

Exploração de Séries Temporais em Processos de Previsão de Vendas

Hélder Borges

helderdanielborges@gmail.com

Orlando Belo

obelo@di.uminho.pt

Centro de I&D ALGORITMI

Departamento de Informática, Escola de Engenharia, Universidade do Minho

Campus de Gualtar, 4710-057 Braga, PORTUGAL

Resumo

Em qualquer meio empresarial as vendas são uma das áreas de atividade mais críticas. A utilização de técnicas de apoio a decisão para aumentar o volume de vendas é algo a que os agentes de decisão das empresas recorrem frequentemente. A previsão de vendas é uma dessas técnicas, talvez mesmo uma das mais utilizadas. Os dados das vendas de uma empresa podem ser vistos como séries temporais, o que torna possível a aplicação de técnicas de mineração de dados em cenários típicos de previsão. Usualmente, os modelos de previsão são desenvolvidos especificamente para cada empresa em particular, requerendo a elaboração de modelos e aplicação de técnicas de mineração que acompanhem e satisfaçam tanto os seus requisitos operacionais como de suporte à decisão. Este artigo foca-se na exploração de modelos de mineração de dados baseados nos algoritmos ARIMA e ARTXP num cenário específico de vendas empresarial. Os resultados que foram obtidos permitem concluir que é possível realizar uma previsão de vendas com um erro percentual absoluto médio aceitável, para uma situação em que se prefira a redução dos custos de adaptação de modelos em relação à melhoria do desempenho da previsão através de modelos específicos.

Palavras-chave: Mineração de Dados, Previsão de Vendas, Séries Temporais, ARIMA, ARTXP e Redes Neurais.

1 Introdução

As séries temporais são sequências de dados ordenados cronologicamente, nas quais um determinado valor é, em termos temporais, posterior ao seu antecedente e anterior ao seu sucessor na série. Este tipo de dados é amplamente usado em muitos domínios de trabalho. Ratanamahatana *et al.* (2005) referem que entre 1974 e 1989 uma amostra de 4000 gráficos de 15 jornais de todo o mundo apresentava 75% dos seus gráficos em forma de séries temporais. Este formato de dados é bastante útil para armazenar informação que implique uma contextualização temporal, como por exemplo, os resultados de um electrocardiograma, uma relação de temperaturas de um determinado local, ou as métricas de negócio de uma dada empresa, durante um dado período de tempo.

As principais características das séries temporais são a sua grande dimensionalidade, o elevado espaço de armazenamento que requerem e o facto de serem de fácil interpretação. Estas características fazem com que este tipo de dados seja cada vez mais usado em aplicações empresariais, simplesmente devido ao seu interesse prático, às suas potencialidades de transmissão de tendências e de padrões e porque, atualmente, dada a evolução que as plataformas computacionais sofreram ao longo dos anos, as empresas têm já as capacidades adequadas para o registo, processamento e apresentação de tais volumes e formatos de dados. Num mundo tão competitivo como o de hoje, as empresas necessitam cada vez mais de ter ao seu dispor ferramentas que permitam acrescentar valor ao seu negócio e marcar a diferença perante os seus concorrentes. As de previsão são, talvez, um dos tipos de ferramentas que podem trazer tais vantagens para o seio de uma organização. Para os agentes de decisão empresariais, como os investidores, a existência e a disponibilidade de tais ferramentas de suporte à decisão, muitas delas já incorporando algumas capacidades preditivas, podem contribuir para a definição de bases mais sólidas para a definição de cenários de previsão que possam sustentar processos de tomada de decisão mais efetivos e que possam ajudar para o estabelecimento de bons processos de negócio [Tsaih *et al.*, 1998]. Trabalhos como [Lin *et al.*, 2011] ou [Valipour *et al.*, 2012], entre outros, são claros testemunhos de uma boa aplicação das técnicas de previsão em vários domínios de trabalho, como a saúde ou a segurança pública.

No presente artigo exploramos um processo de previsão utilizando séries temporais com o objetivo de estudar o seu comportamento num cenário típico de previsão de vendas de um ambiente empresarial. Nas secções seguintes descreveremos o trabalho desenvolvido, abordando, em particular, a aplicação de séries temporais para previsão de vendas, analisando as suas limitações e vantagens em processos de mineração de dados com este tipo de aplicações de previsão (secção 3). De seguida, na secção 4, apresentamos o nosso caso de estudo, descrevendo e caracterizando o processo de mineração de dados levado a cabo, com a análise e discussão dos resultados obtidos nas previsões efetuadas no caso em questão. Por fim, apresentamos as principais conclusões extraídas da realização deste trabalho.

