVOL_PORTALOCP.CELESIOGROUP.COM P
                        WHERE
                            to_date(f.ANO_MES,'yyyymm') BETWEEN TRUNC(LAST_DAY(ADD_MONTHS(SYSDATE, -1 *
(wk_meses_vendas))), 'MM')
                            AND TRUNC(LAST_DAY(ADD_MONTHS(SYSDATE, -1)))
                            AND f.reg_id_estabelecimento = p_reg_id_estabelecimento
                            AND p.reg_id = p_reg_id_produto
                            AND f.produto_id = p.produto_id
                            and NVL(f.quant_pedida, 0)>0
                        GROUP BY
                            p.reg_id, f.ANO_MES, f.reg_id_estabelecimento, TO_CHAR(to_date(f.ANO_MES,'yyyymm'),'Q')
                        ORDER BY
                            p.reg_id, f.ANO_MES, f.reg_id_estabelecimento
                       )  dados
                 ) sazonal
                ORDER BY num_lin
              )
            ) fore ON fore.trimestre = cat.trimestre_fore;
        ---
    BEGIN
        ---
```

```
                wk_meses_vendas  := 24;
                wk_num_meses_forecast  := 3;
                ---
                -- obtem o ano e mes a partir do qual vai calcular a previsão. Serve para eliminar os dados da tabela de forecast
                wk_ano := TO_CHAR(TRUNC(LAST_DAY(ADD_MONTHS(SYSDATE, -1))+1),'yyyy');
                wk_mes := TO_CHAR(TRUNC(LAST_DAY(ADD_MONTHS(SYSDATE, -1))+1),'mm');
                ---
                IF wk_erro IS NULL THEN
                 --
                   FOR r_estab IN c_estab LOOP
                   FOR r_prod IN c_prod LOOP
                      ---
                      wk_reg_id_estabelecimento := r_estab.reg_id_estabelecimento;
                      wk_reg_id_produto := r_prod.reg_id_produto;
                      ---
                      FOR r_fore IN c_fore (r_estab.reg_id_estabelecimento, r_prod.reg_id_produto)
                      LOOP
                         ---
                         wk_conta  := wk_conta + 1;
                         ---
                         BEGIN
                            insert into ocp_maya_vendas_for
                               (ANO_MES, ESTAB_ID, REG_ID_PRODUTO, RL_FOR1)
                            values
                               (TO_CHAR(r_fore.DATA,'yyyy')||TO_CHAR(r_fore.DATA,'mm'), r_estab.reg_id_estabelecimento,
r_fore.reg_id_produto, r_fore.qtd_forecast);
                            ---
                            COMMIT;
                            ---
                         EXCEPTION WHEN OTHERS THEN
                            update ocp_maya_vendas_for
                            set RL_FOR1 = r_fore.qtd_forecast
                            where ANO_MES=TO_CHAR(r_fore.DATA,'yyyy')||TO_CHAR(r_fore.DATA,'mm')
                            and ESTAB_ID = r_estab.reg_id_estabelecimento
                            and REG_ID_PRODUTO = r_fore.reg_id_produto;
                            ---
                            COMMIT;
                         END;
                         ---
                         COMMIT;
                         ---
                      END LOOP;
                      ---
                      COMMIT;
                      ---
                   END LOOP;
                   COMMIT;
                   END LOOP;
                   ---
```

```
   END IF;
   ---
EXCEPTION WHEN others THEN
  wk_erro := SUBSTR(SQLERRM,1,222);
  arm_mail2_pkg.mail (sender => 'olga',
                      recipients => 'augusto.ribeiro@ocp.pt',
                      subject => 'pr_calc_for_rl',
                      MESSAGE => wk_reg_id_estabelecimento|| ' - '||wk_reg_id_produto|| ' - '||wk_erro);
END;

/*****************************************************************************/
/*****************************************************************************/


END;
/
```

Apêndice 7. *Script* da criação da *Materialized View* (ARM_OCP_FACT_CLI_MV)

```
/* Criação da Materialized View agrupa facturação 31 dias, para efeitos performance */
CREATE MATERIALIZED VIEW ARM.ARM_OCP_FACT_CLI_MV
TABLESPACE TBSARMCOMPD
NOCACHE
LOGGING
NOCOMPRESS
NOPARALLEL
BUILD IMMEDIATE
REFRESH FORCE ON DEMAND
AS
select
   c.reg_id reg_id_cliente,
   c.cliente_id,
   c.nome,
   c.CONTRIB_CLASSIF||c.POTENCIAL_CLASSIF classif,
   pa.reg_id_produto,
   fep.produto_id,
   fep.DESIGNACAO_PRODUTO,
   e.reg_id reg_id_estabelecimeno,
   e.estab_id,
   e.abreviatura,
   ARM_OCPP_DW_FUNC_PKG.pesca_pro(fe.reg_id_cliente) cli_limitado,
   ARM_OCPP_DW_FUNC_PKG.fu_tipo_exp_R_NR(pa.reg_id_produto  ) tipo_exp_R_NR,
   ARM_OCPP_DW_FUNC_PKG.lista_infarmed (fep.produto_id) lista_infarmed,
   ARM_OCPP_DW_FUNC_PKG.fu_angariacao(pa.reg_id_produto  ) angariacao,

