APLICAÇÃO INDUSTRIAL DA TÉCNICA DE ANÁLISE COM COMPONENTES PRINCIPAIS (PCA) UTILIZANDO SISTEMAS PIMS[1]

Ludmila Rodrigues Fernandes [2]
Constantino Seixas Filho [3]
Alessandra Rezende Esteves [4]
Artur Patitucci Sobroza [5]

Resumo
Este trabalho tem como objetivo apresentar a técnica estatística de Análise em Componentes Principais aplicada ao Sistema PIMS. A técnica PCA apresenta um enorme potencial em tratamento de um alto volume de dados, tipicamente disponibilizados por estes tipos de sistemas. Através da modelagem de um processo industrial é possível identificar de forma rápida e eficiente as anomalias presentes neste, bem como suas causas. Neste artigo serão apresentados inicialmente os conceitos básicos envolvidos no entendimento desta técnica de análise de dados. Em seguida, será apresentado um caso real de aplicação da técnica de PCA para análise dos dados provenientes de um processo em eletrólise.
Palavras-chave:.PIMS; Análise estatística de dados; PCA; Eletrólise.

INDUSTRIAL APPLICATION OF THE PRINCIPAL COMPONENT ANALYSIS TECHNIQUE (PCA) USING PIMS SYSTEMS

Abstract
This article presents Analysis statistical techniques in Principal Components applied to PIMS System. PCA techniques shows an immense potential in processing of high data volume, typically available in these systems. Through industrial process modeling is possible to identify, in an easy and efficient way, the anomalies of this process as well as their causes. In this work will be presented initially the basic concepts involved in the data analysis techniques. Next, will be presented a real case of the PCA techniques application for data analysis from electrolysis process.
Key words: PIMS; Data statistical analysis; PCA; Electrolysis.

[1] *Contribuição técnica ao XI Seminário de Automação de Processos, 3 a 5 de outubro, Porto Alegre-RS*
[2] *Engenheira de Controle e Automação da ATAN Ciência da Informação LTDA.; ludmila.fernandes@atan.com.br – tel (31) 3289-7728*
[3] *Membro do Conselho Editorial da Revista InTech, Diretor de P&D da ATAN Ciência da Informação, Professor assistente do Departamento de Engenharia Elotrônica da UFMG; constantino.seixas@atan.com.br – tel (31) 3289-7728*
[4] *Engenheira de Controle e Automação da ATAN Ciência da Informação LTDA.; alessandra.esteves@atan.com.br – tel (31) 3289-7728*
[5] *Engenheiro Eletricista, Gerente do Departamento de PIMS da ATAN Ciência da Informação LTDA.; artur.sobroza@atan.com.br – tel (31) 3289-7727*

1 INTRODUÇÃO

1.1 Ferramentas Estatísticas de Análise de Dados

A análise estatística multivariada possui para uma larga faixa de aplicações na supervisão e modelagem de processos industriais. Seu uso para aplicações industriais ainda é muito pouco comum, primeiro por se tratar de tecnologia pouco disseminada e depois por exigir uma ferramenta extra ainda não incorporada aos sistemas industriais mais conhecidos como SDCD, SCADA, PIMS ou MES. Seus benefícios entretanto estão bem alinhados ao conceito vigente de excelência operacional em que se busca conduzir um processo a um ponto operacional ótimo e mantê-lo operando o mais próximo possível desta condição. As duas técnicas mais utilizadas em processos industriais são o PCA – Principal Components Analysis, ou Análise com Componentes Principais e o PLS – Partial Least Squares.

A primeira técnica é utilizada na monitoração estatística multivarada de processos (*MSPM - Multivariate Statistical Process Monitoring*). A segunda técnica (PLS) é usada para inferir valores de variáveis a partir das variáveis medidas, implementando soft-sensors. Os soft sensors podem ser utilizados como back-up de instrumentos ou de procedimentos realizados off-line no laboratório de processo ou para substituí-los completamente. Os resultados calculados podem ser utilizados como variáveis de processo em algoritmos de controle ou por exemplo, para decidir se um processo alcançou o ponto de final de reação. Tanto variáveis de processo como variáveis de qualidade podem ser inferidas com o uso destes algoritmos. Este artigo foca apenas a MSPM.