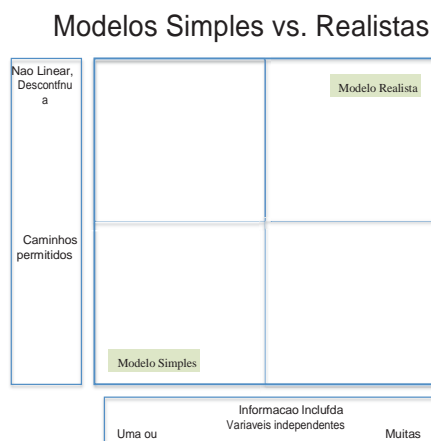


Figura 1 – Representação da complexidade do modelo de previsão – figura extraída e adaptada de [Chereb, 1998]

2 Previsão de Vendas com Séries Temporais

As técnicas de previsão podem ser divididas em dois grupos distintos. Os métodos qualitativos que estabelecem cenários e paralelismo com situações semelhantes, dispensando dados quantificados, e os métodos quantitativos que se baseiam em dados históricos e que permitem realizar uma extrapolação dos padrões identificados para o futuro. Por sua vez, estes últimos podem ainda ser divididos em métodos não causais, que assentam exclusivamente na análise dos valores históricos da variável a prever, sendo um dos modelos mais utilizados o modelo ARIMA; e em métodos causais que procuram relacionar a variável a prever com outras variáveis que possam explicar o seu comportamento, recorrendo para isso a técnicas de regressão [Makridakis, 1998].

Chereb (1998) demonstrou que se for pretendido obter um modelo de previsão realista devem ser utilizadas muitas variáveis independentes (figura 1), defendendo assim a utilização de métodos causais para realizar previsões mais realistas. Contudo, na previsão de vendas devem ser considerados não só os dados históricos, bem como dados históricos de variáveis que possam influenciar as vendas. Existem dois tipos de variáveis que podem influenciar as vendas de uma organização, nomeadamente as variáveis:

- Internas, que dizem respeito a processos internos, como o número de produção de novos artigos, o número de vendedores, ou o número de filiais, entre outros.
- Externas, que têm a ver com fatores externos que podem influenciar as vendas das empresas através da sua influência em elementos da cadeia de negócio da organização, e que podem ser, por exemplo o ciclo económico, o produto interno bruto (PIB) ou as taxas de exportação.

Entre as variáveis a considerar num cenário de previsão, apenas devem ser consideradas variáveis que sejam causais em relação à variável a prever. Isto significa que, uma alteração das variáveis auxiliares implique consequentemente uma alteração da variável a prever. Assim, há a necessidade de se utilizar técnicas capazes de manipular diversas variáveis ao mesmo tempo e que consigam, também, estabelecer regressões entre elas para que se possa fazer o cálculo das previsões. As técnicas de mineração de dados atualmente disponíveis são uma possibilidade para realizar tais processo de previsão de vendas, uma vez que alguns dos algoritmos utilizados nesse domínio permitem realizar tais regressões, bem como fazer o manuseamento de grandes volumes de dados (um aspeto bastante relevante no caso que estudamos).

A experiência prática da aplicação de técnicas de mineração de dados pode trazer grandes vantagens competitivas às empresas [Bose e Mahapatra, 2001]. Para qualquer agente de decisão empresarial será uma mais-valia a utilização de ferramentas que lhe permitam suportar algum tipo de previsão através de uma abordagem mais realista aos problemas e sustentada por métodos e processos de cálculo fiáveis. Usualmente, a informação de

vendas de uma empresa encontra-se associada a um dado período de tempo. Tal caracteriza uma sequência de vendas ao longo do tempo: uma série temporal de vendas. As séries temporais [Box *et al.*, 2008] podem ser definidas através da seguinte fórmula:

$$\mathbf{f} = (t_1, \dots, t_n), t_i \in \mathbf{f}$$

que basicamente determina que, para uma dada variável t_i , o valor t_{i-1} corresponde ao instante de tempo imediatamente anterior, e a variável t_{i+1} ao instante de tempo posterior.

As séries temporais têm vindo a ser utilizadas em diversos tipos de tarefas como a previsão [Song e Li, 2008], a segmentação [Shatkay e Zdonik, 1996], a detecção de anomalias [Keogh e Kasetty, 2002] ou a classificação [Geurts, 2001]. O registo do total de vendas mensal por parte das empresas de retalho pode ser visto, naturalmente, como uma série temporal de vendas. Porém, no caso desta série temporal ser referente, por exemplo, a um conjunto de totais mensais de vendas, a série assume uma dimensão de dados invulgarmente pequena. Neste caso em particular, esta característica permite que as técnicas de indexação e os métodos de representação, bem como as medidas de similaridade se tornem irrelevantes, uma vez que o volume de dados é mais pequeno do que o expectável em relação a outras séries temporais. No caso que trabalhamos, o pré-processamento de dados assumiu um papel importante, uma vez que foram utilizadas outras variáveis por forma a auxiliar na tarefa de previsão.