   sum(nvl(fep.quant_pedida,0) * decode(fe.sinal_factura,'D',1,-1)) pedida,
   sum(nvl(fep.quant_aviada,0) * decode(fe.sinal_factura,'D',1,-1)) aviada
from
   arm_facturas_estab FE, arm_facturas_estab_produtos fep, arm_clientes C,arm_estabelecimentos E, arm_produtos_armaze
ns PA, arm_contratos CT, arm_contratos_cli CTC
where
   fe.data_factura >= trunc(sysdate)-31
   and fe.reg_id = fep.reg_id_factura_estab
   and fe.reg_id_cliente = c.reg_id
   and ARM_OCPP_DW_FUNC_PKG.obtem_cliente_estab_princ(c.reg_id) = e.reg_id
   and PA.reg_id = FEP.reg_id_produto_armazem
   and fe.reg_id_contrato_cli = CTC.reg_id
   and CTC.reg_id_contrato = CT.reg_id
   and nvl(CT.vendas_diarias,'n')='S'
   and CT.contrato_id not in ('EXPORT','VNDCOLAB','TRANSF')
group by
   c.reg_id ,
   c.cliente_id,
   c.nome,
   c.CONTRIB_CLASSIF||c.POTENCIAL_CLASSIF ,
   pa.reg_id_produto,
   fep.produto_id,
   fep.DESIGNACAO_PRODUTO,
   e.reg_id ,
   e.estab_id,
   e.abreviatura,
   ARM_OCPP_DW_FUNC_PKG.pesca_pro(fe.reg_id_cliente) ,
   ARM_OCPP_DW_FUNC_PKG.fu_tipo_exp_R_NR(pa.reg_id_produto  ) ,
   ARM_OCPP_DW_FUNC_PKG.lista_infarmed (fep.produto_id) ,
   ARM_OCPP_DW_FUNC_PKG.fu_angariacao(pa.reg_id_produto  );
/

/* Indice para efeitos performance query calculo outliers */
CREATE INDEX ARM.ARM_OCP_FACT_CLI_MV_I1 ON ARM.ARM_OCP_FACT_CLI_MV
(reg_id_produto,classif)
LOGGING
TABLESPACE TBSARMCOMPI
NOPARALLEL;
/
```

Apêndice 8. *Script* da criação da tabela (ARM_OCP_FACT_CLI_IQR)

```
CREATE TABLE ARM.ARM_OCP_FACT_CLI_IQR
(
  REG_ID_CLIENTE              NUMBER(28)     NOT NULL,
  CLIENTE_ID                  VARCHAR2(30 BYTE) NOT NULL,
  NOME                        VARCHAR2(200 BYTE) NOT NULL,
  CLASSIF                     VARCHAR2(10 BYTE),
  REG_ID_PRODUTO              NUMBER(28)     NOT NULL,
  PRODUTO_ID                  VARCHAR2(32 BYTE) NOT NULL,
  DESIGNACAO_PRODUTO          VARCHAR2(200 BYTE) NOT NULL,
  REG_ID_ESTABELECIMENO       NUMBER(28)     NOT NULL,
  ESTAB_ID                    VARCHAR2(10 BYTE) NOT NULL,
  ABREVIATURA                 VARCHAR2(30 BYTE) NOT NULL,
  CLI_LIMITADO                VARCHAR2(4 BYTE),
  TIPO_EXP_R_NR               VARCHAR2(4 BYTE),
  LISTA_INFARMED              VARCHAR2(4 BYTE),
  ANGARIACAO                  VARCHAR2(4 BYTE),
  PEDIDA                      NUMBER,
  AVIADA                      NUMBER,
  CLASSIF_PERC_50             NUMBER,
  CLASSIF_PERC_75             NUMBER,
  CLASSIF_PERC_25             NUMBER,
  CLASSIF_IQR                 NUMBER,
  CLASSIF_NUM_CLIENTES        NUMBER,
  ESTABELECIMENTO_PERC_50     NUMBER,
  ESTABELECIMENTO_PERC_75     NUMBER,
  ESTABELECIMENTO_PERC_25     NUMBER,
  ESTABELECIMENTO_IQR         NUMBER,
  ESTABELECIMENTO_NUM_CLIENTES  NUMBER,
  CONSUMO_MEDIO_MENSAL_PROD   NUMBER,
  CONSUMO_MEDIO_MENSAL_PROD_EST  NUMBER,
  VALOR_OUTLIER               NUMBER
)
TABLESPACE TBSARMCOMPD;


CREATE INDEX ARM.ARM_OCP_FACT_CLI_IQR_I1 ON ARM.ARM_OCP_FACT_CLI_IQR
(REG_ID_PRODUTO);