1.2 Monitoração Estatística Multivariada de Processos

A idéia de se usar um modelo estatístico ao invés de um modelo fenomenológico se dá pela dificuldade de se construir um modelo preciso do processo e pela falta de sensores on-line para medir variáveis de importância do processo, principalmente as variáveis de qualidade. Modelos baseados em redes neurais são mais fáceis de se produzir, mas requerem data sets específicos para treinar a rede em cada tipo de anormalidade a fim de conseguir emitir diagnósticos confiáveis. Se o processo não for contínuo, mas de batelada, as dificuldades aumentam devido à duração finita deste tipo de processo e à ausência de um estado de regime.

O método de componentes principais é chamado de um método projetivo porque projeta o espaço original de variáveis num espaço de dimensão mais reduzida.

Inicialmente o processo é modelado estatisticamente a partir das variáveis medidas que retratam entradas e variáveis de estado de um processo. O modelo produzido pelo PCA apresenta uma forte redução do número de variáveis, trocando a base definida pelas variáveis medidas inicialmente, por uma outra base ortogonal gerado pela combinação linear destas variáveis, que maximiza a variância das medidas realizadas. Estas novas variáveis são denominadas de componentes principais ou variáveis latentes porque são intrínsecas ao processo e explicam o comportamento do sistema melhor que o conjunto de variáveis inicial. Das N componentes principais encontradas apenas as R componentes que explicam a maior parte da variabilidade são retidas no modelo. A primeira componente principal é que retém a maior parte da variabilidade dos dados, retirada esta componente é

escolhida a segunda que retém a maior variabilidade dos dados remanescentes e assim por diante. Em geral o ruído das medidas é rejeitado junto com as componentes de menor importância do modelo. A simples redução deste número de variáveis já é um benefício de muito valor desta metodologia.

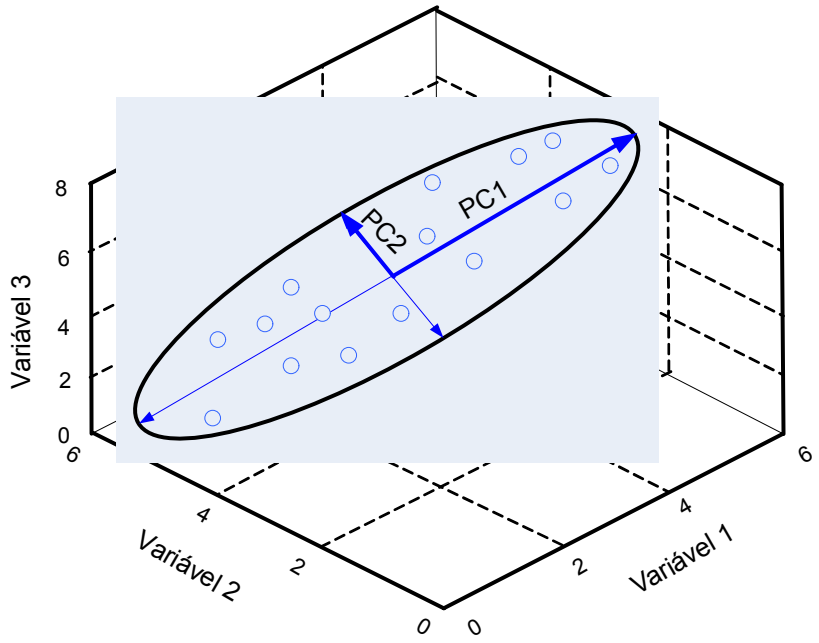

Figura 1 – Espaço inicial formado por três variáveis e novo espaço gerado pelas componentes principais PC1 e PC2 – Melhor explicação para variância dos dados

