Um método de análise de variáveis causais para previsão de demanda no setor farmacêutico

Ricardo Alexandre Feliciano (POLITECNICA/USP) afn250946@terra.com.br
Mauro de Mesquita Spinola (POLITECNICA/USP) mauro.spinola@poli.usp.br
Tamio Shimizu (POLITECNICA/USP) tmshimiz@usp.br

Resumo

O objetivo deste trabalho é descobrir fatores que influenciam a previsão de vendas no setor farmacêutico e analisar a eficácia de um método que se propõe a isso, chamado de análise de componentes principais. Mantendo somente dados relevantes num futuro dispositivo de inteligência artificial, a capacidade de previsão pode ser aumentada e o tempo de processamento reduzido. Neste estudo foi utilizado o método de coleta de dados, verificação de integridade e tratamento de dados, além de submissão destes a um pacote computacional de estatística (The Unscrambler 9.2 ®) que analisou a influência de 5 fatores independentes sobre a demanda. Os principais resultados encontrados foram: (1) a temperatura foi a variável que teve maior correlação com a venda de itens sazonais; (2) a variável temperatura apresentou fortes correlações com a venda de itens não sazonais, o que não era esperado. Estes resultados evidenciam a eficácia da técnica de análise de componentes principais para automatizar a análise de causalidade proposta, embora estas 5 variáveis colhidas sejam insuficientes para análise de correlação de vendas de produtos não sazonais.
Palavras-chave: Previsão de Vendas; Análise de Dados Multivariada; Análise de Componente Principal

1. Introdução

Uma das técnicas mais avançadas para criação de modelos voltados à detecção de padrões e previsão de demanda é a mineração de dados, que inclui métodos de inteligência artificial, redes neurais, entre outras. Atualmente, os gerentes mais eficientes na previsão da demanda são capazes de analisar adequadamente os resultados fornecidos pelas técnicas quantitativas, trazendo otimização de estoques e dos lucros.

O objetivo deste trabalho é descobrir fatores que influenciam a previsão de vendas no setor farmacêutico, a fim de manter somente os dados relevantes na entrada de um dispositivo de inteligência artificial sem perder capacidade de previsão, reduzindo o tempo de processamento. Também é objeto deste trabalho analisar a eficácia de um método que se propõe a isso, chamado de análise de componentes principais (ACP).

O distribuidor farmacêutico objeto deste estudo foi nomeado "*Distribuidor Ltda.*", considerando que informações estratégicas sobre demanda, estoques e preços praticados foram colhidas. Como todo distribuidor farmacêutico, sofre com as incertezas da demanda (variabilidade, sazonalidade, etc) aliadas às limitações de estocagem.

Os tópicos que fazem parte dos Fundamentos Teóricos trazem uma definição genérica de modelo e comentam a importância da previsão de vendas, ao mesmo tempo que ilustram um modelo para tal. Também citam uma metodologia para escolha de variáveis de modelos causais, diferenciam a estatística multivariada da análise de dados multivariada e trazem conceitos de uma ferramenta de análise de dados multivariada: a análise de componentes principais.

2. Fundamentos teóricos

2.1 Modelos

Shimizu (2001) comenta que há vantagem no uso de informações reais de mercado, quando estas existem, eliminando o uso de probabilidades e valores fictícios. As informações reais não conferem uma obrigação, mas o direito de tomada de decisão no futuro.

As empresas perseguem a inteligência organizacional com o uso da heurística, ou seja, método de solução de problemas indutivo baseado em regras derivadas do senso comum ou da experiência de um modelo teórico da matemática, fornecendo uma base geral para a solução de problemas, como por exemplo na exploração e descoberta de fenômenos. O uso de modelos - representação ou interpretação simplificada da realidade - geralmente conduz a rumos novos e importantes.

De acordo com Doyle et al. (1996), o uso de modelos computacionais pode explicar o significado de fenômenos psicológicos, sociais, mercadológicos entre outros. As atuais pesquisas em inteligência artificial (em termos de aprendizado de máquina e adaptação) envolvem: representação do conhecimento de formas eficientes para catalogação e uso posterior, descoberta de técnicas analíticas e estatísticas para extrair tendências, fatores, experiência e dados.

