Inference of variables in the industry of the petroleum with dynamic PLS

Rosalvo Stachiw
CEFET-PR/CPGEI
Av. 7 de Setembro, 3165,
CEP 80230-901 - Curitiba (PR)
rosalvo@cpgei.cefetpr.br

Flávio Neves Junior
CEFET-PR/CPGEI
Av. 7 de Setembro, 3165,
CEP 80230-901 - Curitiba (PR)
neves@cpgei.cefetpr.br

Lucia V. Ramos de Arruda
CEFET-PR/CPGEI
Av. 7 de Setembro, 3165,
CEP 80230-901 - Curitiba (PR)
arruda@cpgei.cefetpr.br

***Abstract* — The present study treats of the inference of variables in the industry of the petroleum with dynamic PLS (Partial Least Square). The dynamic PLS is used to the estimation product composition of a debutanizer column (top and bottom products). The data of debutanizer distillation column process are obtained starting from a chemical and petrochemical processes simulator. The obtained results are presented in a comparative form with standard static PLS. As evaluation criterion, is the ability of the model is approached in the prediction of new data group. The results show the superiority of dynamic PLS in relation to static PLS due to ability of the model in obtaining information of the dynamic behavior of the system.**

***Index Terms* —dynamic PLS, PLS, debutanizer column**

I. INTRODUCÃO

A automação industrial, através da modernização de equipamentos e processos permite a geração de grande quantidade de informações. O tratamento destas informações e a determinação de correlações adequadas permitem que variáveis mensuráveis ou não possam ser estimadas.

Em sistemas químicos industriais as informações são obtidas em ambientes com ruídos e com variáveis correlacionáveis entre si. Em cima disto, diversas técnicas de inferência têm se mostrado eficiente para estimar o comportamento dinâmico destas variáveis.

A identificação de modelos monovariáveis para processos industriais é uma técnica consolidada, mesmo para processos não lineares. Para estes sistemas os algoritmos mais utilizados são os baseados na técnica dos mínimos quadrados seja em suas versões recursivas ou não recursivas e *on-line* ou *off-line* [1].

Para sistemas multivariáveis, a menos que o sistema seja desacoplado ou fracamente acoplado e possa ser tratado como múltiplos modelos monovariáveis, a aplicação da técnica de mínimos quadrados gera modelos polarizados ou com grandes erros de variância, devido à natureza mal condicionada do problema. Para controlar estes problemas, técnicas baseadas na correlação das variáveis do processo podem ser empregados.

Com isso, a técnica estatística multivariável PLS (*Partial Least Square*) [2], uma técnica de Projeção de Estruturas Latentes, pode ser utilizado para resolver problemas da indústria química como ferramenta auxiliar no monitoramento e modelagem de processos, detecção de falhas, etc.

O processo de destilação, como um todo, é amplamente utilizado em processos da indústria química, especialmente na indústria de petróleo. Pequenas melhorias operacionais neste processo podem refletir em significativas reduções de custos, principalmente quando se trata de recuperação de produtos, tornando-o objeto importante de estudo.

O objetivo deste trabalho é a obtenção de um modelo de inferência baseado na versão dinâmica do algoritmo PLS [3] [4]. As variáveis a serem estimadas são os produtos de topo e de fundo das correntes de saída do processo de destilação (debutanização). Os produtos de topo e de fundo representam os produtos GLP (Gás Liquefeito de Petróleo) e Nafta respectivamente.

Os dados de processo da coluna de debutanização são obtidos a partir de um simulador de processos químicos e petroquímicos.

Os resultados obtidos são apresentados de forma comparativa com o PLS estático padrão. Como critério de avaliação, é abordada a habilidade do modelo na predição de variáveis.

II. METODOLOGIA

A. PLS

O PLS, também chamado de Projeção em Estruturas Latentes, é constituído de dois blocos de dados: X e Y, onde X é uma matriz (n x p) com n observações e p variáveis de entrada do sistema e Y, é uma matriz (n x q) com n observações e q variáveis de saída ou respostas associadas ao sistema.