CREATE INDEX ARM.ARM_OCP_FACT_CLI_IQR_I2 ON ARM.ARM_OCP_FACT_CLI_IQR
(REG_ID_CLIENTE)
TABLESPACE TBSARMCOMPI;
```

Apêndice 9. Algoritmo do processo de deteção de *outliers* (ARM_OCPOUT_PKG)

```
CREATE OR REPLACE PACKAGE ARM.ARM_OCPOUT_PKG IS
/* Projecto ANA -  Aplicação Nova Augusto - 2015 */
/*****************************************************************************/
/*****************************************************************************/
function f_cons_mens_avi(i_ano_mes  in number
                        ,  i_reg_id_produto  in number
                        , i_reg_id_estab in number
                         )return number;
/*****************************************************************************/
procedure pr_calculo_out ;
/*****************************************************************************/
procedure pr_mail_aviso;
/*****************************************************************************/
procedure pr_job_out ;
/*****************************************************************************/

END;
/
CREATE OR REPLACE PACKAGE BODY ARM.ARM_OCPOUT_PKG IS
/* Projecto ANA -  Aplicação Nova Augusto - 2015 */
/*****************************************************************************/
/*****************************************************************************/
function f_cons_mens_avi(i_ano_mes  in number
                        , i_reg_id_produto  in number
                        , i_reg_id_estab in number
                        )
                        RETURN NUMBER IS
   ---
   CURSOR c_cons_mensal_olga_estab IS
      SELECT
         ceil(sum(am.consumo_medio_mensal)) consumo
      FROM
         arm_acumulados_mensais_stk am
      WHERE
         am.ano = substr(i_ano_mes,1,4)
         AND am.mes = substr(i_ano_mes,5,2)
         AND am.reg_id_produto = i_reg_id_produto
         and am.reg_id_armazem = arm_arm_pkg.obtem_armazem_por_tipo(i_reg_id_estab,'AVI')
         and am.estab_context = i_reg_id_estab;
   ---
   CURSOR c_cons_mensal_olga_nac IS
      SELECT
         ceil(sum(am.consumo_medio_mensal)) consumo
      FROM
         arm_acumulados_mensais_stk am
      WHERE
         am.ano = substr(i_ano_mes,1,4)
         AND am.mes = substr(i_ano_mes,5,2)
         AND am.reg_id_produto = i_reg_id_produto
         and am.reg_id_armazem = arm_arm_pkg.obtem_armazem_por_tipo(REG_ID_ESTAB,'AVI');
   ---
   wk_consumo number;
   ---
BEGIN
   wk_consumo  := NULL;
   ---
   IF i_reg_id_estab is not null THEN
      OPEN c_cons_mensal_olga_estab;
      FETCH c_cons_mensal_olga_estab INTO wk_consumo;
      CLOSE c_cons_mensal_olga_estab;
   ELSE
      OPEN c_cons_mensal_olga_nac;
      FETCH c_cons_mensal_olga_nac INTO wk_consumo;
      CLOSE c_cons_mensal_olga_nac;
   END IF;
   ---
   RETURN NVL(wk_consumo, 0);
   ---
EXCEPTION WHEN OTHERS THEN
   RETURN NVL(wk_consumo, 0);
END;
```