Shimizu (2001) comenta que a escolha do modelo depende da finalidade da decisão, da limitação do tempo e custo e da complexidade do problema. Um problema pode ser considerado complexo quando os valores das variáveis são definidos de modo impreciso, quando existem riscos ou incertezas nesses valores ou ainda, como no caso desta pesquisa, quando o número de variáveis aumenta – os chamados problemas multidimensionais.

2.2 Previsão de vendas

Tanto decisões estratégicas como operacionais de uma organização requerem a exploração do relacionamento entre os elementos que compõem a realidade em que a organização está inserida. Para apoiar essas decisões citadas, as organizações procuram modelos de previsão com o objetivo de prognosticar o futuro através do exame do passado.

Passari (2003) salienta duas razões básicas para confiar na dependência de observações causais nas séries temporais de demanda de produtos: fatores econômicos que contribuem para a geração de valor não mudam repentinamente e a sazonalidade como padrão de longo prazo é repetitiva.

A Figura 1 ilustra um modelo criado por Mentzer & Kent (1999) e Subrahmanyan (2000) onde buscaram relacionar as variáveis causais com a demanda individual de cada produto, desenvolvendo um modelo de previsão individual para cada produto, que toma como entrada os valores destas variáveis explicativas da demanda ao longo do tempo.

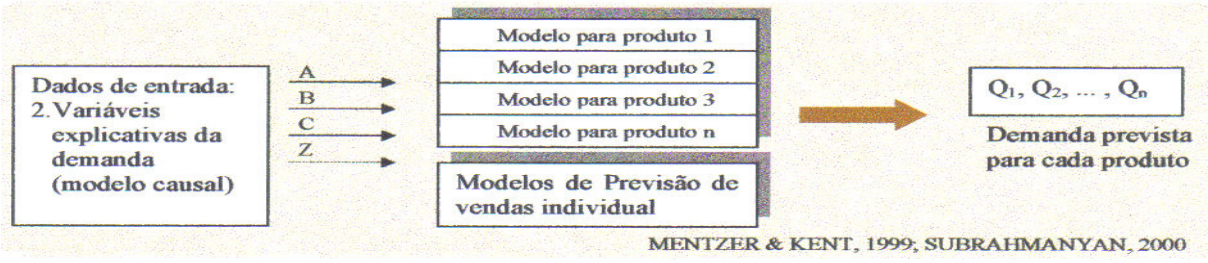

Figura 1 – Modelo de previsão de vendas com uso de dados individuais e modelagem causal (MENTZER & KENT, 1999 e SUBRAHMANYAN, 2000)

2.3 Escolha de variáveis

Makridakis et al. (1983) desenvolveram uma metodologia para a escolha das variáveis a serem utilizadas em modelos causais:

a) Determinação de uma longa lista de variáveis potenciais que possam afetar a variável dependente analisada (problema multidimensional), baseada na opinião de especialistas, na disponibilidade dos dados e no esforço e custo associados à aquisição dos mesmos;

Alguns fatores que alteram o comportamento do consumo (variáveis causais ou independentes) na indústria farmacêutica são:

- Inovações Técnicas
- Influências sazonais
- Preços competitivos dos concorrentes
- Tipos de produtos retirados da linha de produção
- Produtos retidos pela Agência Nacional de Vigilância Sanitária
- Renda per capita
- Promoções, campanhas de marketing
- Disponibilidade de estoque
- Inflação

Neste trabalho foram utilizados: influências sazonais, renda per capita e inflação, além de estoque dos produtos, preço unitário médio e a demanda de produtos em si, a variável dependente.

b) Redução para uma lista curta, eliminando algumas possíveis variáveis menos relevantes;

Isto pode ser feito:

- observando as correlações entre cada variável independente;
- efetuando uma regressão múltipla com todas as variáveis;
- usando métodos sofisticados como regressão stepwise (método iterativo de retirada de variáveis e testes de significância);
- e análise de componentes principais.