O método PLS caracteriza-se pela redução da dimensão de medidas, definindo um número de novas variáveis que

sumarizam a variância (informação) relevante dos blocos *X* e *Y*. Estas novas variáveis são uma combinação linear das variáveis originais. Como resultado ocorre uma predição de *Y* com base em *X*, com um estimador linear da forma (1):

$$Y = XC + resíduos \quad (1)$$

Para a aplicação do algoritmo PLS pode ser necessária escalar os blocos de dados *X* e *Y*, pois os dados experimentais originais podem não ter uma distribuição adequada para a análise (por exemplo, medidas em diferentes unidades). Neste caso, um pré-processamento [5], [6] e [7], nos dados originais pode ser de grande valia, consistindo basicamente em autoescalar os dados. Autoescalar significa centrar os dados na média e dividi-los pelo respectivo desvio padrão de cada variável.

Estas informações escaladas são armazenadas em matrizes de desvio padrão S_X para os dados de *X* e S_Y para os dados de *Y*. Os blocos *X* e *Y* escalados, são então decompostos como uma soma de séries de matrizes com *rank* 1.

Para que se possa compreender melhor o algoritmo PLS dinâmico, a fig. 1 (baseada em [3] [4]) mostra o processo de obtenção dos *scores* e *loadings* de *X* e *Y*.

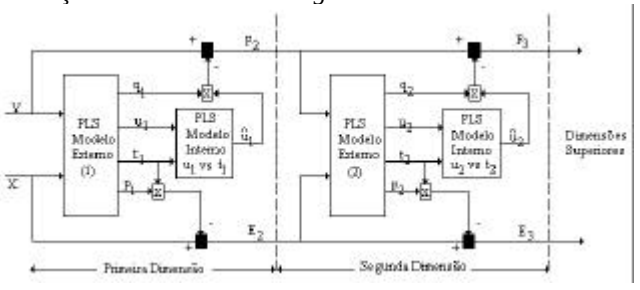


Fig. 1. Procedimento do algoritmo PLS

O primeiro grupo de vetores de *loadings* (pesos), p_1 e q_1 é obtido com maximização de covariância entre os dados de *X* e *Y*. A projeção dos dados de *X* e *Y*, respectivamente em relação a p_1 e q_1 dão o primeiro grupo de vetores de *scores* t_1 e u_1.

O procedimento de determinação dos vetores de *scores* e *loadings* é seguido com os resíduos *E* e *F* armazenados em cada estágio. No início do processamento do PLS $E_1 = X$ e $F_1 = Y$, (ver fig. 1).

Pode-se escrever agora indiretamente as matrizes *X* e *Y* em (2) e (3) através de seus scores com o modelo interno que são justamente uma regressão linear de t_1 e u_1 que pode ser interpretado como parte dos dados de *Y* que estão sendo preditos com a primeira dimensão PLS.

$$X = t_1 p_1' + t_2 p_2' + \cdots + t_n p_n' + E = TP' + E \quad (2)$$

$$Y = u_1 q_1' + u_2 q_2' + \cdots + u_n q_n' + F = UQ' + F \quad (3)$$

onde *T* e *U* representam os *scores*, *P* e *Q* representam os *loadings* para os blocos *X* e *Y*. Os resíduos são dados por *E* e *F*.

B. PLS dinâmico

A técnica estatística multivariável PLS é baseado em modelos lineares, para tratar dados lineares. Quando se tem não linearidade nos dados, o modelo PLS torna-se ineficiente.

Para tratar não linearidade nos dados, ou seja, para tratar dados de sistemas não lineares, algumas técnicas trazem a adição de fatores extras ao PLS, ou seja, funções que retratem não linearidades. Estas funções adicionais são empregas na modelagem das relações internas do algoritmo PLS.

Algumas das técnicas com o objetivo de tratar as não linearidades nos dados são:

- PLS *spline* [8] – Utiliza-se de uma função *spline* (quadrática ou cúbica) para estabelecer a maior covariância entre os *scores* de *X* e *Y*. Tendo excelentes propriedades de aproximação de uma função contínua.

- PLS Neural [9] – Utiliza-se das redes neurais (técnica de modelagem para dados não-linear) para estabelecer as relações internas no algoritmo PLS, entre os *scores* de *X* e *Y*.

- *Dynamic* PLS [3] [4]:

a) Utiliza-se das estruturas de Hammerstein (técnica de modelagem para dados não-linear) para estabelecer as relações internas no algoritmo PLS, entre os *scores* de *X* e *Y*.

b) Utiliza-se dos modelos de séries temporais (ARX – A*uto-Regressive with eXogenous inputs*) para estabelecer as relações internas no algoritmo PLS, entre os *scores* de *X* e *Y*.