```
/*****************************************************************************/
/*****************************************************************************/
procedure pr_calculo_out is
   ---
   wk_erro VARCHAR2(4000) := NULL;
   ---
   cursor c_dados is
      select
         x.REG_ID_CLIENTE,
         x.CLIENTE_ID,
         x.NOME,
         x.CLASSIF,
         x.REG_ID_PRODUTO,
         x.PRODUTO_ID,
         x.DESIGNACAO_PRODUTO,
         x.REG_ID_ESTABELECIMENO,
         x.ESTAB_ID,
         x.ABREVIATURA,
         x.CLI_LIMITADO,
         x.TIPO_EXP_R_NR,
         x.LISTA_INFARMED,
         x.ANGARIACAO,
         x.PEDIDA,
         x.AVIADA,
         PERCENTILE_DISC(0.50) WITHIN GROUP (ORDER BY pedida DESC)  OVER (PARTITION BY classif, x.reg_id_produ
to) classif_perc_50,
         PERCENTILE_DISC(0.75) WITHIN GROUP (ORDER BY pedida DESC)  OVER (PARTITION BY classif, x.reg_id_produ
to) classif_perc_75,
         PERCENTILE_DISC(0.25) WITHIN GROUP (ORDER BY pedida DESC)  OVER (PARTITION BY classif, x.reg_id_produ
to) classif_perc_25,
         PERCENTILE_DISC(0.25) WITHIN GROUP (ORDER BY pedida DESC)  OVER (PARTITION BY classif, x.reg_id_produ
to) - PERCENTILE_DISC(0.75) WITHINGROUP (ORDER BY pedida DESC)  OVER (PARTITION BY classif, x.reg_id_produto) classif_I
QR,
         count(1)   OVER (PARTITION BY classif, x.reg_id_produto) classif_NUM_CLIENTES,
         PERCENTILE_DISC(0.50) WITHIN GROUP (ORDER BY pedida DESC)  OVER (PARTITION BY reg_id_estabelecimen
o, x.reg_id_produto)estabelecimento_perc_50,
         PERCENTILE_DISC(0.75) WITHIN GROUP (ORDER BY pedida DESC)  OVER (PARTITION BY reg_id_estabelecimen
o, x.reg_id_produto)estabelecimento_perc_75,
         PERCENTILE_DISC(0.25) WITHIN GROUP (ORDER BY pedida DESC)  OVER (PARTITION BY reg_id_estabelecimen
o, x.reg_id_produto)estabelecimento_perc_25,
         PERCENTILE_DISC(0.25) WITHIN GROUP (ORDER BY pedida DESC)  OVER (PARTITION BY reg_id_estabelecimen
o, x.reg_id_produto) -
PERCENTILE_DISC(0.75) WITHIN GROUP (ORDER BY pedida DESC)  OVER (PARTITION BY reg_id_estabelecimeno, x.reg_id_prod
uto) estabelecimento_IQR,
         count(1)   OVER (PARTITION BY reg_id_estabelecimeno, x.reg_id_produto) estabelecimento_NUM_CLIENTES
      from
         ARM_OCP_FACT_CLI_mv x, arm_prod_distrib_mv p
      where
         x.REG_ID_PRODUTO = p.REG_ID_PRODUTO
         and estado_id='A'
         and (categoria_prod_id like '01%' or categoria_prod_id like '02%' or categoria_prod_id like '03%' or categoria_prod_id lik
e'09%')
       --        reg_id_produto in ( 5699668,35102990,500080)
        ;
   ---
   wk_CONS_MEDIO_MENSAL_PROD number;
   wk_CONS_MEDIO_MENSAL_PROD_EST number;
   wk_VALOR_OUTLIER number;
   wk_ano_mes_ant number;
   ---
BEGIN
   ---
   delete ARM_OCP_FACT_CLI_IQR;
   commit;
   ---
   wk_ano_mes_ant := to_char(add_months(sysdate,-1),'yyyymm');
   ---
   FOR rec IN c_dados
   LOOP
      ---
      wk_CONS_MEDIO_MENSAL_PROD := ARM_OCPOUT_PKG.f_cons_mens_avi(wk_ano_mes_ant, rec.reg_id_produto, n
ull);
      wk_CONS_MEDIO_MENSAL_PROD_EST := ARM_OCPOUT_PKG.f_cons_mens_avi(wk_ano_mes_ant, rec.reg_id_prod
uto, rec.REG_ID_ESTABELECIMENO);
      ---
      wk_VALOR_OUTLIER:=0;
      ---
      if rec.pedida > rec.CLASSIF_IQR * 3 then
         wk_VALOR_OUTLIER := wk_VALOR_OUTLIER + 1-(1/rec.CLASSIF_NUM_CLIENTES);
      end if;
      ---
      if rec.pedida > rec.ESTABELECIMENTO_IQR * 3 then
         wk_VALOR_OUTLIER := wk_VALOR_OUTLIER + 1-(1/rec.ESTABELECIMENTO_NUM_CLIENTES);
      end if;
      ---
```

```
            if rec.pedida > wk_CONS_MEDIO_MENSAL_PROD_EST then
               wk_VALOR_OUTLIER := wk_VALOR_OUTLIER + 1-(1/rec.ESTABELECIMENTO_NUM_CLIENTES);
            end if;
            ---
            if rec.pedida > wk_CONS_MEDIO_MENSAL_PROD then
               wk_VALOR_OUTLIER := wk_VALOR_OUTLIER + 1;
            end if;
            ---
            wk_VALOR_OUTLIER := round(wk_VALOR_OUTLIER,2);
            ---
            insert into ARM_OCP_FACT_CLI_IQR
               (REG_ID_CLIENTE,CLIENTE_ID,NOME,CLASSIF,
               REG_ID_PRODUTO,PRODUTO_ID,DESIGNACAO_PRODUTO,
               REG_ID_ESTABELECIMENO,ESTAB_ID,ABREVIATURA,
               CLI_LIMITADO,TIPO_EXP_R_NR,LISTA_INFARMED,ANGARIACAO,
               PEDIDA,AVIADA,
               CLASSIF_PERC_50,CLASSIF_PERC_75,CLASSIF_PERC_25,CLASSIF_IQR,CLASSIF_NUM_CLIENTES,
               ESTABELECIMENTO_PERC_50,ESTABELECIMENTO_PERC_75,ESTABELECIMENTO_PERC_25,ESTABELECIME
NTO_IQR,ESTABELECIMENTO_NUM_CLIENTES,
               CONSUMO_MEDIO_MENSAL_PROD,CONSUMO_MEDIO_MENSAL_PROD_EST,
               VALOR_OUTLIER)
            values
               (rec.REG_ID_CLIENTE,rec.CLIENTE_ID,rec.NOME,rec.CLASSIF,
               rec.REG_ID_PRODUTO,rec.PRODUTO_ID,rec.DESIGNACAO_PRODUTO,
               rec.REG_ID_ESTABELECIMENO,rec.ESTAB_ID,rec.ABREVIATURA,
               rec.CLI_LIMITADO,rec.TIPO_EXP_R_NR,rec.LISTA_INFARMED,rec.ANGARIACAO,
               rec.PEDIDA,rec.AVIADA,
               rec.CLASSIF_PERC_50,rec.CLASSIF_PERC_75,rec.CLASSIF_PERC_25,rec.CLASSIF_IQR,rec.CLASSIF_NUM_CLIE
NTES,
               rec.ESTABELECIMENTO_PERC_50,rec.ESTABELECIMENTO_PERC_75,rec.ESTABELECIMENTO_PERC_25,rec.ES
TABELECIMENTO_IQR,rec.ESTABELECIMENTO_NUM_CLIENTES,
               wk_CONS_MEDIO_MENSAL_PROD,wk_CONS_MEDIO_MENSAL_PROD_EST,
               wk_VALOR_OUTLIER);
            ---
            commit;
            ---
         END LOOP;
         ---
      EXCEPTION WHEN others THEN
         wk_erro := SUBSTR(SQLERRM,1,222);
         arm_mail2_pkg.mail (sender => 'olga',
                     recipients => 'augusto.ribeiro@ocp.pt',
                     subject => 'pr_calculo_out',
                     MESSAGE => wk_erro);
      END;
      /*****************************************************************************/
      /*****************************************************************************/
      PROCEDURE pr_mail_aviso is