A previsão de vendas sob o formato de séries temporais tem sido amplamente estudada. Basicamente, esta consiste no cálculo dos possíveis valores para os futuros k períodos de tempo, podendo ser expressa da seguinte maneira:

$$\mathbf{f} = (t_1, \dots, t_n, t_{n+1}, \dots, t_{n+k})$$

Na literatura do domínio podemos encontrar inúmeros registos sobre a aplicação de diversas técnicas com o intuito de fazer a previsão de vendas utilizando séries temporais. Frequentemente, as técnicas de previsão são comparadas entre si, variando o seu resultado com as características específicas de cada problema sobre o qual são aplicadas, tal como esperado. Entre as referências das diversas técnicas aplicadas encontram-se as *support vector machines* (SVM), as redes neuronais, o raciocínio baseado em casos (CBR), a segmentação (*clustering*), as árvores de decisão e as *radial basis functions* (RBF). Existem diversos modelos de previsão baseados em auto regressão. ARIMA [Box e Jenkins, 1970], ou ASTAR [Lewis *et al.*, 1994] são apenas alguns exemplos desse tipo de modelos. O modelo *Auto Regressive Tree* apresentado em [Meek *et al.*, 2002] tem a particularidade de adaptar modelos auto regressivos a árvores. Nesse trabalho, os autores apresentam um modelo ART (*auto regressive tree*) e um modelo ART(p) (*autoregressive tree model of lenght p*), nos quais utilizam uma árvore de decisão, que possui modelos auto-regressivos nas folhas, para limitar a abrangência do modelo. Este modelo pode suportar um processo de regressão linear, se a árvore for restringida a apenas um ramo.

3 Um Caso Concreto de Utilização de Séries Temporais

3.1 Contexto aplicacional e caracterização de variáveis

Para a realização deste trabalho foram escolhidas três empresas com atividades empresariais bastante heterogêneas. Por questões de confidencialidade, as empresas selecionadas estão aqui designadas, simplesmente, por A, B e C. A empresa A é uma empresa de comércio por grosso de equipamentos electrónicos e de telecomunicação, que possui um volume de negócios entre 2.000.001€ a 10.000.000€. Esta é uma empresa que tanto desenvolve as suas atividades na prestação de serviços de instalação e suporte de sistemas de alarme, como na revenda de equipamentos. A análise da série temporal obtida para esta empresa não apresentou qualquer sazonalidade nas vendas, o que reflete o tipo de empresa que é. Por sua vez, a empresa B, com um volume de negócio entre 10.000.001€ e 50.000.000€, desenvolve as suas atividades no domínio das agências de viagem. Na análise dos seus dados foi detetada sazonalidade nas vendas que foram realizadas. Este facto, já esperado, revela que as viagens tendem a ser realizadas nos mesmos períodos em todos os anos que passam. De referir que, esta empresa apenas se foca na venda de serviços. Por fim temos a empresa C que desenvolve as suas atividades empresariais no fabrico de artigos para uso doméstico de faiança, de porcelana e de grés, gerando anualmente um volume de negócios entre 10.000.001€ a 50.000.000€. Esta última, é uma empresa de retalho especializada em porcelanas, dedicando-se a um tipo de comércio um pouco diferente das duas empresas anteriores, uma vez que se dedica a produção de produtos que não é afetada por qualquer tipo de sazonalidade. As variáveis internas e externas que consideramos foram variáveis com características causais sobre as vendas, isto é, variáveis cuja flutuação influencia a variação das vendas. A identificação destas variáveis foi realizada juntamente com alguns especialistas das áreas de gestão de produto e *marketing* de uma empresa produtora de soluções ERP. As variáveis internas consideradas foram o número de novos itens por mês, o número de novos clientes por mês, o número de sucursais e o número de vendedores. Como variáveis externas, foram utilizadas o PIB, o índice de confiança dos consumidores, a taxa de exportações, o investimento em obras privadas e o ciclo da economia.

3.2 Análise e preparação dos dados

Na tabela 1 podemos observar alguma informação sobre as variáveis utilizadas, vendo a sua caracterização em termos de tipo, da fonte de onde são extraídas, da necessidade ou não de se preparar estes dados para ser utilizados e uma breve descrição sobre si.

Tabela 1 - Descrição dos dados utilizados pelos algoritmos para previsão de vendas

Vendas	Bigint	ERP	Normalização sob a fórmula: $\frac{x - \min}{\max - \min}^*$	Totais mensais da empresa
Novos Produtos	Int	ERP	Não necessário	Lançados pela empresa no presente mês
Novos Clientes	Int	ERP	Não necessário	Adicionados pela empresa neste mês
Número de vendidos	Int	ERP	Não necessário	A atuar no presente mês
Número de filiais	Int	ERP	Não necessário	Em funcionamento no presente mês
P.I.B.	Int	Banco de Portugal	Não necessário	Variação em relação ao mês anterior
Ciclo Económico	Int	Banco de Portugal	Não necessário	Variação em relação ao mês anterior
Índice de confiança do Consumidor	Int	Banco de Portugal	Não necessário	Variação em relação ao mês anterior
Investimento Privado na Construção	Int	Banco de Portugal	Não necessário	Variação em relação ao mês
Taxa de exportação	Int	Banco de Portugal	Não necessário	Variação em relação ao mês

* Obs.: max e min representam o valor máximo e o valor mínimo da série temporal.