A metodologia do PLS dinâmico a ser utilizada para a obtenção do modelo de inferência é baseada em [3] [4] com uso dos modelos de séries temporais. O algoritmo PLS dinâmico é baseado na modificação direta do algoritmo interno do PLS (ver fig.1). Em vez de relacionar os *scores* de entrada e saída (*t* e *u*) usando modelos estáticos lineares ou não lineares, utiliza-se um modelo ARX.

A partir de (1), obtém (4):

$$Y = XC_{din} + resíduos \quad (4)$$

A diferença significativa entre (1) e (4), reside no fato que o *C* é uma matriz constante em (1), em (4) *C* é C_{din}, uma

matriz que retrata um mapeamento dinâmico, não variante no tempo, das entradas manipuladas e as saídas controladas.

Assim, a analogia dinâmica de (3) é dada por (9):

$$Y = G_1(t_1)q_1' + G_2(t_2)q_2' + \cdots + G_n(t_n)q_n' + F \qquad (9)$$

onde G é o modelo dinâmico linear (um modelo ARX) identificado entre os *scores* de X e Y para cada variável latente.

Uma comparação importante entre o PLS dinâmico e o PLS estático padrão, é que o PLS estático é caracterizado pela habilidade em reduzir a dimensionalidade do espaço de medidas, removendo redundâncias no conjunto de dados. Contudo, esta característica do PLS não é utilizada no esquema de modelamento dinâmico proposto, devido ao fato que na identificação empírica do modelo todas as informações de entrada são utilizadas na construção do modelo.

A obtenção do modelo PLS dinâmico difere da obtenção do PLS estático. Inicialmente os dados são processados com o PLS estático para a obtenção dos *scores* e *loadings*. Posteriormente, tenta-se achar um modelo que descreva a relação matemática dinâmica entre os *scores* de X e de Y. Segundo [3] e [4], o melhor modelo que descreve esta relação é um modelo ARX.

A identificação do comportamento dinâmico é feita entre o 1º vetor de *scores* de X e 1º vetor de *scores* de Y, ou seja, entre t_1 e u_1, seguindo a identificação até que a última variável latente seja identificada. No caso da coluna de debutanização foram identificados 15 modelos ARX (existem 15 variáveis latentes). A exemplificação do processo de identificação por ARX pode ser mais bem compreendido através da fig. 2.

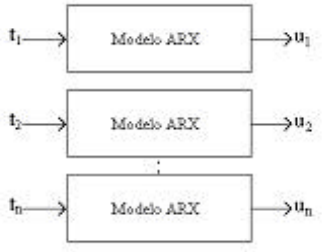


Fig. 2. Identificação dos modelos ARX entre os loadings de X e Y.

A necessidade de se obter modelos ARX adequados, ou seja, que retratem matematicamente u, a partir de t, é uma condição elementar para o sucesso do PLS dinâmico.

O critério de avaliação dos modelos ARX identificados foi o erro quadrático médio de predição (RMSEP (*Root Mean Squared Error of Prediction*)).

O esquema de obtenção do modelo de inferência por PLS dinâmico é dado na fig. 3, baseado em [3] e [4].

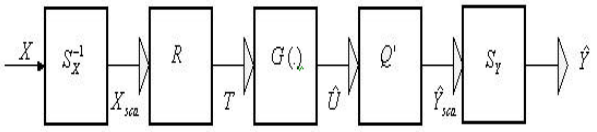


Fig. 3. Esquema do modelamento do PLS dinâmico

onde:

- S_X e S_Y são as matrizes de desvio padrão de X e Y respectivamente;
- X é a matriz de dados de entrada do sistema;
- R é o peso de compressão ($R^{-1} = P'$);
- T é a matriz de *scores* de X;
- G é o elemento dinâmico identificado em cada dimensão PLS;
- $\hat{U}$ indica os *scores* de Y preditos a partir de G;
- Q' é a matriz transposta do *loadings* de Y;
- *sca* significa que os dados da planta são escalados;
- $\hat{Y}$ indica os valores preditos na saída do modelo.

Uma explicação completa destes procedimentos de identificação por modelos ARX e a modelagem do PLS dinâmico pode ser encontrado em [4] e [10].