         ---
         wk_erro VARCHAR2(4000);
         wk_linha_txt VARCHAR2(32000):=NULL;
         ---
         wk_mail_to          VARCHAR2(4000) := NULL;
         wk_mail_cc          VARCHAR2(4000) := NULL;
         wk_mail_bcc         VARCHAR2(4000) := NULL;
         wk_mail_subject     VARCHAR2(4000) := NULL;
         wk_conn             UTL_SMTP.connection;
         wk_linha            NUMBER := 0;
         wk_old_tipo         VARCHAR2(1) := NULL;
         wk_dados            VARCHAR2(4000) := NULL;
         wk_data             VARCHAR2(10) := NULL;
         wk_hora             VARCHAR2(6) := NULL;
         l_boundary          VARCHAR2(255) DEFAULT 'a1b2c3d4e3f2g1';
         l_temp              VARCHAR2(32767) DEFAULT NULL;
         ---
         wk_cont number:=0;
         ---
         cursor C_produto is
            select
               temp.reg_id_produto, temp.produto_id, temp.DESIGNACAO_PRODUTO
            from
               (
               select
                  reg_id_produto, produto_id, DESIGNACAO_PRODUTO, count(1)
               from
                  ARM_OCP_FACT_CLI_IQR
               where
                  valor_outlier>0
                  and pedida>aviada
               group by
                  reg_id_produto, produto_id, DESIGNACAO_PRODUTO
               having
                  count(1)>0
               order by
```

```
                    count(1) desc
                ) temp
            where rownum <= 10;
        ---
        cursor c_produto_detalhe (i_reg_id_produto in number) is
            select
                x.produto_id,
                x.DESIGNACAO_PRODUTO ,
                p.descricao,
                x.CLASSIF,
                x.ABREVIATURA,
                x.CLI_LIMITADO,
                x.TIPO_EXP_R_NR,
                x.LISTA_INFARMED,
                x.ANGARIACAO,
                sum(x.PEDIDA) pedida,
                sum(x.AVIADA) aviada,
                sum(x.VALOR_OUTLIER) valor_outlier
            from
                ARM_OCP_FACT_CLI_IQR x, arm_prod_distrib_mv p
            where x.reg_id_produto = i_reg_id_produto
                and x.valor_outlier>0
                and x.pedida>x.aviada
                and x.reg_id_produto = p.reg_id_produto
            group by
                x.produto_id,
                x.DESIGNACAO_PRODUTO ,
                p.descricao,
                x.CLASSIF,
                x.ABREVIATURA,
                x.CLI_LIMITADO,
                x.TIPO_EXP_R_NR,
                x.LISTA_INFARMED,
                x.ANGARIACAO
            order by
                x.CLASSIF,
                x.ABREVIATURA,
                x.CLI_LIMITADO;
        ---
        cursor C_cliente is
            select
                temp.reg_id_cliente, temp.cliente_id, temp.nome, ARM_MORADAS_PKG.OBTEM_LOCALIDADE_CLIENTE(temp.reg_i
d_cliente) LOCALIDADE
            from
                (
                select
                    reg_id_cliente, cliente_id, nome, count(1)
                from
                    ARM_OCP_FACT_CLI_IQR
                where
                    (valor_outlier>0
                    and pedida>aviada)
                    OR
                    /* adicionar este cliente */
                    (cliente_id in (select cliente_id from arm_clientes where cliente_id='100105' or numero_anf='10588'))
                group by
                    reg_id_cliente, cliente_id, nome
                having
                    count(1)>0
                order by
                    count(1) desc
                ) temp
            where rownum <= 11;
        ---
        cursor c_cliente_detalhe (i_reg_id_cliente in number) is
            select
                x.produto_id,
                x.DESIGNACAO_PRODUTO ,
                p.descricao,
                x.CLASSIF,
                x.ABREVIATURA,
                x.CLI_LIMITADO,
                x.TIPO_EXP_R_NR,
                x.LISTA_INFARMED,
                x.ANGARIACAO,
                sum(x.PEDIDA) pedida,
                sum(x.AVIADA) aviada,
                sum(x.VALOR_OUTLIER) valor_outlier
            from
                ARM_OCP_FACT_CLI_IQR x, arm_prod_distrib_mv p
            where x.reg_id_cliente = i_reg_id_cliente
                and x.valor_outlier>0
                and x.pedida>x.aviada
                and x.reg_id_produto = p.reg_id_produto
            group by
```