O limitador deste estudo foi a ausência de dados de concorrentes do "*Distribuidor Ltda*." e outros dados sócio-econômicos como, por exemplo, crescimento vegetativo da população, acesso à saneamento básico, registros de epidemias, taxa do dólar. Tais dados contribuiriam para a construção de um modelo mais completo, explicando fenômenos que, porventura, possam ficar ocultos num modelo de apenas 5 variáveis.

2.4 Análise de dados multivariada versus estatística multivariada

A estatística é uma ferramenta útil para a área de descoberta de comportamentos. Quando fenômenos são estudados a partir de dados coletados em muitas variáveis, os métodos estatísticos delineados para obter informações a partir desses conjuntos de dados são denominados de métodos de análise de dados multivariada.

Esbensen (2002) explica que a estatística multivariada e a análise de dados multivariada são campos cuja fronteira não é clara, embora cada um possua leves diferenças. O foco da primeira é observar a parte da estrutura de dados onde se encontram os ruídos, ou seja, os erros aleatórios ou não explicados. A segunda foca principalmente a parte da estrutura dos dados onde se encontra a explicação para os fenômenos pesquisados. A questão é que as

observações são soma de ambos. Não podemos, de imediato, ver o que deve ser mantido e o que deve ser descartado.

O objetivo da análise de dados multivariada é usar as correlações intrínsecas às variáveis de um conjunto de dados para separar a parte da estrutura que interessa à pesquisa e o ruído. A análise de componente principal é um método freqüentemente usado para análise de dados multivariada de dados de uma matriz de dados dimensional genérica, de forma a decompor dados para detectar fenômenos ocultos.

Schoeder & Noy (2001) e Lino (2004) mostram interessantes características de sistemas com múltiplos agentes (variáveis independentes) e aplicações reais da análise de dados multivariada.

2.5 Análise de componentes principais (ACP)

Considerando uma matriz X com 'n' amostras - observações, objetos ou pontos - e somente 3 variáveis, a representação espacial terá 3 eixos: X1, X2 e X3. O gráfico a seguir representa esta matriz plotada:

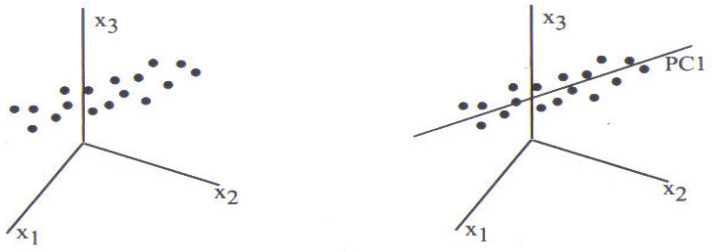

Figura 2 - Dados plotados analisados por um eixo chamado componente principal (ESBENSEN. 2002)

No exemplo, percebemos uma tendência entre os objetos, que podemos chamar de 'associação linear oculta'. Um eixo central pode ser desenhado através deste pontilhado de forma a representar os dados de forma tão eficiente quanto os 3 eixos originais. Isso porque as variáveis estão fortemente correlacionadas. Este eixo central não precisa necessariamente ser paralelo a nenhum dos eixos originais e independentemente da "nova variável" que o eixo possa representar, ele é chamado de primeiro componente principal (PC1). Cada ponto é projetado perpendicularmente sobre o eixo e essa distância 'ei' é chamada de distância residual. Este eixo minimiza a soma de todas as distâncias transversais quadráticas Σ(ei)2 ou, em outras palavras, maximiza a variância.

Se a distribuição dos pontos não puder ser bem correlacionada por um único componente principal, novos componentes principais devem ser descobertos.