C. Modelagem e simulação

A modelagem da coluna de destilação debutanizadora bem como os dados de simulação do processo foram obtidos através de um simulador de processos químicos e petroquímicos. Os dados foram obtidos em malha fechada (malha fechada significa que houve controladores atuando para que as variáveis de saída se mantivessem dentro dos padrões), aplicando diversas formas de perturbações. As variáveis coletadas são as temperaturas de todos os pratos (15 ao total) como variáveis de entrada (X) e a fração molar do componente principal dos produtos de topo e fundo como variável de saída (Y).

A simulação do processo de debutanização, para a coleta de dados, foi dividida em duas etapas: calibração (construção do modelo) e predição (validação do modelo). Para a etapa de calibração foram coletadas 1748 amostras do processo e para a etapa de predição 1401 amostras.

A composição das correntes de saída dos produtos de topo e fundo, correspondentes aos produtos GLP e Nafta respectivamente são dados na tab. 1.

A modelagem da coluna de debutanização pode ser mais bem entendida analisando a fig. 2, onde é possível observar a construção da matriz de dados de *X*, com as temperaturas dos pratos da coluna, e de *Y* com a fração molar dos produtos de topo e de fundo.

Com base nos dados da tab. 1 é possível compreender melhor o uso dos compostos n-butano e i-pentano para representar as correntes de saída Nafta e GLP respectivamente.

Todos os componentes (do propano ao n-octano) estão presentes, tanto na fração de topo quanto na fração de fundo da coluna. O n-butano é o composto pertencente ao produto GLP, contudo também é encontrado na Nafta e possui a maior fração molar entre os compostos do GLP que estão presentes na Nafta. O mesmo acontece com o i-pentano, composto pertencente ao produto Nafta que é encontrado também no GLP e possui a maior fração molar entre os compostos da Nafta que estão presentes no GLP.

TABELA I
COMPOSIÇÃO DAS CORRENTES DE SAÍDA EM ESTADO ESTACIONÁRIO - TOPO E FUNDO

	Componente	Fração topo	Fração fundo
Principais produtos do GLP	Propano	0.0311	$0.2749e^{-5}$
	i-Butano	0.2689	$0.1365e^{-2}$
	n-Butano	0.2748	**$0.5114e^{-2}$**
	i-Buteno	0.1641	$0.2367e^{-3}$
Principais produtos da Nafta	i-Pentano	**0.1363**	0.1479
	n-Pentano	0.1242	0.2798
	n-Hexano	$0.6720e^{-3}$	0.1977
	n-Heptano	$0.1126e^{-4}$	0.2033
	n-Octano	$0.1345e^{-6}$	0.1646
	Total	1.0000	1.0000

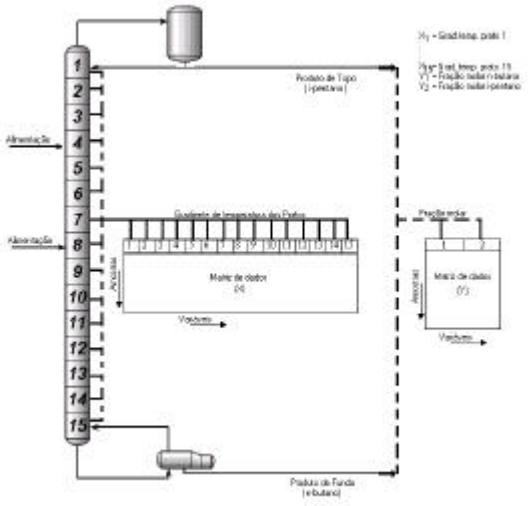


Fig. 2. modelagem da coluna de debutanização

III. RESULTADOS

Os resultados obtidos com o PLS dinâmico são apresentados de forma comparativa com o PLS estático padrão.

Os dados coletados do simulador de processos químicos e petroquímicos foram divididos em duas etapas: etapa de calibração e etapa de predição. Os dados da etapa de calibração foram utilizados para a construção do modelo de predição PLS dinâmico, e os dados da etapa de predição foram utilizados para a validação do modelo obtido.