```
                    x.produto_id,
                    x.DESIGNACAO_PRODUTO ,
                    p.descricao,
                    x.CLASSIF,
                    x.ABREVIATURA,
                    x.CLI_LIMITADO,
                    x.TIPO_EXP_R_NR,
                    x.LISTA_INFARMED,
                    x.ANGARIACAO
                  order by
                    sum(x.VALOR_OUTLIER) desc;
            begin
              --- ENVIO MAIL PRODUTOS
              FOR rec0 in C_produto
              LOOP
                ---
                wk_cont:=0;
                ---
                wk_dados :=NULL;
                wk_mail_subject := 'OUTLIER Produto '||rec0.produto_id||'-'||rec0.DESIGNACAO_PRODUTO||'-
'||TO_CHAR(SYSDATE,'dd-mm-yyyy');
                wk_mail_to := 'augusto.ribeiro@ocp.pt ';
                ---
                BEGIN
                  wk_conn  :=
                    arm.arm_mail2_pkg.begin_mail(arm_mail2_pkg.mailer_sender
                                ,wk_mail_to
                                ,wk_mail_subject
                                ,arm_mail2_pkg.multipart_mime_type
                                ,1
                                , wk_mail_cc
                                , wk_mail_bcc
                                );
                EXCEPTION
                  WHEN UTL_SMTP.transient_error OR UTL_SMTP.permanent_error THEN
                    wk_mail_subject  := NULL;
                END;
                ---
                IF wk_mail_subject IS NOT NULL THEN
                  arm.arm_mail2_pkg.attach_text(conn => wk_conn, DATA => '
', mime_type => 'text/html', INLINE => TRUE, filename => NULL, LAST => FALSE);
                  wk_dados  := '<html>
                    <HEAD>
                      <STYLE TYPE="text/css">
                      <!--
                      .datagrid TABLE { border-collapse: collapse;
                                text-align: LEFT;
                                 }
                      .datagrid {
                              font-family: Courier NEW;
                              background: #fff;
                              OVERFLOW: hidden;
                              border: NONE;
                              -webkit-border-radius: 3px;
                              -moz-border-radius: 3px;
                              border-radius: 3px; }
                      .datagrid TABLE td,
                      .datagrid TABLE th { padding: 4px 4px; }
                      .datagrid TABLE thead th {background-color:#2F662D;
                                          color:#FFFFFF;
                                          font-SIZE: 10px;
                                          font-weight: bold;
                                          border-LEFT: 1px solid #E1EEF4; }
                      .datagrid TABLE thead th:FIRST-CHILD { border: NONE; }
                      .datagrid TABLE tbody td { color: #001E2B;
                                           border-LEFT: 1px solid #E1EEF4;
                                           font-SIZE: 10px;
                                           font-weight: normal; }
                      .datagrid TABLE tbody
                      .alt td { background: #B0D9B6;
                                color: #001E2B; }
                      .altf td { background: #2F662D;
                                color:#FFFFFF;
                                font-SIZE: 10px;
                                font-weight: bold;
                                border: 1px solid #000000;
                                 }
                      .datagrid TABLE tbody td:FIRST-CHILD { border-LEFT: NONE; }
                      .datagrid TABLE tbody tr:LAST-CHILD td { border-bottom: NONE; }
                      .grayRow{
                         background: #d8d8d8;
                      }
                      --->
                      </STYLE>
                    </HEAD>
```