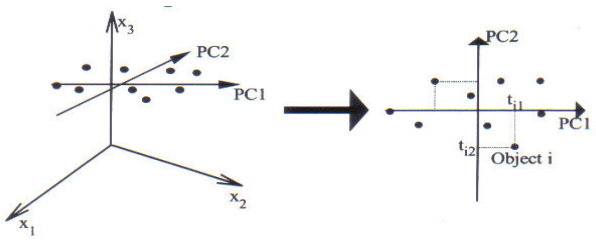

Figura 3 - Coordenadas no espaço PC (ESBENSEN. 2002)

Uma transição entre o espaço multidimensional e o espaço de componentes principais pode ser visualizada em gráficos como o "X-loadings". Esses gráficos conhecidos como gráficos de

carregamento são mapas de variáveis e mostra quanto cada variável contribui com cada PC. Nota-se na figura a seguir que os pontos plotados são variáveis ao invés de observações (objetos). Os gráficos de carregamento trazem uma projeção dos relacionamentos inter-variáveis (par a par) e ajuda na interpretação dos mesmos.

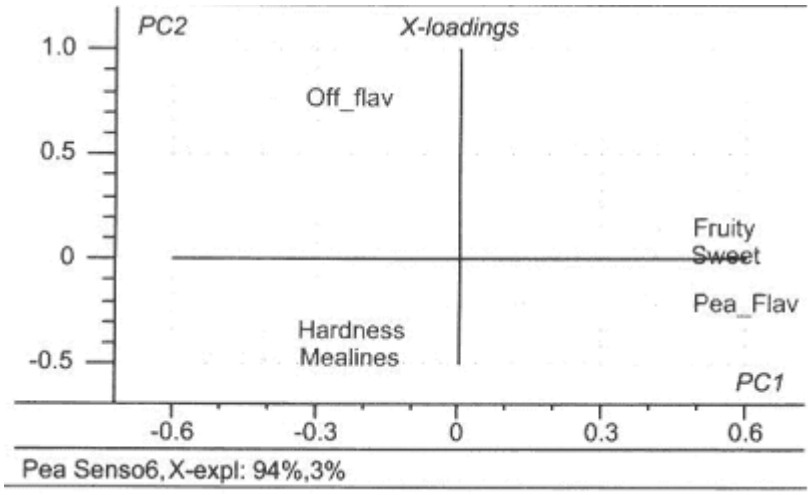

Figura 4 – Gráfico de carregamento para o par PC1 e PC2 (ESBENSEN. 2002)

A variância explicada é a porção da variância total que é levada em conta pelo modelo. É computada como diferença entre a variância total e a residual – que resulta a variância explicada - dividida pela variância total, expressa em porcentagem. Como exemplo, uma variância explicada de 90% significa que 90% da variação nos dados é entendida pelo modelo, enquanto que os 10% restantes são ruído (erro). A variância explicada e a não explicada sempre somam 100% da variância total. Sendo uma função crescente a cada PC acrescentado – pois cada PC acrescentado traz mais variância explicada ao sistema - o gráfico de Validação pela Variância Explicada é usado para descobrir um número ótimo de PC's para a modelagem. No exemplo da figura a seguir, 2 PC's explicam aproximadamente 90% do fenômeno pesquisado.

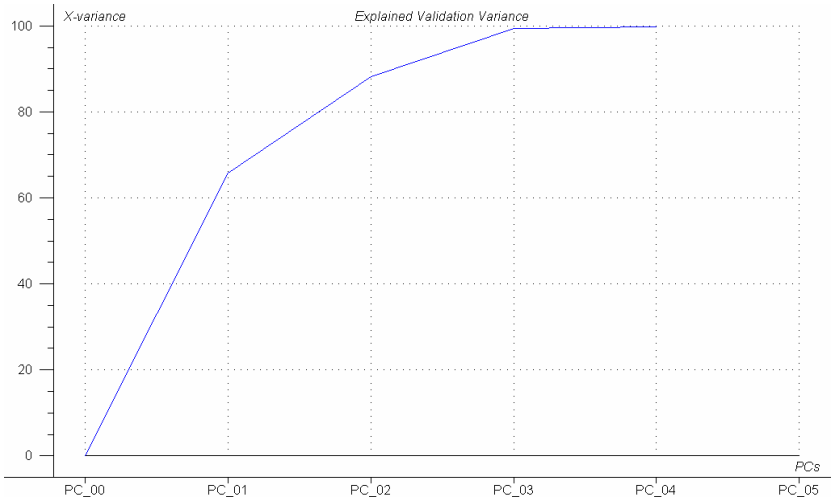

Figura 5 - Gráfico de Validação pela Variância Explicada (ESBENSEN. 2002)

A normalização é um pré-tratamento matemático realizado na matriz de dados para compensar as diferenças de ordem de grandeza. Cada elemento da linha é dividida pela

diferença entre o maior e o menor valor da linha da matriz. Análises específicas requerem pré-tratamentos específicos, embora a normalização seja um pré-tratamento comum a todas.

