Análise Exploratória de Hierarquias em Base de Dados Multidimensionais

Nuno Ferreira,
Faculdade de Economia da Universidade do Porto,
nuno.ferreira@fe.up.pt

João Gama,
Faculdade de Economia da Universidade do Porto,
jgama@fep.up.pt

Resumo

Cada vez mais, as empresas e organizações utilizam bases de dados multidimensionais e ferramentas OLAP como forma de organizar informação proveniente de sistemas transacionais com o objetivo de aceder e analisar dados com elevada flexibilidade e desempenho. No modelo de dados OLAP, a informação é concetualmente organizada em cubos. Cada dimensão do cubo tem uma hierarquia associada, o que possibilita analisar os dados em diferentes níveis de agregação. Apresenta-se uma metodologia que explora os diferentes níveis de agregação das hierarquias para uma análise exploratória assim como previsões para diferentes horizontes temporais. Esta metodologia mostrou-se muito eficiente, apresentando melhores resultados em comparação com as técnicas usuais de previsão. Os resultados das previsões realizadas são promissores e coerentes com a respetiva análise exploratória.

palavras-chave: dimensões hierárquicas, OLAP, *dynamic time warping*, previsão

Abstract

Increasingly, companies and organizations use multidimensional databases and OLAP tools to structure and organize information from transactional systems, with the objective of accessing and analyzing data with high level of flexibility and performance. In the OLAP models, data is conceptually organized into cubes. Each cube's dimension has an associated hierarchy, which allows for data analysis at different levels of aggregation. A methodology is presented, which explores the different levels of aggregation of the hierarchies for an exploratory analysis as well as the forecasts for different time horizons. This methodology proved to be very efficient, with better results than those obtained from the usual techniques of forecasting. The forecasting results are promising and in line with the respective exploratory analysis.

keywords: hierarchical dimensions, OLAP, dynamic time warping, prediction

1 Introdução

O funcionamento normal de empresas, entidades e instituições, envolve, cada vez mais, um maior volume de dados e a maioria destas entidades já possuem uma base de dados multidimensional com suporte OLAP. O rápido desenvolvimento da tecnologia OLAP levou à necessidade de uma capacidade de análise de dados mais sofisticada, como a previsão, tendências e deteção de exceções [Chen, et al., 2002].

O intuito deste trabalho é apresentar um método que analise simultaneamente várias séries temporais hierarquicamente ligadas, aproveitando as hierarquias criadas por peritos do domínio e com experiência do negócio. Realizamos uma análise exploratória sobre as séries temporais no sentido de identificar tendências e correlações entre séries hierarquicamente ligadas. O estudo realizado proporciona uma forma de prever os próximos elementos da série temporal, usando uma agregação das previsões das séries descendentes na hierarquia associada a essa dimensão.

As questões que investigamos vão para além da pergunta, "*Can we predict the future by looking into the past?*" [Makridakis, et al., 1998]. Será possível, para além de olhar para o passado, usar os dados históricos dos descendentes hierárquicos, que a compõem, para prever essa mesma série? Será que é possível prever a informação de um dado indicador referente ao país, se olhar para os indicadores de cada distrito/cidade/localidade? Os resultados obtidos demonstram que é possível usar a informação dos descendentes hierárquicos.

O documento está organizado da seguinte forma: no próximo capítulo vai ser descrito o trabalho relacionado; no terceiro capítulo vai ser apresentada a metodologia seguida para a análise exploratória e para a elaboração das previsões; no quarto capítulo são apresentadas experiências realizadas para teste à metodologia apresentada e uma análise dos resultados desses testes; no quinto capítulo são apresentadas as conclusões do trabalho realizado e trabalhos futuros; no sexto capítulo é apresentada a bibliografia utilizada.

2 Trabalho Relacionado

O trabalho realizado no âmbito deste estudo é transversal a algumas áreas de estudo. Serão retratados o estado da arte nas áreas: cubos OLAP e o cálculo de distâncias entre séries.

2.1 Online Analytical Processing (OLAP)

A esmagadora maioria das empresas usa sistemas de base de dados para a manutenção das operações do dia-a-dia. Estes sistemas são denominados *Online Transaction Processing* (OLTP). O uso massivo levou a um grande desenvolvimento neste tipo de sistemas e uma explosão do tamanho das bases de dados durante as últimas décadas [Ramakrishnan, et al., 2000]. Com o uso intensivo veio uma degradação do desempenho em consultas com dados agregados, esta levou ao aparecimento das *Datawarehouses* (DW) e as respetivas ferramentas de análise [Chen, et al., 2002].

Para a implementação dos cubos OLAP há a necessidade de consolidação, integração de dados, passando, muitas vezes, pela deteção e limpeza de erros dos mesmos [Chen, et al., 2002]. O OLTP tem demasiado detalhe para suporte à decisão, ao invés o DW sumariza e agrega com diferentes granularidades, oferecendo consultas úteis aos diferentes níveis organizacionais [Ramakrishnan, et al., 2000]. As DW têm outra propriedade importante, para além dos dados, estes sistemas têm um repositório de metadados. Este repositório guarda a informação sobre a estrutura (como dimensões, níveis, hierarquias, origens de dados, etc.), tipos de dados dos campos, termos de negócio e operadores de agregação.

Os sistemas que são caracterizados por consultas, que tipicamente envolvem agrupamentos, agregações e funcionalidades para análises de séries temporais, são denominados *Online Analytic Processing* (OLAP) [Ramakrishnan, et al., 2000]. As consultas nestes sistemas são análises multidimensionais, para permitir estas consultas todas as métricas têm de ter representação em todas as dimensões [Han, et al., 2006]. Os dados são definidos por um tabela de factos, esta tabela é constituída pelas métricas e a representação, obrigatória, de cada entrada, em cada dimensão do universo do cubo. Usualmente, cada dimensão é composta por vários níveis de abstração ou granularidade definidos e ordenados num conceito de hierarquias [Ramakrishnan, et al., 2007]. Atualmente, as ferramentas OLAP permitem que o cubo seja composto por mais do que uma tabela de factos, permitindo um leque de utilização ainda mais alargado. Por exemplo, criar estruturas de relações de muitos para muitos entre dimensões.

As DW são criadas especificamente para a estrutura, as políticas e processos de cada empresa. Para a construção destas DW, é necessário ter os dados numa forma desnormalizada, para que a tabela de factos consiga comportar as ligações para as tabelas de dimensões, ao contrário das bases de dados tradicionais que devem estar na terceira forma normal. Esta característica das DW faz com que o número de relações entre tabelas necessárias decresça, o que se traduz numa diminuição do tempo de resposta numa consulta, mas uma maior necessidade de espaço em disco. Devido ao elevado número de registos do um DW típico, o tempo de resposta é uma característica muito importante.

Os pontos de utilização mais comuns são melhorar o *focus* no cliente, reposicionamento de produto, manutenção de *portfolios*, análise de operações, gestão de clientes e recursos empresariais [Ramakrishnan, et al., 2000].