```
            <BODY>';

        arm.arm_mail2_pkg.write_text(wk_conn, wk_dados);
        ---
              wk_dados  :=      '<div CLASS="datagrid"><TABLE>
                    <thead><tr>
              <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">Produto</th>
              <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">Designacao</th>
              <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">Categoria</th>
              <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">CLASSIF</th>
              <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">ABREVIATURA</th>
              <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">LIMIT</th>
              <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">R_NR</th>
              <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">INFARMED</th>
              <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">ANGARIA</th>
              <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">Pedida</th>
              <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">Aviada</th>
              <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">Val_Outlier</th>
                      </tr></thead>
                    <tbody>';
              arm.arm_mail2_pkg.write_text(wk_conn, wk_dados);
        --
    END IF;
    ---
    wk_dados :=NULL;
    wk_erro  := NULL;
    wk_linha_txt := NULL;
    wk_linha:=0;
    ---
    FOR REC IN c_produto_detalhe (rec0.reg_id_produto)
    LOOP
        ---
              wk_linha := wk_linha+1;
              wk_dados       :=
                  '<tr>
                            <td nowrap>'|| rec.produto_id|| '</td>
                    <td align="left" nowrap>'|| rec.DESIGNACAO_PRODUTO|| '</td>
                    <td align="left" nowrap>'|| rec.descricao|| '</td>

                    <td align="left" nowrap>'|| rec.CLASSIF|| '</td>
                    <td align="left" nowrap>'|| rec.ABREVIATURA|| '</td>

                    <td align="left" nowrap>'|| rec.CLI_LIMITADO|| '</td>
                    <td align="left" nowrap>'|| rec.TIPO_EXP_R_NR|| '</td>

                    <td align="left" nowrap>'|| rec.LISTA_INFARMED|| '</td>
                    <td align="left" nowrap>'|| rec.ANGARIACAO|| '</td>

                    <td align="left" nowrap>'|| rec.Pedida|| '</td>
                    <td align="left" nowrap>'|| rec.Aviada|| '</td>

                    <td align="left" nowrap>'|| rec.Valor_Outlier

                  || '</tr> ';
              arm.arm_mail2_pkg.write_text(wk_conn, wk_dados);
        ---
    END LOOP; -- c in c1
    ---
    IF NVL(wk_linha,0)> 0 THEN
        wk_dados  := '</TABLE>';
        arm.arm_mail2_pkg.write_text(wk_conn, wk_dados);

            arm_mail2_pkg.write_text(wk_conn, '<BR> <BR>');
            wk_dados  := ' ';
            arm_mail2_pkg.write_text(wk_conn, wk_dados);

        wk_dados  := '</BODY>';
        arm.arm_mail2_pkg.write_text(wk_conn, wk_dados);
        wk_dados  := '</html>';
        arm.arm_mail2_pkg.write_text(wk_conn, wk_dados);
        ---

    END IF;
    ---
    arm.arm_mail2_pkg.end_mail(wk_conn);
    ---
END LOOP;
---
---
---
--- ENVIO MAIL CLIENTES
FOR rec0 in C_cliente
LOOP
    ---
    wk_cont:=0;
```

```
                    ---
                    wk_dados :=NULL;
                    wk_mail_subject := 'OUTLIER Cliente '||rec0.cliente_id||'-'||rec0.nome||' ('||rec0.localidade||')-'||TO_CHAR(SYSDATE,'dd-
mm-yyyy');
                    wk_mail_to := 'augusto.ribeiro@ocp.pt ';
                    ---
                     BEGIN
                       wk_conn  :=
                         arm.arm_mail2_pkg.begin_mail(arm_mail2_pkg.mailer_sender
                                      ,wk_mail_to
                                      ,wk_mail_subject
                                      ,arm_mail2_pkg.multipart_mime_type
                                      ,1
                                      , wk_mail_cc
                                      , wk_mail_bcc
                                      );
                     EXCEPTION
                       WHEN UTL_SMTP.transient_error OR UTL_SMTP.permanent_error THEN
                         wk_mail_subject  := NULL;
                     END;
                    ---
                    IF wk_mail_subject IS NOT NULL THEN
                       arm.arm_mail2_pkg.attach_text(conn => wk_conn, DATA => '
', mime_type => 'text/html', INLINE => TRUE, filename => NULL, LAST => FALSE);
                       wk_dados  := '<html>
                           <HEAD>
                             <STYLE TYPE="text/css">
                             <!--
                             .datagrid TABLE { border-collapse: collapse;
                                          text-align: LEFT;
                                           }
                             .datagrid {
                                       font-family: Courier NEW;
                                       background: #fff;
                                       OVERFLOW: hidden;
                                       border: NONE;
                                       -webkit-border-radius: 3px;
                                       -moz-border-radius: 3px;
                                       border-radius: 3px; }
                             .datagrid TABLE td,
                             .datagrid TABLE th { padding: 4px 4px; }
                             .datagrid TABLE thead th {background-color:#2F662D;
                                                  color:#FFFFFF;
                                                  font-SIZE: 10px;
                                                  font-weight: bold;
                                                  border-LEFT: 1px solid #E1EEF4; }
                             .datagrid TABLE thead th:FIRST-CHILD { border: NONE; }
                             .datagrid TABLE tbody td { color: #001E2B;
                                                   border-LEFT: 1px solid #E1EEF4;
                                                   font-SIZE: 10px;
                                                   font-weight: normal; }
                             .datagrid TABLE tbody
                             .alt td { background: #B0D9B6;
                                       color: #001E2B; }
                             .altf td { background: #2F662D;
                                      color:#FFFFFF;
                                      font-SIZE: 10px;
                                      font-weight: bold;
                                      border: 1px solid #000000;
                                       }
                             .datagrid TABLE tbody td:FIRST-CHILD { border-LEFT: NONE; }
                             .datagrid TABLE tbody tr:LAST-CHILD td { border-bottom: NONE; }
                             .grayRow{
                                background: #d8d8d8;
                             }
                             --->
                             </STYLE>
                           </HEAD>
                           <BODY>';