Esbensen (2002) ilustra a teoria de análise de componentes principais ao mesmo tempo que mostra o uso do pacote computacional The Unscrambler 9.2 ® e a interpretação de gráficos e resultados.

3. Método

O método adotado para desenvolvimento do estudo seguiu a seguinte linha de raciocínio: coleta de dados, verificação da integridade dos dados, tratamento de dados, submissão dos dados a uma ferramenta estatística e obtenção de resultados.

O levantamento de dados reais objetiva retratar o dinamismo de uma situação particular, focado em fenômenos contemporâneos do segmento em questão.

Os dados de alguns itens relativos a vendas, preços praticados e estoques, disponíveis no banco de dados da empresa, foram exportados para planilhas Microsoft Excel. Abrangem o período de maio de 2004 a outubro de 2005, por ser este o período em que o "*Distribuidor Ltda*" se estabilizou em termos de vendas desde sua abertura em abril de 2003. Dezoito meses são o suficiente para demonstrar possíveis sazonalidades (inverno ou verão), verificando-se picos e vales de vendas. A variável preço unitário refere-se ao preço unitário médio praticado no mês, em virtude de descontos, promoções e campanhas de marketing. A variável estoque refere-se ao percentual de dias do mês com estoque suficiente para atender à demanda média, o que explica sua variação entre 0 e 1 (100%). A temperatura média em Belo Horizonte (local onde se situa o distribuidor), renda per capita e índices de inflação na mesma capital e relativos ao período acima citado foram coletados em sítios de órgãos oficiais na Internet (IBGE – Instituto Brasileiro de Geografia e Estatística, INPE – Instituto Nacional de Pesquisas Espaciais, CPTEC – Centro de Previsão de Tempo e Estudos Climáticos e FGV – Fundação Getúlio Vargas).

Em virtude do número reduzido de itens estudados, verificou-se, sem o auxílio de ferramentas computacionais, itens com células em branco ou com números inconsistentes (por exemplo, venda, estoque ou preço negativos). Tais itens foram excluídos das planilhas.

Os dados coletados foram submetidos ao pacote computacional de estatística The Unscrambler 9.2 ®.

4. Resultados

O item Caladryl é marca registrada e foi usado para ilustração de um caso prático de comportamento mercadológico, sendo um produto sazonal muito usado no verão, especialmente no tratamento de queimaduras solares. A matriz de dados deste produto foi submetida ao pacote computacional, assim como a de outros itens. Os resultados aqui apresentados explicam pontualmente o comportamento deste produto.

		qtde	$unit	estq	temp_BH	inflação	renda
		1	2	3	4	5	6
mai/04	1	5.0000	12.0920	1.0000	20.0000	1.3100	869.6974
jun/04	2	3.0000	11.6133	1.0000	18.6667	1.3800	884.4944
jul/04	3	8.0000	11.7525	1.0000	17.3333	1.3100	894.7520
ago/04	4	7.0000	11.5014	1.0000	19.0000	1.2200	902.1222
set/04	5	9.0000	11.7511	1.0000	22.0000	0.6900	880.7152
out/04	6	12.0000	11.5792	1.0000	22.8333	0.3900	876.5785
nov/04	7	41.0000	11.3851	1.0000	23.5000	0.8200	863.7265
dez/04	8	23.0000	11.5509	1.0000	23.6667	0.7400	1.0416e+03
jan/05	9	43.0000	11.8663	0.9048	24.0000	0.3900	888.0967
fev/05	10	27.0000	12.0667	1.0000	23.5000	0.3000	897.1308
mar/05	11	19.0000	12.0042	1.0000	23.8333	0.8500	914.4780
abr/05	12	22.0000	12.7327	1.0000	23.5000	0.8600	905.8129
mai/05	13	16.0000	12.9144	1.0000	20.1667	−0.2200	893.4351
jun/05	14	16.0000	12.9456	1.0000	19.0000	−0.4400	912.8052
jul/05	15	9.0000	12.9822	1.0000	18.1667	−0.3400	897.0289
ago/05	16	17.0000	13.4200	1.0000	20.3333	−0.6500	901.9808
set/05	17	29.0000	13.5038	1.0000	22.0000	−0.5300	875.6723
out/05	18	75.0000	13.5039	0.7000	24.5000	0.6000	874.0675