As ferramentas OLAP permitem aos executivos organizar, entender e tomar decisões estratégicas com base nos dados da empresa. Toda a informação passa a ficar acessível a todos em tempo útil, de uma forma clara e em vários níveis de agregação [Han, et al., 2006]. Permitindo aos utilizadores finais fazer um grande leque de consultas orientadas para o negócio de uma forma fácil, rápida e intuitiva, intituladas consultas *ad-hoc*, a diferentes granularidades, sem ajuda ou intervenção de um programador [Ramakrishnan, et al., 2000].

Para consultas as DW oferecem operações que são gerais a qualquer implementação:

- ***roll-up*** – Quando se está num nível e pretende-se subir para o nível do ascendente. Ou seja retirar detalhe (agregar dados) na análise.
- ***drill-down*** – Quando se está num nível e pretende-se saber informação de um nível descendente, obter mais detalhe.
- **slicing** – Fazer um "Corte" no cubo numa dada dimensão, isto é, na prática, aplicar um filtro a uma dimensão.
- ***dicing*** – O *dicing* é a capacidade de juntar os vários cortes (*slices*) de forma a obter uma consulta com filtros diversos.
- ***Pivot*** – Numa consulta trocar as dimensões num dos eixos com as dimensões no outro eixo. Por exemplo, numa consulta *ad-hoc*, trocar as dimensões que estão em colunas pelas dimensões que estão em linhas.

Na área de extração de conhecimento (ECD) com recurso a sistemas OLAP, alguns dos trabalhos mais significativos passam pela criação de novos cubos para previsão, combinando tendências nas várias dimensões do cubo. Ferramentas OLAP são utilizadas em [Sarawagi, et al., 1998] e [Han, et al., 2006] para assinalar exceções guiando o utilizador para os tuplos de maior interesse. [Chen, et al., 2006b] utiliza cubos para previsão de tendências e deteção de exceções em modelos de regressão. [Palpanas, et al., 2005] aplica o uso do OLAP para determinar os desvios de cada descendente e qual dos descendentes apresenta o maior desvio. [Rabaséda, et al., 2011] utiliza sistemas OLAP para a previsão de uma métrica de um acontecimento futuro respondendo à questão "e se…". [Imielinski, et al., 2002] usa OLAP com regras de associação para a obtenção de novas regras. De realçar o trabalho de [Chen, et al., 2006a], que usa sistemas OLAP para prever valores de elementos agregados em função de elementos descendentes que mais se assemelhem, usando árvores de decisão como modelo base. Por exemplo, para prever os valores das vendas de um produto nos Estados Unidos, usa o valor das vendas de uma cidade cujo histórico é o mais semelhante ao do país.

2.2 Semelhança entre séries temporais - *Dynamic Time warping* (DTW)

Os dados das empresas, instituições ou organização têm um enquadramento temporal, por exemplo, uma fatura tem, obrigatoriamente, uma data. O uso da dimensão tempo possibilita o uso de ferramentas e técnicas de análise de séries temporais no estudo dos dados de empresas [Berndt, et al., 1994].

Uma das técnicas de análise de séries temporais consiste no estudo da semelhança entre séries. O cálculo da semelhança é feito através do cálculo de distâncias entre séries. Entre os modelos mais comuns, temos a distância euclidiana, a *Discrete Fourier Transformation*, a *Discrete Wavelet Transformation* e o *Dynamic Time Warping* (DTW) [Niels, 2004]. Alguns investigadores afirmam que o DTW é a melhor métrica na maioria dos domínios [Rakthanmanon, et al., 2012]. Uma característica que torna o DTW apetecível é a interpretabilidade dos resultados, pois as comparações de curvas apresentam resultados que fazem sentido [Niels, 2004].

A distância euclidiana é largamente utilizada e divulgada. Consiste em comparar todos os pontos da série com o ponto correspondente (com o mesmo índice) da série base. Para cada índice é calculada a distância euclidiana. A distância entre as duas séries é dada pela soma das distâncias de todos os índices. O painel superior da Ilustração 1. Este método é rápido e tem um crescimento temporal linear, ao invés o DTW tem um

crescimento quadrático. O método da distância euclidiana é sensível a distorções, compressões, a desfasamentos e obriga a que as duas séries tenham o mesmo comprimento [Keogh, et al., 2000]. O DTW é majorado pela distância euclidiana.

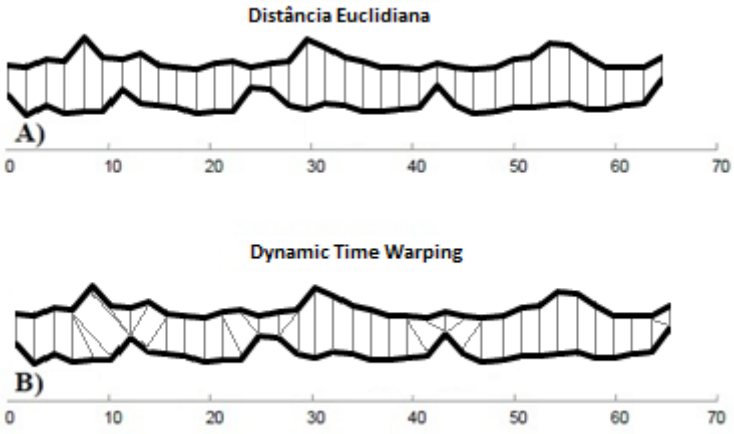

Ilustração 1 – Diferença entre os modelos *DTW* e Distância euclidiana

2.2.1 O Algoritmo DTW

A técnica DTW usa uma programação dinâmica que alinha, estica e comprime as séries de forma a obter uma medida de distância menor possível [Berndt, et al., 1994] [Sakoe, et al., 1978].

Para alinhar duas séries que possam estar desfasadas temporalmente ou com uma compressão diferente, sem estar predisposto a variações ou atraso em toda ou parte da série, foi apresentado, em [Sakoe, et al., 1978], um método de programação dinâmica na área de processamento de fala. Sakoe e Chiba apresentam o termo "*warp*", e a forma de o calcular em programação dinâmica. Sakoe e Chiba referiam-se a este método como *DP-matching* (*Dynamic programming Matching*) [Sakoe, et al., 1978].

Mais tarde, em [Berndt, et al., 1994], é introduzida a técnica de DTW na área de ECD. No mesmo trabalho onde se apresentou a utilidade desta técnica, também foi mostrado o seu problema de performance, o crescimento quadrático do número de operações com o crescimento do comprimento da série.