                       arm.arm_mail2_pkg.write_text(wk_conn, wk_dados);
                       ---
                             wk_dados  :=      '<div CLASS="datagrid"><TABLE>
                                       <thead><tr>
                             <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">Produto</th>
                             <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">Designacao</th>
                             <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">Categoria</th>
                             <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">CLASSIF</th>
                             <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">ABREVIATURA</th>
                             <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">LIMIT</th>
                             <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">R_NR</th>
                             <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">INFARMED</th>
                             <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">ANGARIA</th>
                             <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">Pedida</th>
                             <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">Aviada</th>
```

```
                <th align="left" bgcolor="#C0C0C0" bordercolor="#FFFFFF">Val_Outlier</th>
                    </tr></thead>
                  <tbody>';
                arm.arm_mail2_pkg.write_text(wk_conn, wk_dados);
            --
        END IF;
        ---
        wk_dados :=NULL;
        wk_erro  := NULL;
        wk_linha_txt := NULL;
        wk_linha:=0;
        ---
        FOR REC IN c_cliente_detalhe (rec0.reg_id_cliente)
        LOOP
            ---
                wk_linha := wk_linha+1;
                wk_dados       :=
                    '<tr>
                            <td nowrap>'|| rec.produto_id|| '</td>
                      <td align="left" nowrap>'|| rec.DESIGNACAO_PRODUTO|| '</td>
                      <td align="left" nowrap>'|| rec.descricao|| '</td>

                      <td align="left" nowrap>'|| rec.CLASSIF|| '</td>
                      <td align="left" nowrap>'|| rec.ABREVIATURA|| '</td>

                      <td align="left" nowrap>'|| rec.CLI_LIMITADO|| '</td>
                      <td align="left" nowrap>'|| rec.TIPO_EXP_R_NR|| '</td>

                      <td align="left" nowrap>'|| rec.LISTA_INFARMED|| '</td>
                      <td align="left" nowrap>'|| rec.ANGARIACAO|| '</td>

                      <td align="left" nowrap>'|| rec.Pedida|| '</td>
                      <td align="left" nowrap>'|| rec.Aviada|| '</td>

                      <td align="left" nowrap>'|| rec.Valor_Outlier

                    || '</tr> ';
                arm.arm_mail2_pkg.write_text(wk_conn, wk_dados);
            ---
        END LOOP; -- c in c1
        ---
        IF NVL(wk_linha,0)> 0 THEN
            wk_dados  := '</TABLE>';
            arm.arm_mail2_pkg.write_text(wk_conn, wk_dados);

              arm_mail2_pkg.write_text(wk_conn, '<BR> <BR>');
              wk_dados  := ' ';
              arm_mail2_pkg.write_text(wk_conn, wk_dados);

            wk_dados  := '</BODY>';
            arm.arm_mail2_pkg.write_text(wk_conn, wk_dados);
            wk_dados  := '</html>';
            arm.arm_mail2_pkg.write_text(wk_conn, wk_dados);
            ---

        END IF;
        ---
        arm.arm_mail2_pkg.end_mail(wk_conn);
        ---
    END LOOP;
    ---

exception when others then
        /* ENVIAR MAIL AVISO ERRO */
        BEGIN
            wk_erro  := 'ERRO:' || SUBSTR(SQLERRM, 1, 200);
            ---
              arm.arm_mail2_pkg.mail(sender      => 'OLgA@ocp.pt'
                          ,recipients  => 'augusto.ribeiro@ocp.pt'
                          ,subject     => 'pr_mail_aviso'
                          ,MESSAGE     => wk_erro
                          );
        EXCEPTION
            WHEN OTHERS THEN
              NULL;
        END;
end;
---
/*****************************************************************************/
/*****************************************************************************/
procedure pr_job_out is
    ---
    wk_erro VARCHAR2(4000) := NULL;
    ---
begin
```

```
   ---
   dbms_snapshot.refresh('ARM_OCP_FACT_CLI_MV');
   ARM_OCPOUT_PKG.pr_calculo_out ;
   ARM_OCPOUT_PKG.pr_mail_aviso;
   ---
EXCEPTION WHEN others THEN
   wk_erro := SUBSTR(SQLERRM,1,222);
   arm_mail2_pkg.mail (sender => 'olga',
                       recipients => 'augusto.ribeiro@ocp.pt',
                       subject => 'pr_job_out',
                       MESSAGE => wk_erro);
END;
/*****************************************************************************/
/*****************************************************************************/

END;
/
```