Os passos a serem seguidos na monitoração estatística são:

- Construção de um modelo estatístico a partir de um conjunto de dados característico do processo operando corretamente, isto é, produzindo produtos com a qualidade desejada. O PIMS é uma ferramenta fundamental nesta fase, produzindo as séries históricas necessárias a esta análise.
- Construção de um índice de controle estatístico que permita monitorar o processo e detectar condições anormais. Os principais indicadores para este fim são as estatísticas de Hotelling ou índice T2 e a estatística Q também chamada de Squared Prediction Error (SPE).
- Identificação de quais variáveis contribuíram mais fortemente para o índice que revelou o estado de anomalia através de diagramas de contribuição (contribution plots).
- Diagnóstico da situação de anormalidade através da análise dos diversos gráficos gerados e do conhecimento fenomenológico do processo sendo controlado.

A projeção de um ponto na nova base formada pelos componentes principais se denomina score. Os scores podem ser exibidos para cada duas componentes principais em um gráfico denominado score plot. Pontos perto da origem (média estatística das variáveis) denotam uma operação normal. Os limites de confiança são representados por elipses.

Dois índices são utilizados para detectar distúrbios, o índice de Hotelling (T2) e o índice Q. O índice ou estatística de Hotelling mede a distância estatística de um ponto ao centro da elipse. Se T2 for maior que um o score está fora da elipse que denota uma confiança de 99%. Isso significa que o processo está sob distúrbio. O

índice Q por sua vez mede a distância de um ponto ao plano que representa o modelo gerado pela nova base. Se o índice Q for grande isso significa que o modelo levantado para o processo não mais serve para representá-lo e deve ser recalculado.

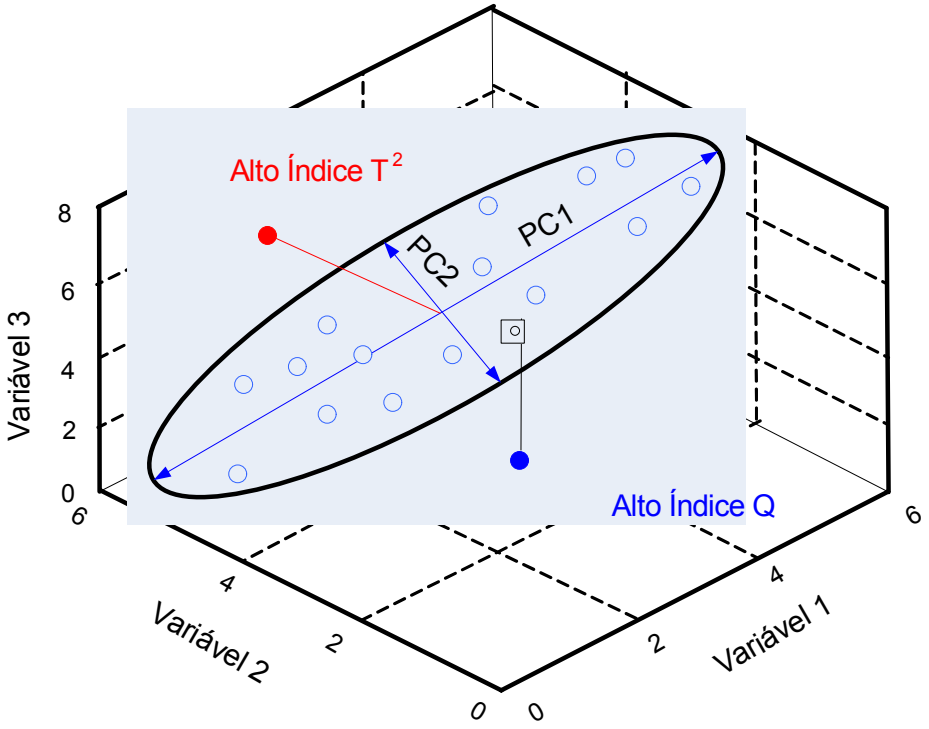

Figura 2 – Interpretação geométrica de Q e T2

Dos gráficos gerados os mais importantes são:

Score Plots

Servem para examinar que variáveis estão fugindo ao controle. Estes gráficos sempre apresentam como eixos duas componentes principais de cada vez, para análises bidimensionais. Se por exemplo tivermos três componentes principais P1, P2 e P3, existirão três gráficos de score: P1 x P2, P1 x P3 e P2 x P3. No caso de um sistema de bateladas os score plots permitem detectar que uma batelada está sofrendo distúrbios e que portanto poderá fornecer um produto fora de especificações muito antes da batelada se completar. Isso traz um grande ganho e permite que uma análise seja conduzida para descobrir a condição de distúrbio e reconduzir o processo ao seu estado de controle. É possível marcar regiões em um score plot denotando padrões de distúrbios previamente identificados por especialistas no processo, se tornando uma boa ferramenta para diagnósticos.

Loading Plots ou Gráficos de Carregamento

Se duas medidas estão próximas no gráfico de carregamento elas estão altamente correlacionadas. Se estiverem distantes a correlação entre elas é baixa.

T2 Plot

É uma carta de controle mostrando se o índice T2 se encontra dentro da faixa permitida. Se ultrapassar o valor 1 estão o processo está sofrendo um distúrbio que deverá ser analisado.

Q Plot

Também é uma carta de controle estatístico. Um valor alto de Q indica que ao projetar uma determinada amostra no plano do modelo de componentes principais, houve um grande erro devido às variáveis não preservadas do modelo. É comum que num sistema contendo cerca de 20 variáveis medidas apenas 3 componentes principais sejam mantidas por representarem um grande percentual da variabilidade do sistema num dado momento. Entretanto após algum tempo as características estatística do processo podem mudar e uma variável que não apresentava nenhum peso passa a influir mais no processo. Neste caso Q aumenta e deve ser interpretado.

Contribution Plots ou Diagramas de Contribuição

Mostram como as variáveis originais medidas participam da composição dos scores e dos índices de desempenho T2 ou Q e quais variáveis devem ser apontadas como as maiores responsáveis pelo seu aumento. Esta análise ajuda a encontrar os culpados pelos distúrbios e propicia uma ação corretiva.

MSPM traz como benefício não apenas um aumento da qualidade, mas um aumento da segurança, redução de custos e aumento da eficiência dos ativos de produção.

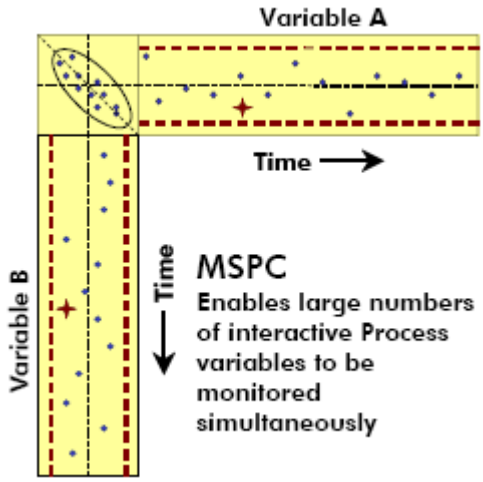

Figura 3 – Multivariate Statistical Process Monitoring

2 DESENVOLVIMENTO

2.1 Aplicação da Análise em Componenetes Principais

Torna-se cada vez mais notável, em usinas metalúrgicas de todo o mundo, um crescente e rigoroso controle de qualidade sobre o produto final. Um determinado tipo de metal a ser produzido carrega consigo um conjunto de especificações que não devem e nem podem deixar de ser cumpridas. E o alcance de resultados finais tão desejáveis pode ser obtido apenas mediante uma operação muito bem sintonizada de toda a linha de produção.