Especificamente, consiste em calcular a distância dos valores da série *S*, com os valores da série *T*:

$$S = s_1, s_2, \dots, s_n \qquad T = t_1, t_2, \dots, t_m \tag{1}$$

As séries são organizadas numa matriz de distâncias com uma dimensão de n*m, onde cada célula dessa matriz corresponde à combinação do par (s_i, t_j), sendo *i* e *j* o índice das séries *S* e *T*, respetivamente. O caminho *W* é constituído pelas combinações dos pontos das séries *S* e *T* que minimizam a distância.

$$W = w_1, w_2, \dots, w_k \qquad onde \max(n, m) \leq k < n + m - 1 \tag{2}$$

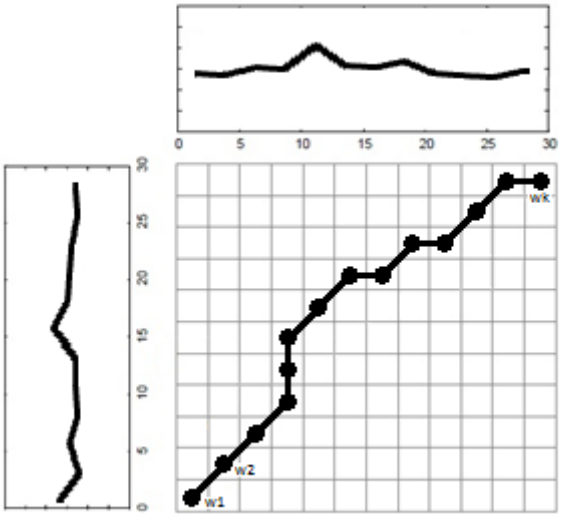

Ilustração 2 – Exemplo de uma matriz de *warp*

Para determinar cada ponto da série *W* é necessário definir a métrica da distância a aplicar, na fórmula apresentada, a distância euclidiana. Tendo *n* como o número de métricas para o cálculo da distância.

$$\delta(i,j) = \sqrt{\sum_{i=1}^{n} \left(s_i - t_j\right)^2} \quad \textbf{(3)}$$

$$\Rightarrow dado\ que\ n = 1$$

$$\Leftrightarrow \delta(i,j) = \sqrt{\left(s_i - t_j\right)^2}, logo$$

$$\Leftrightarrow\ \delta(i,j) = \left|s_i - t_j\right| \quad \textbf{(4)}$$

Para casos em que só há uma métrica a avaliar, pode usar-se a versão da distância euclidiana mais simples (Equação (4)), caso contrária será necessário usar a fórmula original (Equação (3)).

Podemos definir o problema como uma minimização de um caminho potencial com base no valor acumulado da distância, desde o ponto inicial.

$$DTW(S,T) = min_w \left[\sum_{k=1}^{p} \delta(w_k)\right] \quad \textbf{(5)}$$

Na equação (5), *p* representa o comprimento da série/caminho *W* em questão.

Para melhorar a eficiência e o tempo de cálculo do algoritmo, foram apresentados alguns pressupostos que, não sendo obrigatórios, representam boas práticas [Berndt, et al., 1994] [Sakoe, et al., 1978].

Este algoritmo tem sido usado nas mais diversas áreas, tais como, análise da similaridade de músicas [Muller, et al., 2006], análise de caracteres escritos à mão [Niels, 2004], extração de conhecimento [Keogh, et al., 2000].

Alguns das evoluções ao algoritmo de DTW estudados foram o *Multiscale* DTW de Muller [Muller, et al., 2006], método iterativo que em cada iteração, vai-se pormenorizando mais. Permite ainda a separação em várias subséries, reduzindo o problema da dimensionalidade; FastDTW [Salvador, et al., 2007], parte de um princípio de iteração, mas com base numa janela teórica, onde as células escolhidas, dependem do raio de ação definido. Entre outras derivações, ainda se realçam *Piecewise* DTW [Keogh, et al., 2000], *Iterative Multiscale* DTW [Zinke, et al., 2006] e *Derivated* DTW [Keogh, et al., 2001].

3 Metodologia

Neste trabalho é apresentada uma forma de analisar dados estruturados hierarquicamente, da série de referência e séries descendentes. Além da análise exploratória propõe-se prever valores de alguns períodos de tempo.

3.1 Análise exploratória

A análise exploratória consiste na elaboração do cronograma da série de referência, esta representação permite verificar a tendência de evolução da série, identificar a sazonalidade, conhecer o perfil (*layout*) da série e aferir sobre a estacionaridade. Além do cronograma, gera-se a reta da regressão linear. Esta possibilita confirmar o crescimento/retração médio do valor da métrica da série a analisar.

Em seguida aferem-se as autocorrelações da série. Este teste indicará se o valor de um período está ou não relacionado com outro período, qual o fator dessa relação e se é possível usar os dados históricos na previsão. Se a série não apresentar correlações, o uso dos dados históricos não é benéfico, pois esses valores não terão qualquer influência nos valores dos períodos futuro.

Para caracterizar a relação entre a série de referência e as séries descendentes, *drill-down* numa determinada dimensão, representa-se graficamente o cronograma das séries com os valores normalizados. Os valores devem estar normalizados pois a análise ao perfil é o objetivo. Para normalizar seguiu-se a equação (6).

$$Z = \frac{X - \mu}{\sigma} \quad \textbf{(6)}$$

X representa o valor a normalizar, μ é o valor da média da série e σ corresponde ao desvio padrão da série.

Com as séries normalizadas, representa-se um cronograma onde se pode observar, no mesmo período, como se comporta cada série descendente perante a série de referência. Trata-se de uma observação importante para verificar qual dos descendentes é mais semelhante à série de referência. No caso de existência de desfasamentos, compressões ou expansões da série, exige um espírito crítico por parte do interveniente, para

determinar se se trata de uma diferença real e repetível ou de apenas um pico ou depressão, de origem esporádica.

Esta análise permite entender melhor o negócio, em especial sazonalidades e o panorama da evolução da atividade ao longo de um período temporal.

3.2 Análise preditiva

As experiências realizadas no âmbito da análise preditiva das séries vão prever dados usando as hierarquias. Para a avaliação do modelo preditivo apresentado neste trabalho foi realizada uma comparação com os métodos de previsão *baseline* e ARIMA.

Antes da previsão, é necessário determinar o número de elementos históricos que definem cada subsérie, este depende da sazonalidade dos dados. A série é separada em subséries como exemplificada na Ilustração 3.

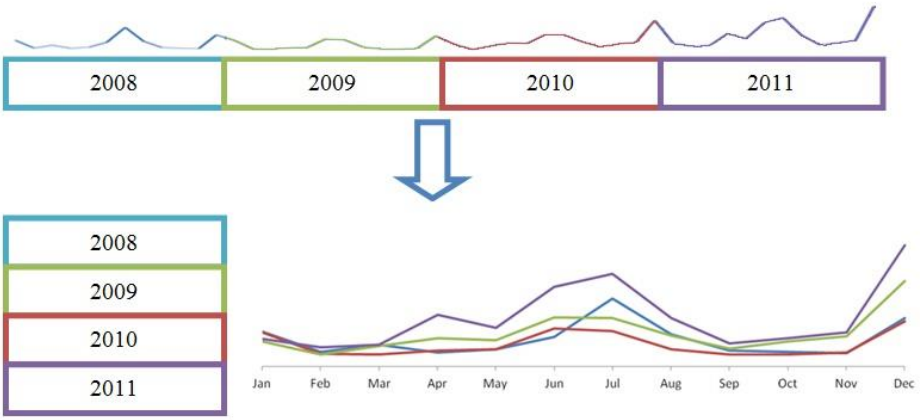

Ilustração 3 - Conversão da série em várias séries mais curtas de período definido

Para cada série, vai ser criada uma nova subsérie adicional que consiste na soma de todas as subséries, excetuando a subsérie a comparar, no exemplo usado neste trabalho, a subsérie referente ao ano de 2011. Esta agregação habilita-nos com mais uma série na altura da comparação, com a particularidade de ter os picos e depressões atenuados (Ilustração 4).