Inicialmente, fez-se o pré-processamento dos dados e o procedimento padrão para a obtenção das matrizes de desvio padrão, *scores* e *loadings* dos dados de *X* (temperatura dos pratos = 15) e *Y* (fração molar do produto de topo - i-pentano e produto de fundo - n-butano) e posteriormente a identificação dos 15 modelos ARX (existem 15 variáveis latentes).

As matrizes de S_X, S_Y, *P*, *Q*, *R* e *G* (com todos os 15 modelos ARX identificados) estão disponíveis em [10].

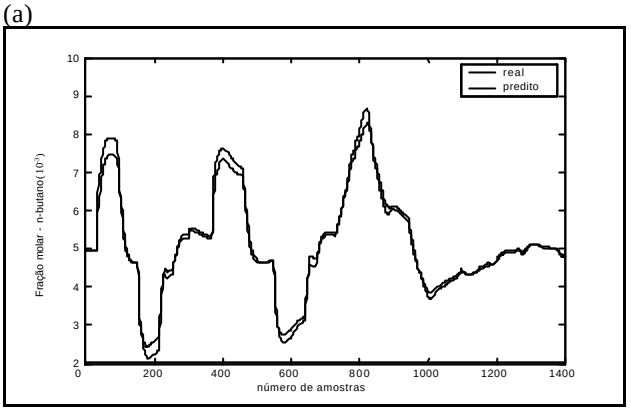
Os resultados obtidos com o modelo PLS dinâmico para os produtos de fundo (fração molar do n-butano) e de topo (fração molar do i-pentano) são demonstrados na fig. 5 (a) e fig. 5 (b) respectivamente.

Os resultados obtidos, de forma comparativa para os mesmos dados (fração molar do produto de fundo e produto de topo) e as mesma condições de simulação, com o PLS estático padrão são demonstrados na fig. 6 (a) e fig. 6 (b) respectivamente.

A fig. 7 mostra os erros de predição para ambos os modelos PLS dinâmico e o PLS estático padrão para os

produtos de fundo (fração molar do n-butano) e topo (fração molar do i-pentano).

(a)



(b)

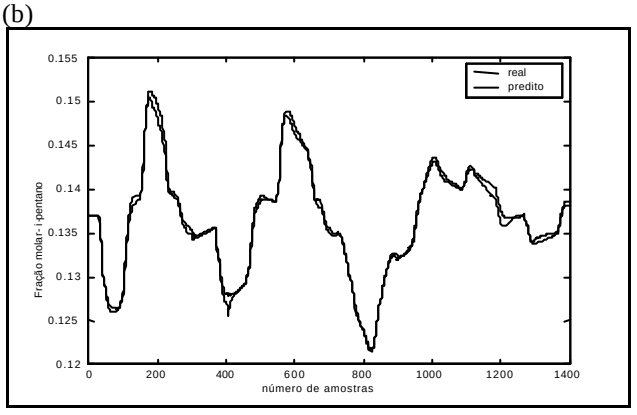
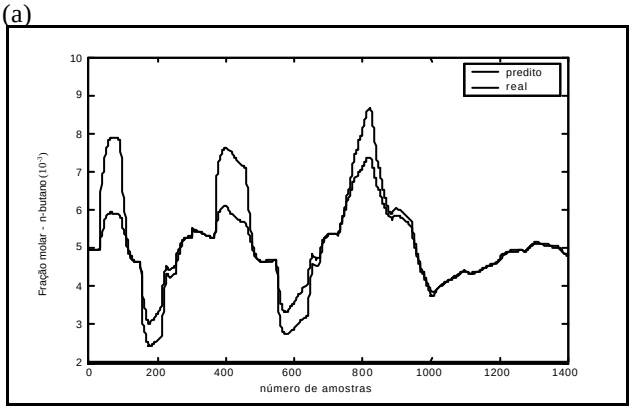


Fig. 5. Resultados obtidos com PLS dinâmico das correntes de saída: (a) produto de fundo e (b) produto de topo

(a)



(b)