Os tipos de dados são referentes aos tipos de dados com que as variáveis foram utilizadas no *software*. O tipo de fonte “ERP” deve-se ao facto de alguns dos dados serem retirados diretamente do *software* ERP utilizado por cada uma das empresas em questão.

3.3 Implementação do modelo

Uma vez realizada a recolha de dados passámos à fase de implementação dos modelos propriamente ditos. Numa primeira parte, e após ter sido escolhido o software *SQL Server Analysis Services* [SQL Server, 2013] para as tarefas de mineração de dados a levar a cabo, foram escolhidos os algoritmos *Microsoft Time Series* e *Microsoft Neural Networks* para a realização dos processos de previsão. Esta escolha deveu-se ao facto do algoritmo *Microsoft Time Series* permitir a implementação de um modelo ARIMA, de um modelo ARTXP e de um modelo misto, enquanto que o algoritmo *Microsoft Neural Networks* permite a implementação de redes neuronais com diversos parâmetros. Com esta escolha é possível avaliar o desempenho de um modelo estatístico ARIMA, um dos modelos tradicionalmente mais usados, de um modelo ARTXP, que se perfilou como uma alternativa para analisar séries temporais, e de um modelo baseado em redes neuronais, modelo amplamente usado na bibliografia para tratar dados de previsão baseada em séries temporais.

Os algoritmos foram testados sob diferentes combinações de variáveis e de parâmetros. As combinações de variáveis que considerámos para cada um dos modelos desenvolvidos estão apresentadas na tabela 2. O modelo baseado em redes neuronais realiza previsões estabelecendo relações matemáticas entre todas as suas variáveis. Desta forma, para realizar previsões para um período k o modelo necessitou de ter variáveis referentes a esse período para que fosse possível estabelecer relações entre as variáveis e apresentar uma previsão. Como o objetivo era o de prever as vendas que pudessem ser realizadas no futuro, seria necessário estimar as vendas desse período k para que as estimativas de vendas permitam utilizar uma variável de vendas como fator a ter em consideração nas previsões futuras. Por isso, considerámos que não seria correto utilizar unicamente como variável a estimativa de vendas, pois estaríamos a introduzir demasiado *bias* no modelo. No que diz respeito aos parâmetros dos algoritmos utilizados, os do algoritmo *Microsoft Time Series* mantiveram-se sempre iguais, enquanto que os parâmetros do algoritmo *Microsoft Neural Networks* variaram para que o modelo pudesse ser testado perante diferentes configurações do algoritmo. Na Tabela 3 podemos ver os parâmetros que foram configurados e as variações que ocorreram bem como a sua descrição.

Tabela 2 - Variáveis utilizadas pelos modelos para realizar a previsão de vendas

Modelo	Variáveis
ARTXP	Vendas
ARTXP	Vendas e variáveis internas
ARTXP	Vendas e variáveis externas
ARTXP	Vendas+Variáveis Internas+Variáveis Externas
ARTXP + ARIMA	Vendas
ARTXP + ARIMA	Vendas e variáveis internas
ARTXP + ARIMA	Vendas e variáveis externas
ARTXP + ARIMA	Vendas+Variáveis Internas+Variáveis Externas
ARIMA	Vendas
Redes Neuronais	Vendas+Variáveis Internas+Variáveis Externas
Redes Neuronais	Vendas e variáveis internas
Redes Neuronais	Vendas e variáveis externas

Tabela 3 - Parâmetros usados pelos algoritmos para configuração dos modelos a utilizar

Algoritmo	Parâmetro	Valores	Descrição
Microsoft Time Series	<i>Auto Detect Periodicity</i>	1	Procura padrões na variação dos dados que permitam identificar um possível comportamento futuro. O valor 1 representa uma procura de periodicidade totalmente automática.
Microsoft Time Series	<i>Forecast Method</i>	ARIMA, ARTXP, MIXED	Permite processar o modelo desejado através de um dos algoritmos disponibilizados
Microsoft Time Series	<i>Missing Value Substitution</i>	Mean	No caso de um valor de uma série temporal ser nulo, este é substituído por uma média.
Microsoft Neural Networks	<i>Hidden Node Ratio (HNR)</i>	1,4,8	Define o número de <i>hidden layers</i> *
Microsoft Neural Networks	<i>Holdout Percentage (HP)</i>	30,60	Percentagem dos dados que serão usados para treino do algoritmo.
Microsoft Neural Networks	<i>Maximum Input Attributes</i>	60000	Determina a quantidade de variáveis de entrada que o modelo aceita antes de começar a considerar para efeitos de previsão apenas valores estatisticamente relevantes.