Figura 6 – Matriz de dados para o produto Caladryl

A normalização foi o pré-tratamento matemático aplicado à matriz de dados.

Plotando um gráfico de carregamento para o par PC1 e PC2 verificamos quanto cada uma das 5 variáveis contribuem com cada componente principal. Seguem os gráficos:

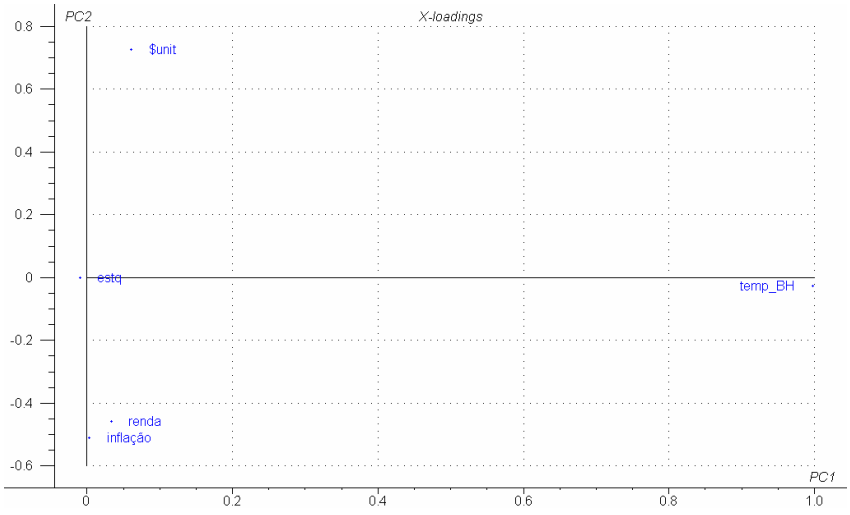

Figura 7 – Gráfico de Carregamento para o par PC1e PC2 para o produto Caladryl

De acordo com a Figura 7, observa-se que a variável “temperatura” contribuiu em 99,80% com o PC1, ou seja, a demanda do produto Caladryl e a temperatura estão fortemente correlacionadas. Ainda, o PC1 é um eixo praticamente paralelo ao eixo das temperaturas no espaço multidimensional, com distâncias residuais mínimas entre eles (eixos).

O preço unitário contribui 72,70% com o PC2, a inflação -51,10% e a renda, -45,90%. Este segundo eixo correlaciona moderadamente estas três variáveis entre si e com a demanda, tendo outras duas variáveis “estoque” e “temperatura” praticamente não correlacionadas entre si e nem com a demanda. Mesmo assim, é o segundo melhor eixo que explica a correlação entre todas, atravessando o espaço multidimensional.