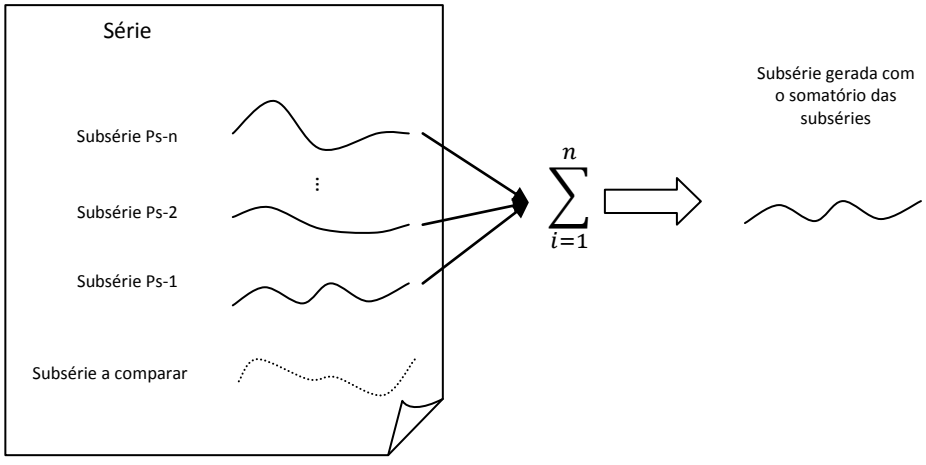

Ilustração 4 – Série adicional

3.2.1 Soma das previsões das séries descendentes

Este método de previsão consiste em prever um número de períodos para cada série descendente hierárquica da série de referência, e somar todas as previsões obtidas para determinar qual o valor previsto para a série de referência. Por exemplo, considerando a série <Mês, Todas Cores> correspondente ao par ordenado das dimensões <Tempo, Cores> a série ascendente contém todas as cores e as séries descendentes contêm as cores individuais: branco, preto, etc.

Este método consiste em calcular, para todas as séries descendentes, a semelhança entre as subséries descendentes históricas e a subsérie descendentes correspondente ao período de tempo a prever. No caso de dados mensais e uma previsão dos 3 últimos meses de 2011, os dados seriam comparados com os 9 primeiros meses de 2011 (Ilustração 5). Após a previsão de todos os valores das séries descendentes, as previsões devem ser somadas para se obter o valor da previsão da série de referência.

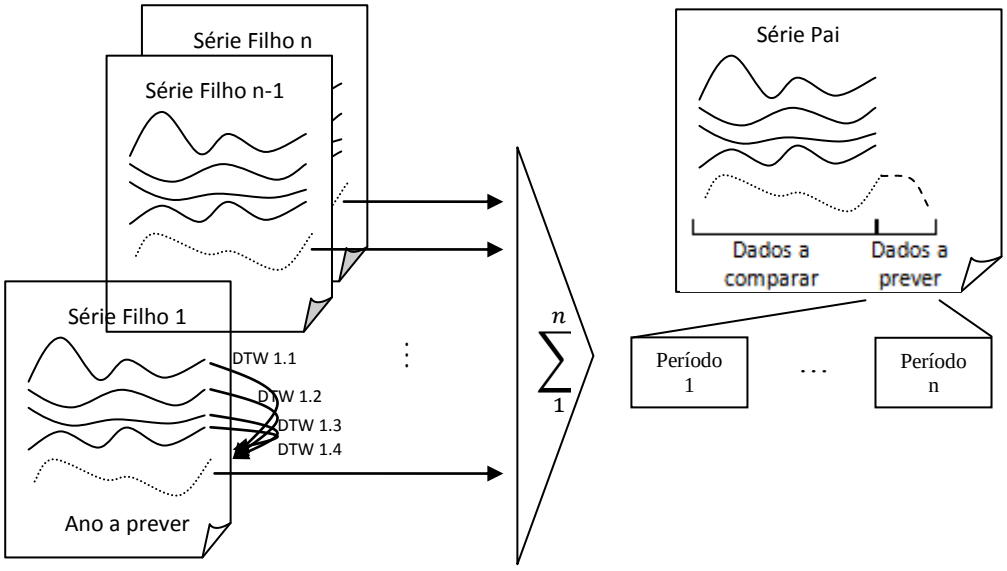

Ilustração 5 – Previsão de valores com base na soma das séries descendentes

mês a prever e do mês anterior (Equação (7)). Após o cálculo da previsão em todas as séries descendentes, agregam-se as previsões, por período, para obter os valores previstos da série de referência (Equação (8)).

$$V_{prever\ x} = \frac{V_{semelhante\ x}}{V_{semelhante\ x-1}} * V_{prever\ x-1} \quad \textbf{(7)}$$

Sendo "V prever" o valor da série do ano em que queremos prever, "V semelhante" o valor da subsérie mais semelhante, *x* o período a prever, e *x-1* o período anterior. Para previsões superiores a um período, o valor a prever do período anterior é a previsão previamente realizada.

$$F\_ref_{t+p} = \sum_{i=1}^{n} F\_desc(i)_{t+p} \quad \textbf{(8)}$$

Sendo *F_ref* e *F_desc* os valores de previsões das séries de referência e descendentes, respetivamente. Onde *t* último período de tempo conhecido, *p* o número de períodos a prever e n o número de séries descendentes.

De seguida, por uma questão de validade e quantificação da qualidade da previsão, compara-se o valor previsto com o valor real (valor não usado na previsão), ou seja, determina-se o erro da previsão.

A métrica de erro usada será o erro absoluto relativo médio (Equação (9)). Usado com vista a penalizar, igualmente, erros de previsão excessiva como negativa e não cancelar um erro positivo com um negativo.

$$\varepsilon = \frac{1}{n}\sum_{p=1}^{n}\left|\frac{A_{t+p} - F\_ref_{t+p}}{A_{t+p}}\right| \quad \textbf{(9)}$$

A_{t+p} representa o valor real num período de tempo $t+p$, F_ref_{t+p} o valor previsto para o período de tempo $t+p$ e n o número de períodos previstos.

3.2.2 Métodos comparativos

O método comparativo *baseline* é o método de previsão mais simples possível. Consiste em tomar os valores visionados no período sazonal anterior, no exemplo apresentado, os valores dos últimos três meses do ano 2010, e usá-los como previsão.

A implementação do ARIMA requer um conjunto de parâmetros elevado, aplicou-se o método da análise do valor da variância das diferenças das séries, tanto sazonais como do período anterior. O método permite, com facilidade, chegar aos valores dos parâmetros, usualmente denominados por d e D (os parâmetros das diferenças). Neste ponto o utilizador já deve conhecer a sazonalidade da série.