* O número de *hidden layers* é calculado pela fórmula $HIDDEN_NODE_RATIO * \sqrt{Total\ input\ neurons * Total\ output\ neurons}$ sendo *input neurons* as variáveis de entrada e *output neurons* as variáveis de saída.

A atribuição do valor “1” à variável *Auto Detect Periodicity* relaciona-se com o facto de termos procurado um modelo genérico, sendo portanto oportuno maximizar este parâmetro. Aplicando valores de médias móveis sob valores desconhecidos, é possível corrigir a falta de informação sem comprometer o resultado global. O parâmetro *Hidden Layer Ratio* permite controlar a utilização de um maior ou menor número de nodos intermédios. Esta variação de nodos intermédios permite verificar se o aumento de nodos intermédios, e consequente aumento do número de cálculos do algoritmo, influencia positiva ou negativamente um *trade-off* entre velocidade de processamento e qualidade da previsão. Quanto ao valor de *Holdout Percentage*, utilizámos o valor 60, sendo este um valor pertinente uma vez que a partir de 50% de *holdout* começa a ser propício o fenómeno de *overfitting*. *Overfitting* é um fenómeno no qual existiu um treino tão intenso do modelo que este é capaz de realizar previsões com erros muito baixos na fase de treino, mas não é capaz de generalizar o seu desempenho obtendo erros mais elevados em testes perante novos dados [Demuth e Beale, 2004]. Com 60% de *holdout* será permitido realizar previsões com uma baixa probabilidade de *overfitting*. O objetivo de colocar um *maximum input attribute* com um valor tão elevado, prende-se com o facto de ser necessário evitar que este comece a descartar as variáveis estatisticamente menos significativas, uma vez que num contexto económico é necessário disponibilizar a introdução de valores bastante dispares em relação aos valores mais comuns.

3.4 Resultados obtidos

Os resultados apresentados comparam o erro percentual absoluto médio (EPAM) da previsão, considerando a diferença entre o valor previsto e os valores reais, para os últimos dez períodos conhecidos. O erro médio absoluto foi calculado utilizando a seguinte fórmula:

$$EPAM = \frac{\sum_{i=1}^n |EP_i|}{n}$$

na qual EP corresponde ao erro percentual. Este, por sua vez, foi calculado pela seguinte fórmula:

$$EP = \frac{X_t - F_t}{F_t} * 100$$

em que X_t é o valor observado no período t e F_t é o valor previsto para o período t .

De seguida, foram comparados os resultados obtidos com os modelos ARIMA, ARTXP, ARTXP + ARIMA e o modelo baseado em redes neuronais, com base nas suas diferentes configurações de parâmetros e de variáveis. O modelo baseado em redes neuronais teve a particularidade de ter sido testado recorrendo ao conhecimento dos valores de vendas para os dez períodos temporais em análise. Ao contrário dos outros modelos, este modelo não realiza uma extrapolação dos dados para o futuro. Desta forma, necessita sempre de recorrer a valores das variáveis para realizar a previsão. Assim, achamos ser relevante para o estudo em causa verificar o seu desempenho perante o facto de este modelo recorrer ou não a uma estimativa das vendas para o período a prever uma vez que este algoritmo necessita de possuir valores conhecidos em um determinado período para realizar previsões.

Tabela 4 - EPAM para os modelos ARTXP + ARIMA, ARTXP e ARIMA, resultante da aplicação do modelo a uma empresa e do mesmo modelo a todas as empresas

Modelo	Variáveis	Empresas			
		A	B	C	Média
ARTXP + ARIMA	Vendas	269,79%	18,28%	54,55%	114,21%
	Vendas + Int	269,79%	18,28%	54,55%	114,21%
	Vendas + Ext	430,95%	19,51%	54,98%	168,48%
	Todas	430,95%	21,21%	57,52%	169,89%
ARTXP	Vendas	341,48%	28,80%	64,72%	145,00%
	Vendas + Int	205,94%	28,80%	64,72%	99,82%
	Vendas + Ext	1234,07%	17,28%	64,50%	438,62%
	Todas	1234,07%	19,04%	64,50%	439,20%
ARIMA	Vendas	388,67%	21,50%	50,83%	153,67%

Na tabela 4 é visível ver que, na comparação individual, nenhum modelo é melhor do que os restantes, pois cada empresa obtém um melhor desempenho em modelos diferentes. A análise permite identificar que, quanto à utilização dos vários conjuntos de variáveis, existiu um melhor

desempenho dos modelos que utilizaram as vendas juntamente com as variáveis internas. Esta constatação foi reforçada pelo melhor EPAM referente à aplicação do mesmo modelo a todas as empresas (99,82%). O EPAM de 100% foi o melhor obtido com a aplicação de um mesmo modelo a todas as empresas, sendo este referente à utilização de um modelo ARTXP sobre os dados provenientes de vendas e de variáveis internas. No que diz respeito aos dados apresentados na tabela 5, estes não permitem identificar qual a melhor combinação de variáveis em relação aos melhores resultados individuais. Tal como sucedeu

na tabela 4, o melhor resultado obtido na aplicação do mesmo modelo às três empresas foi o modelo que recorreu às vendas e às variáveis internas (neural networks com vendas, holdout de 60%, 1 nível de *hidden layer* e apenas recorrendo a variáveis internas obtém um EPAM médio de todas as empresas de 44,18%).