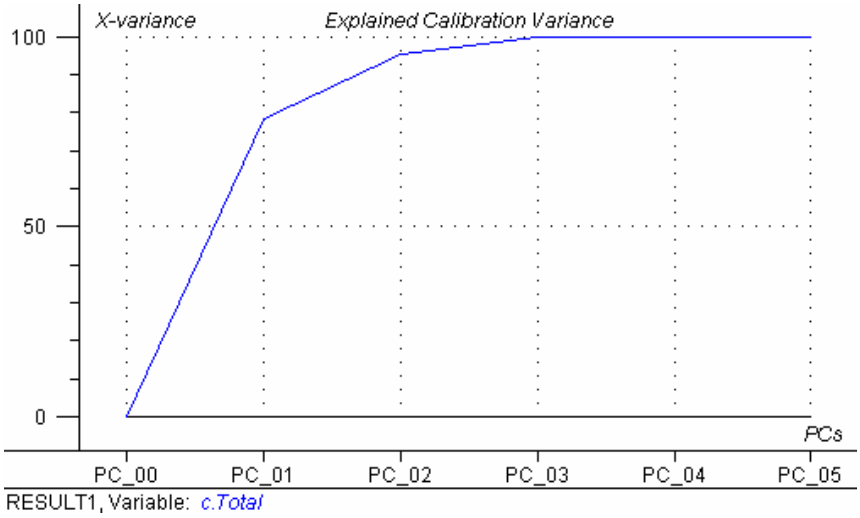

Figura 8 – Validação pela Variância Explicada para o produto Caladryl

De acordo com a Figura 8, o PC1 sozinho corresponde a 78,25% da variância explicada pelo sistema. Os dois primeiros componentes juntos correspondem a 95,36% da variância explicada pelo sistema. Observa-se que dois componentes são suficientes para explicar o fenômeno adequadamente.

5. Conclusões

Este trabalho aborda um problema de real importância econômica: a otimização de estoques e dos lucros em distribuidores farmacêuticos através da previsão de vendas.

Pela análise pontual dos resultados para o produto Caladryl, conclui-se que a temperatura por si só explica 78,25% da variação da demanda e que temperatura aliada a renda per capita, preço e inflação explicam pelo menos 95,36% da variação da demanda. A análise de resultados de todos os itens disponíveis mostra que a variável estoque não influenciou a demanda. A variável temperatura apresentou fortes correlações com a demanda de itens não sazonais, o que não era esperado. Estes resultados comprovam a eficácia da técnica de análise de componentes principais para automatizar a análise de causalidade proposta, embora as 5 variáveis colhidas sejam insuficientes para análise de correlação de vendas de produtos não sazonais.

Como proposta para trabalhos futuros, propõe-se coleta e análise de dados da concorrência e maior variedade de dados sócio-econômicos, o que foi justamente o limitador de estudo desta pesquisa.

Este trabalho pode servir como referência para trabalhos futuros de previsão de vendas no setor farmacêutico com o uso de inteligência artificial, submetendo às entradas destes dispositivos as variáveis relevantes descobertas nesta pesquisa.

6. Bibliografia

DEAN, T.; DOYLE, J. *Strategic Directions in Artificial Intelligence.* ACM Computing Surveys. Vol. 28, n.4, p. 653-670, 1996.

ESBENSEN, K.H. *Multivariate Data Analysis.. 5. ed.* Esbjerg: CAMO Process AS, 2002.

LINO, M.M. *Qualidade de Vida e Satisfação Profissional de Enfermeiras de Unidades de Terapia Intensiva.* USP: São Paulo, 2004.

MAKRIDAKIS, S.; WHEELWRIGHT, S.C.; MCGEE, V.E. *Forecasting: Methods and Application..* 2.ed. New York: John Wiley & Sons, 1983.

MENTZER, J.T.; KENT, J.L. *Forecasting demand in the Longaberger Company.* Marketing management. Vol. 8, n. 2, p.46-50, 1999.

NOY, P.; SCHOEDER, M. Multi-*Agent Visualisation Based on Multivariate Data.* Proceedings of the fifth internactional conference on Autonomous agents. Agents'01, May 2001; p.85-91, 2001.

PASSARI, A.F.L. Exp*loração de Dados Atomizados para Previsão de Vendas no Varejo Utilizando Redes Neurais*. USP: São Paulo, 2003.

SHIMIZU, T. *Decisão nas Organizações.*. 2.ed. São Paulo: Atlas, 2002.

SUBRAHMANYAN, S. *Using quantitative models for setting retail prices.* Journal of Product and Brand Management. Vol. 9, n. 5, p.304-320, 2000.