Através do método apresentado no trabalho [Hyndman, et al., 2008], iterativamente testam-se combinações de parâmetros usando a métrica do AIC (*Akaike Information Criterion*) como avaliador, esta deverá ser o menor possível.

Para a criação das várias combinações da série usando o ARIMA, utilizou-se código em R, e duas funções desenvolvidas por especialistas o SARIMA e o SARIMA.FOR [Shumway, et al.]. A primeira função define o modelo ARIMA e permite avaliar o valor do modelo gerado através do AIC. A segunda função serve para fazer as previsões dos valores usando o modelo escolhido.

4 Experiências Realizadas

4.1 Descrição dos dados

Os dados usados para o teste desta metodologia têm origem na informação recolhida do *Enterprise Resource Planning* (ERP) de uma empresa da área do retalho de vestuário e moda. Os dados estão guardados numa base de dados relacional, posteriormente, limpos, transformados e carregados para um sistema com suporte OLAP. O processo de limpeza e transformação segue um esquema como o apresentado na Ilustração 6.

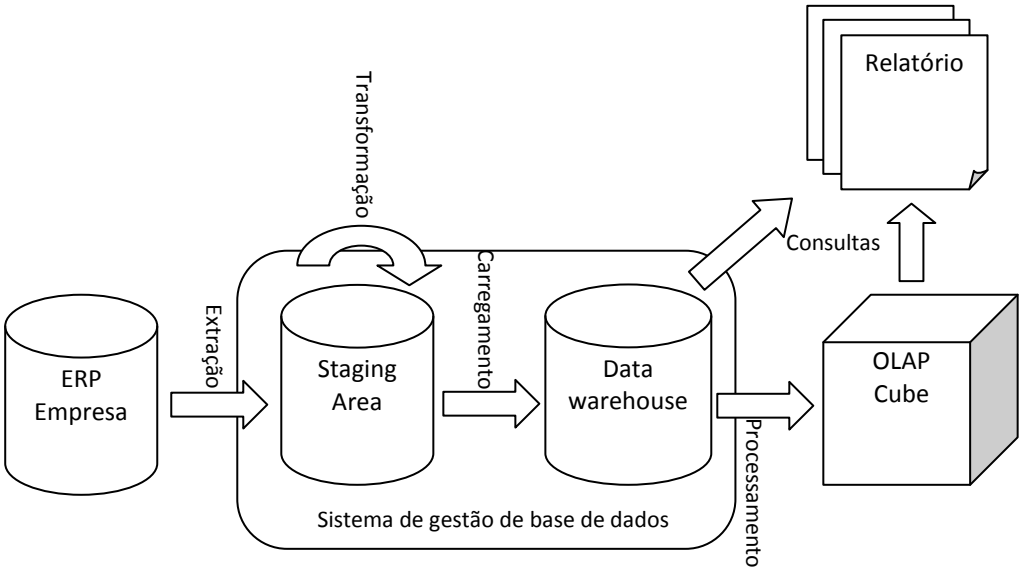

Ilustração 6 – Esquema de transformação dos dados para OLAP

O primeiro passo consiste em extrair os dados para o sistema de apoio, com duas bases de dados, a *Staging area* e a *Data warehouse*. A *Staging area* é uma base de dados de apoio para o processamento e transformação dos dados, adaptando-os à estrutura de dimensões definida pelos peritos do domínio. Seguidamente, é carregada a base de dados *Data warehouse*, esta tem como finalidade alimentar os cubos OLAP e alguns relatórios.

O trabalho proposto vai ser realizado usando os dados do cubo OLAP referente aos dados das vendas das lojas desde 2008 até ao final de 2011, sendo que os três últimos meses de 2011 serão o conjunto de validação da qualidade do método de previsão apresentado. Mais tarde serão usados dados referentes a 2012 para verificar a distribuição do erro de previsão.

4.2 Análise exploratória

Como exemplo ilustrativo analisamos a série de referência definida pelo conjunto <Meses, Todas cores> como representação do par ordenado das dimensões <tempo, cores>.

O primeiro passo corresponde à representação da série através de um cronograma (Ilustração 7), auxiliado por outro cronograma com a série referida e as séries descendentes (Ilustração 8).

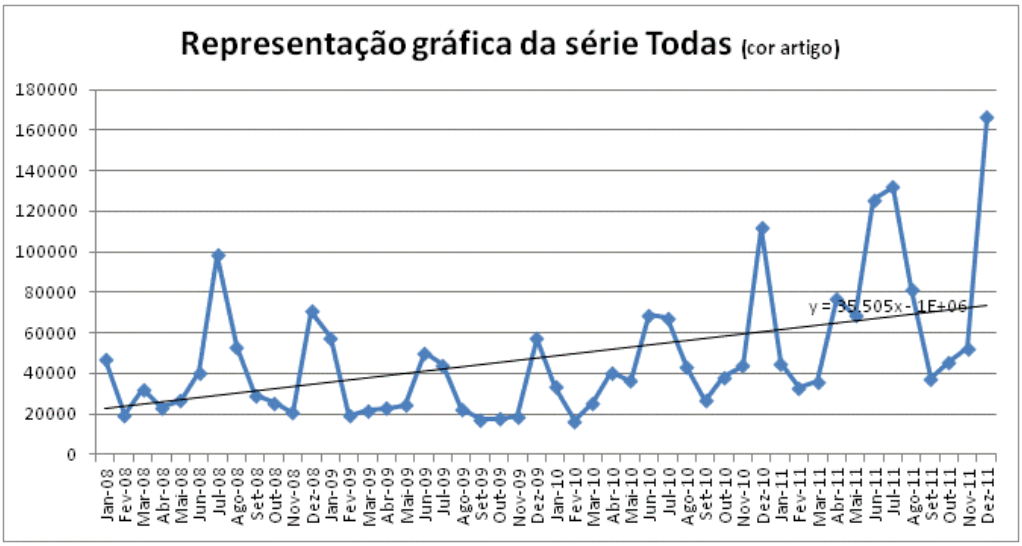

Ilustração 7 – Gráfico da série e reta da regressão linear

No gráfico podemos observar uma sazonalidade vincada com picos constantes em julho e dezembro de todos os anos, uma tendência crescente desde janeiro de 2010 (caso se retire a sazonalidade), assim como um aumento na variância ao longo do tempo. A tendência crescente dos valores dos dados da série é reforçada pela fórmula da reta da tendência linear. Em média, a tendência linear apresentaria resultados satisfatórios, mas pouco informativa acerca do consumo em cada mês. Como se pode depreender pelo gráfico e pela tendência da série, esta não é estacionária (Ilustração 7).