Tabela 5 - EPAM resultante da aplicação dos modelos baseados em redes neuronais a cada empresa e do mesmo modelo **referente à média** de todas as empresas

	HP	HNR	Variáveis	Empresas			
				A	B	C	Média
C\Vendas	30	1	Todas	207,23%	35,00%	36,58%	92,93%
			Int	118,41%	22,77%	121,16%	87,45%
			Ext	177,08%	42,60%	29,01%	82,89%
		4	Todas	169,66%	35,51%	30,93%	78,70%
			Int	160,98%	19,46%	40,12%	73,52%
			Ext	182,14%	37,45%	40,12%	86,57%
		8	Todas	142,91%	34,58%	27,49%	68,33%
			Int	160,26%	21,04%	50,61%	77,30%
			Ext	143,65%	33,00%	44,45%	73,69%
	60	1	Todas	185,45%	25,26%	38,74%	83,15%
			Int	81,75%	17,76%	33,01%	44,18%
			Ext	208,71%	34,87%	26,57%	90,05%
		4	Todas	159,51%	29,39%	25,97%	71,63%
			Int	124,80%	16,71%	47,34%	62,95%
			Ext	134,07%	36,81%	16,76%	62,55%
		8	All	148,30%	21,01%	23,86%	64,39%
			Int	158,16%	16,92%	48,68%	74,59%
			Ext	107,92%	29,12%	28,60%	55,21%
s\Vendas	30	1	Todas	114,34%	19,73%	79,33%	71,13%
			Int	83,21%	21,00%	85,57%	63,26%
			Ext	308,69%	17,42%	74,96%	133,69%
		4	Todas	138,47%	16,43%	79,46%	78,12%
			Int	81,51%	20,17%	77,04%	59,57%
			Ext	133,78%	26,64%	77,04%	79,15%
		8	Todas	100,48%	26,16%	74,87%	67,17%
			Int	200,97%	19,10%	75,27%	98,45%
			Ext	102,22%	19,86%	83,19%	68,42%
	60	1	Todas	146,23%	17,21%	80,49%	81,31%
			Int	76,45%	19,32%	83,78%	59,85%
			Ext	148,30%	24,17%	75,17%	82,54%
		4	Todas	119,00%	19,27%	82,22%	73,50%
			Int	90,00%	18,04%	77,04%	61,66%
			Ext	148,30%	24,17%	76,51%	82,99%
		8	Todas	148,30%	24,76%	77,25%	83,44%
			Int	1058,60%	20,57%	75,00%	384,70%
			Ext	250,35%	24,62%	81,31%	118,76%

A tabela 5 não é, pois, conclusiva quanto à utilização ou não de estimativa de vendas para os períodos de previsão por parte dos modelos baseados em redes neuronais, isto porque, para duas empresas, o melhor resultado foi obtido através de modelos sem estimativas de vendas e para a empresa restante foi para o modelo que utiliza variáveis externas e utiliza estimativas de vendas para auxiliar a previsão. O melhor resultado foi obtido considerando que foi realizada uma estimativa de vendas para o período a prever. A comparação do melhor EPAM conseguido por modelos aplicados a cada empresa e a média de todas as empresas em conjunto está apresentada na tabela 6. Esta tabela apresenta o melhor resultado conseguido por cada modelo para as diversas configurações adotadas para os seus parâmetros e variáveis. Desta forma, foi possível identificar que os modelos baseados em redes neuronais produzem uma melhor qualidade de previsão, uma vez que os melhores resultados, quer individuais quer da aplicação de um modelo a todas as empresas, foram obtidos por eles. O melhor resultado de todos foi obtido com a aplicação de um modelo

baseado em redes neurais recorrendo a 60% dos dados para treino e a um parâmetro de *Hidden Node Ratio* de 1, o que origina dois níveis de *hidden layers* na estrutura, recorrendo a variáveis internas para realizar a previsão.

Tabela 6 - Melhor EPAM obtido por cada modelo para uma dada empresa e a média referente a todas as empresas

Algoritmo	Empresa			Melhor Média
	A	B	C	
ARTXP + ARIMA	270%	18%	55%	114%
ARTXP	206%	17%	65%	100%
ARIMA	389%	22%	51%	154%
NN 30 C\ Vendas	118%	19%	27%	68%
NN 60 C\ Vendas	82%	17%	17%	44%
NN 30 S\ Vendas	82%	16%	75%	60%
NN 60 S\ Vendas	76%	17%	75%	60%

Os gráficos apresentados na figura 2 permitem analisar a proximidade dos resultados das previsões alcançadas com os valores reais das vendas realizadas pelas empresas. Neles são comparados os resultados da melhor previsão para cada empresa com a previsão obtida pelo modelo com melhores resultados, quando aplicado a todas as empresas. Esta comparação permite identificar o quanto se aproximam as previsões dos valores reais.