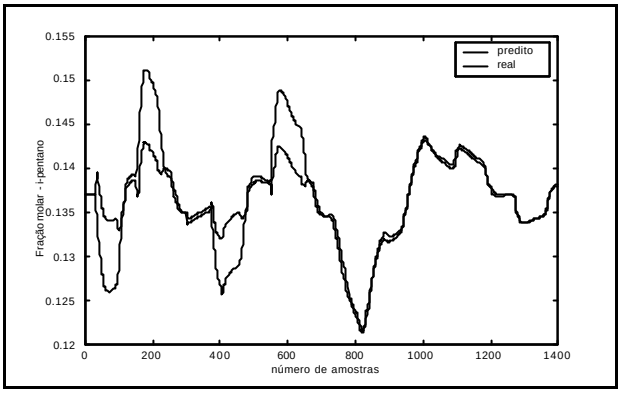


Fig. 6. Resultados obtidos com PLS estático das correntes de saída: (a) produto de fundo e (b) produto de topo

Conforme visto nos resultados, para as mesmas condições de teste, o PLS dinâmico mostrou maior habilidade do que o PLS estático padrão na predição de variáveis, visto pelas fig. 5 (a) e 5 (b), 6 (a) e 6 (b), verificados também na fig. 7, onde encontram-se os erros de predição (RMSEP) dos produtos de topo e fundo.

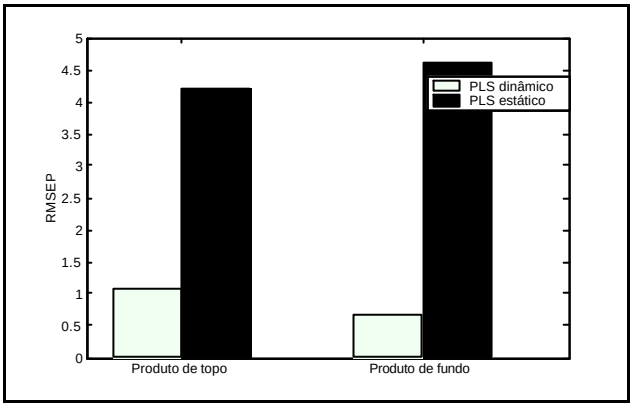


Fig. 7. Resultado dos erros de predição para o PLS dinâmico e o PLS estático padrão para os produtos de topo (i-pentano) e fundo (n-butano)

O melhor resultado do PLS dinâmico frente ao PLS estático é devido ao modelo PLS dinâmico possuir a habilidade de adicionar ao modelo informações ou particularidades do processo (informações do comportamento dinâmico do sistema) que foram identificadas com os modelos ARX. O PLS estático padrão não possui tal habilidade, por se tratar de um modelo linear baseado apenas nas informações estáticas obtidas na construção do modelo.

AGRADECIMENTOS

Os autores agradem o apoio financeiro da Agência Nacional do Petróleo - ANP - e da Financiadora de Estudos e Projetos - FINEP - por meio do Programa de Recursos Humanos da ANP para o Setor Petróleo e Gás - PRH-ANP/MCT (PRH10-CEFET-PR)

REFERÊNCIAS

[l] L. Ljung, *System Identification: Theory for the user*. Prentice-Hall, NJ: 1987, p.143.

[2] H. Wold, "Estimation of principal components and related models by iterative least squares". *Multivariate Analysis Academic Press, NY*, 1966, pp. 391-420.

[3] S. Lakshminarayanan, S.L. Shah and K. Nandakumar, "Modeling and Control of Multivariable Processes: Dynamic PLS Approach." *AIChE Journal*, vol.43, 1997, pp. 2307-2322.

[4] S. Lakshminarayanan, "Process characterization and control using multivariate statistical techniques", PhD thesis, University of Alberta, Canada, 1997.

[5] H. Martens and T. Naes, *Multivariate Calibration*. John Wiley & Sons, New York: 1989, p. 56.

[6] E.R. Malinowski, *Factor Analysis in Chemistry*. John Wiley & Sons, New York: 1991, p. 44.

[7] K.R. Beebe, R.J. Pell and M.B. Seasholtz, *Chemometrics: A Practical Guide*, Wiley, New York: 1998, p.112.

[8] S. Wold, "Nonlinear Partial Least Squares Modeling II: Spline Inner Relation". *Chemometrics and Intelligent Laboratory Systems*, vol.14, 1992, p71-84.

[9] S.J. Qin and T.J. McAvoy, "*Nonlinear PLS Modelling Using Neural Networks*". *Computers and Chemical Engineers*. Vol.16(4), 1992, p379-391.

[10] R. Stachiw, Inferência de Variáveis na Indústria do Petróleo com PLS dinâmico. Tese de mestrado. CEFET/PR, Brasil, 2004.