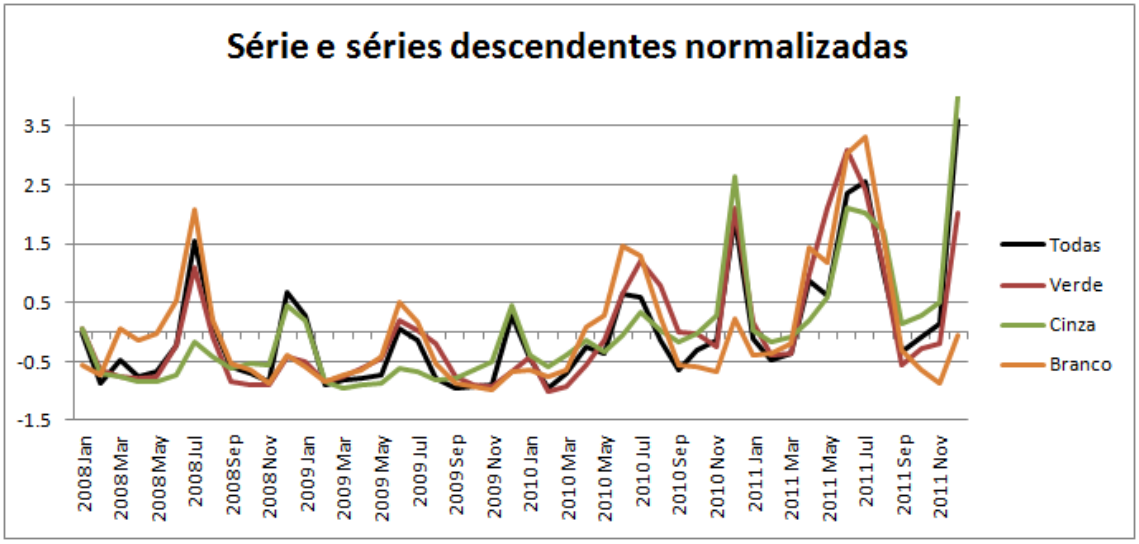

Ilustração 8 - Série e séries descendentes com os valores normalizados

Observando as séries, a maior semelhança é dada pela série Verde, mas não significa que seja a mais capaz para prever os períodos futuros. Para efeitos de análise e de apresentação de um gráfico mais legível, serão representadas, após normalização, só algumas das séries descendentes (Ilustração 8): Verde, Cinza e Branco.

O cálculo da distância entre séries foi obtido através do algoritmo DTW com os resultados apresentados (Tabela 1).

Tabela 1 - Valores da distância usando DTW

Série descendente	Distância DTW
Verde	11,90
Cinza	14,69
Vermelho	14,71
Preto	15,57
Rosa/Lilás	15,77
Castanho	18,30
Azul	19,08
Salmão/Laranja	20,42
Crú/Bege	21,03
Branco	21,23
Amarelo	21,86
Outra	32,04

Os valores representados indicam o valor da distância entre séries, no caso entre as séries descendentes e a série ascendente.

No cronograma normalizado verifica-se que as séries descendentes já não seguem a série de referência de uma forma consistente (Ilustração 8). Como se pode verificar, em julho de 2010, as séries Branco e Verde têm valores mais elevados que o valor da série de referência. Em dezembro de 2010, a série referente à cor Branca tem um valor que, apesar do ter crescimento, é muito baixo quando comparado com as restantes séries. Existe a possibilidade de uma série ser exatamente igual nos primeiros anos, mas depois ser totalmente diferente, fazendo com que esta análise possa falhar. Nota-se que a série descendente Branco não apresenta dois picos vincados por ano: apresenta um só pico no mês de julho, indicando que esta cor não vende bem no inverno, mas sim no Verão.

Seguidamente, apresenta-se um gráfico com quatro autocorrelogramas sobrepostos. As barras representam a autocorrelação da série de referência e as linhas representam os autocorrelogramas de cada série descendente.

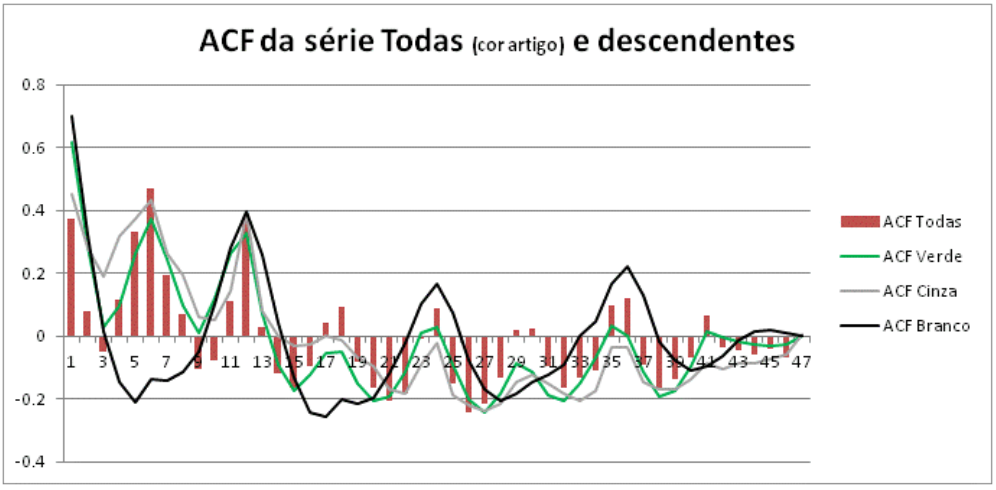

Ilustração 9 - Autocorrelações da série de referência e das séries descendentes

O correlograma apresenta uma sazonalidade semestral vincada (Ilustração 9), nomeadamente nos meses com uma correlação negativa. Esta correlação semestral surge devido aos períodos de coleções dos artigos e inerentes ao negócio instituído. A indústria do retalho de vestuário apresenta duas coleções anuais, Primavera-Verão e

Outono-Inverno, que seguem uma distribuição semelhante, pois há, oficialmente, dois períodos de fim de estação, os saldos e natal. O gráfico mostra flutuações sazonais com oscilações periódicas e não estacionário.

O correlograma mostra também um comportamento bem distinto das séries descendentes em relação à série ascendente (Ilustração 9). A autocorrelação da série ACF Verde tem, nos primeiros anos, uma curva semelhante à série pai, mas com o passar do tempo tende a distanciar-se. A particularidade mais interessante neste gráfico é a frequência apresentada pela série ACF Branco, que é metade da frequência das outras séries: apresenta uma sazonalidade anual e não semestral como as outras.

4.3 Previsão

O teste consiste em prever os últimos 3 meses do ano de 2011 usando o método da soma das previsões das séries descendentes. Os resultados serão comparados com as previsões do ARIMA e do método *baseline*. A série de referência usada para a previsão de valores é a mesma sobre a qual foi realizada a análise exploratória.

A previsão dos valores da série de referência é a soma das previsões das séries descendentes. Neste caso é necessário fazer a previsão para cada uma das cores. Aplicou-se o DTW para determinar o valor das diferenças entre séries. Para cada cor calculou-se a diferença da subsérie de cada ano com a do ano de 2011, resultando nos valores apresentados na Tabela 2.