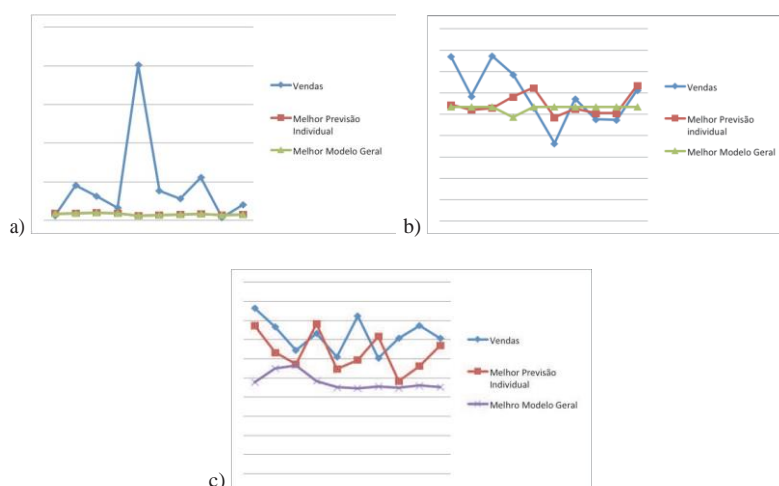


Figura 2 - Comparação entre as previsões obtidas e os valores reais para as empresas a) A, b) B e c) C, respetivamente

No caso da empresa A ambas as previsões encontram-se distanciadas dos valores reais. No segundo gráfico é visível que apenas em dois períodos a previsão do melhor modelo não segue a tendência dos valores reais, conseguindo, porém, nos últimos quatro períodos ter resultados próximos dos valores reais. Por sua vez a previsão do melhor modelo geral, embora obtenha um EPAM próximo da melhor previsão para a empresa (18% e 16%, respetivamente), não segue as tendências económicas dos valores reais. Por fim, a empresa C apresenta inicialmente um resultado da melhor previsão bastante próxima do modelo real nos primeiros períodos, acabando por ter uma tendência diferente no sétimo período que se reflete depois nos restantes períodos. O resultado obtido pelo modelo com melhores resultados globais é distante do resultado obtido pelo melhor modelo. Porém, esta diferença era expectável devido à diferença entre os EPAM, 33% para o modelo com melhor resultado global e 17% para o melhor resultado obtido para a empresa.

3.5 Análise e discussão de resultados

Os resultados obtidos na tabela 6 revelam que os modelos baseados em redes neuronais conseguem realizar melhores previsões com base em séries temporais do que os modelos ARIMA, ARTXP e ARIMA+ARTXP. Porém, tal desempenho tem uma pequena desvantagem: exige informação das variáveis sobre os k períodos a prever. Assim, é necessário avaliar a possibilidade de estimar as variáveis a utilizar nos k períodos e ter em conta que uma má estimativa pode comprometer o resultado da previsão. Embora o melhor resultado obtido tenha sido através de um modelo baseado em redes neuronais, que recorre à estimativa de vendas para os k períodos a prever, devemos ser um pouco críticos no que a ele diz respeito. Este resultado foi obtido essencialmente devido à grande diferença na qualidade de previsão na empresa C, sendo que o melhor resultado é um EPAM de 51%, descartando o modelo de *Neural Networks* que utiliza estimativa de vendas, o que é bastante superior ao EPAM conseguido no melhor caso que foi de 17%. Esta diferença pode ser justificada com o facto de o último valor registado, antes de serem realizadas as previsões, apresentar um incremento de mais do dobro do valor de vendas. Este incremento não foi acompanhado pelos restantes modelos, que não lhe atribuíram o peso devido, uma vez que todos os outros valores estavam na mesma gama de valores - 0 e 2M€.

Ao contrário dos outros modelos, o modelo que utilizou uma estimativa de vendas, conseguiu utilizá-la para influenciar os valores da previsão. Ao serem utilizadas estimativas próximas do valor real obtido, o modelo foi influenciado beneficiando bastante a qualidade de previsão para os k períodos desejados. Se os modelos baseados em redes neuronais, com recurso a estimativa de vendas fossem descartados devido a esta última empresa, o melhor modelo a realizar uma previsão, sobre todas as empresas, seria igualmente um modelo baseado em redes neuronais. Porém, este melhor resultado foi obtido tanto por uma configuração de parâmetros de 30 HP e 4 HNR, e de 60 NP e 1 HNR. Como a utilização de 60% dos dados poderá levar a uma situação de *overfitting*, caso não tivesse sido considerado um modelo com recurso a estimativas de vendas, o melhor modelo teria sido um modelo baseado em redes neuronais com 30% dos dados para treino e dez níveis de *hidden layers*.

4 Conclusões

Através das comparações efectuadas com os resultados obtidos foi possível retirar algumas ilações quanto aos modelos utilizados nos processos de previsão realizados. Os modelos baseados em redes neuronais obtiveram sempre os melhores resultados. Porém, aquando da sua escolha, é necessário considerar o esforço extra necessário para produzir a estimativa de variáveis desconhecidas em relação aos k períodos que foram alvos do processo de previsão. Contudo é necessário ter sempre em conta o *bias* inserido no modelo associado às estimativas. Se considerarmos os modelos que não utilizam estimativas de variáveis, o

modelo ARTXP perfila-se como a melhor escolha, de acordo com os resultados obtidos. Porém é necessário salguardar que os melhores resultados obtidos por modelos que utilizavam variáveis internas apenas conseguiram melhores desempenhos devido às variáveis utilizadas. Diferentes empresas e diferentes variáveis podem produzir resultados diferentes. Sendo a previsão apenas um instrumento de apoio à decisão, o melhor EPAM obtido, 44%, pode ser considerado por alguns agentes de decisão empresariais como aceitável perante o *trade-off* de tempo de implementação e execução para com a qualidade da previsão. Através da análise dos resultados realizada pode-se concluir que embora as previsões realizadas com foco numa só empresa proporcionam melhores resultados. Todavia, a utilização de modelos genéricos de previsão pode ser útil perante um cenário no qual seja privilegiado o baixo custo de implementação e a configuração de modelos.

5 Referências

- Bose, I., Mahapatra, R. K., Business Data Mining – A Machine Learning Perspective, Information & Management, Volume: 39 Issue: 3 pp.211-225, 2001.
- Box, G., Jenkins, G., Time Series Analysis, Forecasting and Control Holden-Day, San Francisco, CA (1970)
- Box, G., Jenkins, G., Andreinsel, G., Time Series Analysis: Forecasting and Control, Wiley, 4 edition, 2008.
- Chereb, D., Does Data Mining Improves Business Forecasting?, The 18th International Symposium on Forecasting, Edinburgh, Scotland, 1998.
- Demuth, H & Beale, M, 2004, Neural Network Toolbox for Use with MATLAB,. The mathworks inc.,Natick, MA.
- Geurts, P., Pattern extraction for time series classification, In Proceedings of the 5th European Conference on Principles of Data Mining and Knowledge Discovery (PKDD '01), Luc De Raedt and Arno Siebes (Eds.). Springer-Verlag, London, UK, UK, 115-127, 2001.
- Keogh, E., Kasetty, S., On the need for time series data mining benchmarks: a survey and empirical demonstration. In Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining (KDD '02). ACM, New York, NY, USA, 102-111, 2002
- Lewis, P., Ray, B. e Stevens, J. 1994. Modeling time series by using mars. In Timeseries prediction, pages 297-318. AddisonWesley, New York, 1994.
- Lin MI, Groves WA, Freivalds A, Lee EG, Harper M., Comparison of artificial neural network (ANN) and partial least squares (PLS) regression models for predicting respiratory ventilation: an exploratory study, Eur J Appl Physiol, 112(5):1603-11, 2012.
- Makridakis, S., Wheelwright, S., Hyndman, R., Forecasting: Methods and Applications, John Wiley & Sons Inc., 1998.
- Meek, C., Chickering, D., Heckerman, D.,, “Autoregressive Tree Models for Time-Series Analysis,” Proc. Second SIAM Int'l Conf. Data Mining (SDM '02), 2002.
- Ratanamahatana, C., Lin, J., Gunopulos, D., Keogh, E., Vlachos, M., Das, G., Mining Time Series Data, Data Mining and KNowledge Discovery Handbook, pp 1049-1077, 2010.
- SQL Server 2013. Microsoft SQL Server Analysis Services [online] Disponível em: <http://www.microsoft.com/en-us/sqlserver/solutions-technologies/business-intelligence/analysis.aspx> [Acedido em 18 de Junho de 2013].
- Shatkay, H. Zdonik, S., Approximate Queries and Representations for Large Data Sequences. Proc. of the Int. Conf. on Data Engineering (ICDE), pp. 536-545, 1996.
- Song, H., LI, G., Tourism demand modelling and forecasting – A review of recent research, Tourism Management 29, no. 2, 203-220, 2008.
- Tsaih, R., Hsu, Y., Lai, C., Forecasting S&P 500 stock index futures with a hybrid AI system, Decision Support Syst., 23, pp. 161–174, 1998.
- Valipour, M., Banihabib, M. E., & Behbahani, S. M. R., Comparison of the ARMA, ARIMA, and the autoregressive artificial neural network models in forecasting the monthly inflow of Dez dam reservoir. Journal of Hydrology, 2012.