Tabela 2 – Matriz das dissemelhanças

Série	Preto	Branco	Cinza	Bege	Cast.	Amar.	Laranja	Verm.	Rosa
adicional	**3,22**	1,64	**2,05**	**2,34**	5,31	**2,73**	3,59	**3,00**	2,02
2008	4,16	3,18	2,63	3,31	**3,74**	3,87	3,38	2,99	2,76
2009	3,59	2,56	4,92	4,17	5,36	3,37	**2,00**	4,81	1,87
2010	3,86	**1,24**	4,00	4,40	6,36	3,08	4,13	2,55	**1,63**

Para a previsão de cada série descendente será usada a subsérie descendente mais semelhante à sua subsérie do ano 2011. Para a previsão referente à série cor Preto é usada a subsérie adicional, para a série referente à cor Branco é usada a subsérie de 2010 e as restantes estão assinaladas na Tabela 2 com os valores a negrito.

Somando os valores previstos por cor, obtemos a previsão para 2011 da série Todas da dimensão cor artigo (Tabela 3).

Tabela 3 - Resultados da previsão da soma dos descendentes

	Set	Out	Nov	Dez	Erro Médio	Erro acumulado
Previsão		51.861,8	60.981,5	158.209,9		271.053,2
Série Todas 2011	37.476	45.587	52.468	166.303		264.358
Erro absoluto		**13,76%**	**16,22%**	**4,87%**	**11,62%**	**2,53%**

Os valores previstos pelo método *baseline* são os mesmos valores de 2010 para os períodos analisados, outubro, novembro e dezembro.

Os valores previstos pelo modelo ARIMA, usando os parâmetros (0,1,1,1,1,2,12), são apresentados na Tabela 4.

Tabela 4 - Resultados dos métodos comparativos

	Set	**Out**	**Nov**	**Dez**	**Erro Médio**	**Erro acumulado**
Série Todas 2011	37.476	45.587	52.468	166.303		264.358
baseline		38.288	43.744	111.726		193.758
Erro absoluto *baseline*		**16,01%**	**16,63%**	**32,82%**	**21.82%**	**26,71%**
ARIMA (0,1,1,1,1,2,12)		66.370,0	70.463,4	145.735,8		282.569,1
Erro absoluto ARIMA		**45,59%**	**34,30%**	**12,37%**	**30,75%**	**6,89%**

Para uma comparação mais fácil da capacidade preditiva dos métodos, criou-se um quadro resumo com os erros médios e acumulados (Tabela 5).

Tabela 5 - Erros de previsão

	Erro Médio	**Erro acumulado 3M**
Soma das previsões das séries descendentes	**11,62%**	**2,53%**
Erro absoluto *Baseline*	**21.82%**	**26,71%**
Erro absoluto ARIMA	**30,75%**	**6,89%**

O erro médio do método de previsão através da soma das previsões dos descendentes apresenta-se como o mais preciso, inclusive na previsão acumulada.

Para suportar estes resultados foram testadas mais duas séries de referência. As séries são definidas pelos pares ordenados <Meses, Todos Géneros> e <Meses, Masculino>, correspondente à estrutura de dimensões <Tempo, Sexo artigo> (Tabela 6).

Tabela 6 – Erros de previsão, séries adicionais

	<Meses, Todos>		**<Meses, Masculino>**	
	Erro Médio	**Erro acumulado em 3M**	**Erro Médio**	**Erro acumulado em 3M**
Soma das previsões das séries descendentes	**11,28%**	**4,44%**	**8,89%**	**6,92%**
Erro absoluto *Baseline*	**22,29%**	**27,54%**	**17.54%**	**23,49%**
Erro absoluto ARIMA	**26,43**	**1,73%**	**38.57%**	**27,84%**

Como se pode observar o erro médio do método apresentado é o menor em todos os casos apresentados.

Para determinar a distribuição do erro do método realizou-se um último teste com a série definida por <Semanas, Todos>, da estrutura de dimensões definida por <Tempo, Sexo artigo>.

Tabela 7 – Tabela das frequências

Erro	Freq.
<=-0,9	**3**
]-0,9; -0,7]	**0**
]-0,7; -0,5]	**1**
]-0,5; -0,3]	**4**
]-0,3; -0,1]	**4**
]-0,1; 0,1]	**17**
]0,1; 0,3]	**22**
]0,3; 0,5]	**9**
]0,5; 0,7]	**0**
]0,7; 0,9]	**0**
0,9>	**0**

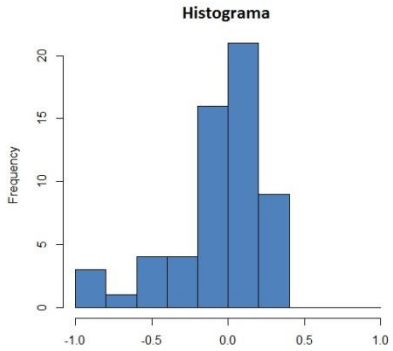

Ilustração 10 – Histograma dos erros de previsão

A distribuição apresenta-se enviesada à direita, mostrando que não segue uma distribuição normal. Para testar a normalidade da distribuição foi realizado o teste *Saphiro Wilk's*, este obteve o *p_value* de 2,95e-10, ou seja o teste da hipótese nula é rejeitado.

5 Conclusões e Trabalhos Futuros

Neste trabalho desenvolveu-se uma metodologia para a análise exploratória e previsão de uma série usando diferentes granularidades dos dados.

A análise exploratória consistiu em analisar os cronogramas com linhas de tendências para obter a sazonalidade e tendência da série de referência e as suas séries descendentes. Nos cronogramas, onde estão presentes a série de referência e séries descendentes, aferimos a relação entre as séries de diferentes granularidades. Adicionalmente, realizou-se a análise das correlações aferida pela informação dos correlogramas da série de referência e descendentes. A análise exploratória, usando DTW, permitiu definir qual das séries descendentes são as mais semelhantes à série de referência. Também se procedeu à previsão de valores usando as previsões das séries descendentes, agregando-se esses valores. Para aferir a capacidade do método de previsão compararam-se os resultados obtidos com os resultados dos métodos *baseline* e do ARIMA.

O método de previsão apresentado mostra uma boa eficácia na previsão, quando comparados com o método de *baseline* e com o ARIMA. Este método de previsão apresenta, por norma, um erro absoluto médio inferior aos outros métodos.

O problema da falta de vários valores das séries foi amenizado com a criação de uma subsérie adicional. Em situações como a Páscoa, correspondentes a picos flutuáveis no período de sazonalidade, o erro tende a sair fora da gama de valores normais. Prevêem-se dados com base numa subsérie com um pico de vendas, aumentando o valor previsto que deveria corresponder a um pico que já ocorreu, ou ainda estará para ocorrer. Esta situação é bem visível nos testes feitos à série com dados semanais, onde o efeito da Páscoa é mais notório.

Para o conjunto de dados utilizado, a metodologia testada para a previsão de dados usando as séries descendentes mostrou uma boa capacidade de previsão, excluindo algumas situações já identificadas. Além de se tratar de um conceito simples, apresenta, mesmo assim, valores com uma taxa de erro inferior à dos métodos de controlo.

O trabalho apresentado permite, por exemplo, que as empresas de retalho possam prever o consumo e com isso controlar melhor a compra de artigo, gerir os *stocks* e controlar promoções. Entender as relações entre diferentes níveis hierárquicos ajuda a entender o artigo que se vende e o segmento de mercado.

Apesar do trabalho realizado ser promissor, surgem ainda alguns pontos que podem a ser aperfeiçoados.

Como trabalhos futuros, destaca-se a fórmula da previsão de valores que poderá usar os valores das autocorrelações parciais para determinar o peso dos períodos antecessores, ao invés de ser o peso total no período anterior [Kittler, 1998].

Outra possível melhoria seria a avaliação da semelhança de séries com base em mais do que uma métrica. No caso apresentado, a métrica adicional a ter em conta poderia ser a quantidade de peças em *stock*, visto que este número pode influenciar o número de peças vendidas.

BIBLIOGRAFIA

Berndt, D. e Clifford, J. (1994), "Using dynamic time warping to find patterns in time series", *AAAI-94 Workshop on Knowledge Discovery in Databases (KDD-94)*, pp. 359-370, Seattle, Washington.

Chen, B.-C., Ramakrishnan, R., Shavlik, J. e Tamma, P. (2006a), "Bellwether Analysis: Predicting Global Aggregates from Local Regions", *VLDB '06 Proceedings of the 32nd international conference on Very large data bases*, pp. 655-666, Seul, Coreia: ACM Press.

Chen, Y., Dong, G., Han, J., Pei, J., Wah, B. e Wang, J. (2006b), "Regression Cubes with lossless compression and aggregation". *IEEE Trans. Knowledge and Data Engineering* , Vol. 18, Nº 12, pp. 1585-1599.

Chen, Y., Dong, G., Han, J., Wah, B. e Wang, J. (2002), "Multi-Dimensional Regression Analysis of Time-Series Data Streams", *Proceedings of the 28th VLDB conference*, pp. 323-334, Hong Kong, China.

Han, J. e Kamber, M. (2006), *Data Mining Concepts and Techniques* (2ª ed.), Morgan Kaufmann Publishes.

Hyndman, R. e Khandakar, Y. (2008), "Automatic time series forecasting: the forecast package for R". *Journal of Statistical Software* , Vol. 27, Nº 3.

Imielinski, T., Khachiyan, L. e Abdulghani, A. (2002). "Cubegrades: Generalizing Association Rules", *Data Mining and Knowledge Discovery*, Vol. 6, pp. 219-258, Netherlands: Kluwer Academic Publishers.

Keogh, E. e Pazzani, M. (2000), "Scaling up dynamic time warping for datamining applications", *6th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, pp. 285-289, Boston: ACM.

Keogh, E. e Pazzani, M. (2001), "Derivative Dynamic Time Warping", *In Proc. of the First SIAM International Conference on Data Mining* , Vol. 1, Chicago, IL USA: SIAM.

Kittler, J. (1998), "Combining classifiers: A theoretical framework", *Pattern Analysis & Applications*, Vol. 1, Nº 1, pp. 18-27.

Makridakis, S., Wheelwright, S. e Hyndman, R. (1998). *Forecasting: Methods and Applications* (3ª ed.), John Wiley & Sons, Inc.

Muller, M., Mattes, H. e Kurth, F. (2006), "An efficient multiscale approach to audio synchronization", *In Proceedings of the 6th International Conference on Music Information Retrieval*, pp. 192-197.

Niels, R. (2004), "Dynamic Time Warping - An intuitive way of handwriting recognition", Tese de Mestrado, Radboud University Nijmegen, Department of Artificial Intelligence.

Palpanas, T., Koudas, K. e Mendelzon, A. (2005), "Using Datacube Aggregates for Approximate Querying and Deviation Detection", *IEEE Transactions on Knowledge and Data Engineering* , Vol. 17, Nº 11, pp. 1465-1477.

Rabaséda, S., Boussaid, O., Niemczuk, A. e Messaoud, R. (2011), "Prédiction dans les cubes de données OLAP", *Conférence Méditerranéenne sur l'Ingénierie Sure des Systèmes Complexes*, Agadir, Marrocos.

Rakthanmanon, T., Campana, B., Mueen, A., Batista, G., Westover, B., Zu, Q., et al. (2012), "Searching and Mining Trillions of Time Series Subsequences under Dynamic Time warping", *The 18th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 262-270, Pequim, China: ACM.

Ramakrishnan, R. e Chen, B.-C. (2007), "Exploratory mining in cube space", *Data mining and Knowledge Discovery* , Vol. 15, 29-54.

Ramakrishnan, R. e Gehrke, J. (2000). *Database Management Systems* (2ª ed.). McGraw-Hill.

Sakoe, H. e Chiba, S. (1978), "Dynamic programming algorithm optimization for spoken word recognition", *Trans. Acoustics, Speech and Signal Proc. ASSP-26*, pp. 159-165, IEEE.

Salvador, S. e Chan, P. (2007), "FastDTW: Toward Accurate Dynamic Time Warping in Linear Time and Space", *Intelligent Data Analysis* , Vol. 11, Nº 5, pp. 561-580.

Sarawagi, S., Agrawal, R. e Megido, N. (1998), "Discovery-driven Exploration of OLAP Data Cubes", *Proceedings of the 6th International Conference on Extending Database Technology: Advances in Database Technology*, pp. 168-182, Valência, Espanha: Springer-Verlang.

Shumway, R. e Stoffer, D. (s.d.), "Time Series Analysis and Its Applications: With R Examples", http://www.stat.pitt.edu/stoffer/tsa2, acedido em 26 de Janeiro de 2012.

Zinke, A. e Mayer, D. (2006), "Iterative Multi Scale Dynamic Time Warping", *Computer Graphics Technical Reports - CG-2006/1* ,Vol. 1.

Nuno Ferreira, licenciado em Engenharia Informática e Computação pela Faculdade de Engenharia da Universidade do Porto, em Setembro de 2003. Desempenhou as funções de analista programador na *Infineon Technologies*. De Junho de 2005 a Maio de 2009 foi consultor, na área de *Business Intelligence*, na Novabase. Desde aquela data, desempenha o cargo de *IT Specialist* na Cofemel - Sociedade de Vestuário. Em Maio de 2007 obteve a certificação em *Sql Server 2005*. Em Dezembro de 2012, terminou o Mestrado em Análise de Dados e Sistemas de Apoio à Decisão, na Faculdade de Economia da Universidade do Porto.

João Gama é Professor Associado da Universidade do Porto e investigador no LIAAD / INESC TEC, trabalhando no grupo de Extração de Conhecimento de Dados. O seu principal interesse de investigação é a descoberta de conhecimento a partir de fluxos de dados. Foi co-presidente do ECML 2005, DS09, ADMA09 e de uma série de Workshops em KDDS e Descoberta Conhecimento de dados de sensores no ACM SIGKDD. É autor de um livro recente sobre Descoberta de Conhecimento de Fluxos de Dados.