Além de condição fundamental para obtenção de qualidade, uma operação adequada pode conduzir, ainda, à cada vez mais almejada redução de custos. Portanto, um processo sob controle é fator indispensável na produção de metais e o único meio para alcançá-lo é conhecer profundamente todos os eventos que o conduzem.

Todas as informações que regem a produção e são responsáveis pelos seus resultados, sejam eles bons ou ruins, estão contidas em suas variáveis de processo e, definitivamente, estas não são poucas. Avanços na área de Instrumentação têm gerado um altíssimo volume de medições disponíveis, contemplando praticamente a totalidade de tudo aquilo que determina o comportamento do processo. Complementando o cenário de facilitação de acesso a todas estas medidas, os sistemas PIMS (Plant Information Management System), responsáveis por concentrá-las e armazená-las, encontram-se altamente difundidos nas usinas metalúrgicas, e compõem uma robusta e consistente base de dados. Ou seja, a tecnologia de hardware e software já faz a sua parte, disponibilizando toda e qualquer informação necessária, mas para alcançar os objetivos principais, ainda é preciso difundir um outro tipo de tecnologia, a Engenharia de Dados, responsável pela geração do conhecimento.

De fato, já pode ser notado nas indústrias de metalurgia, destacando-se as siderúrgicas, um significativo aumento de investimento em controle de qualidade, em geral, por meio de ferramentas univariadas de CEP (Controle Estatístico de Processo). Isto é, as variáveis são analisadas de forma individual, o que, comprovadamente, pode conduzir a conclusões insuficientes ou até mesmo errôneas. A causa disto reside no fato de que o comportamento do processo é conseqüência de um conjunto de variáveis e, principalmente, da forma como elas interagem entre si.

Como a oferta de variáveis é muito alta, o que pode tornar a análise simultânea de muitas delas uma tarefa completamente inviável, a técnica PCA, por todas suas características apresentadas, mostra-se como uma excelente solução para análise de dados do processo metalúrgico.

Similar a qualquer outro processo industrial, a metalurgia é composta por variáveis de entrada e saída que, naturalmente, são altamente correlacionadas entre si, favorecendo ainda mais a análise de componentes principais, em substituição às suas variáveis originais.

A seguir, é apresentada uma metodologia proposta para utilização da técnica PCA aplicada a uma processo industrial. Tratam-se de dados reais referentes à eletrólise de cubas de uma usina hidrometalúrgica, para produção de determinado tipo de metal.

2.2 Estudo de Caso

Para a análise dos dados de processo de eletrólise, algumas etapas foram seguidas:

Análise do processo

Uma prévia análise do processo é fundamental para que se obtenha uma massa de dados robusta o suficiente para atuar como modelo operacional. Desta forma, uma análise conjunta de especialistas do processo deve conduzir a um período de dados cujo processamento esteja tão próximo do ideal quanto possível.

O processo exemplificado é composto por várias cubas nas quais o metal deposita-se. A principal variável de interesse para análise do comportamento do processo é a resistência apresentada por cada cuba à medida que o metal é depositado (a corrente é constante e conhecida).

Para este exemplo, foram coletadas 850 amostras (coletadas a uma taxa de 1 minuto) referentes às resistências de 5 cubas, em um período cuja operação era considerada bastante satisfatória. Os dados foram coletados a partir de um sistema PIMS.

Como trata-se de uma análise bastante focalizada, não é considerado um número muito alto de variáveis. No entanto, embora não seja tão significativa a vantagem da redução do número de variáveis a serem analisadas, objetiva-se que o PCA propicie a construção de um modelo operacional e, a partir dele, condições anormais possam ser detectadas.

Normalização dos dados

Para que todas as cubas apresentem a mesma importância na análise, é feita uma prévia normalização dos dados, obtida através da equalização das médias e variâncias, conforme ilustrado a seguir, para as resistências das cubas A, B, C, D e E:

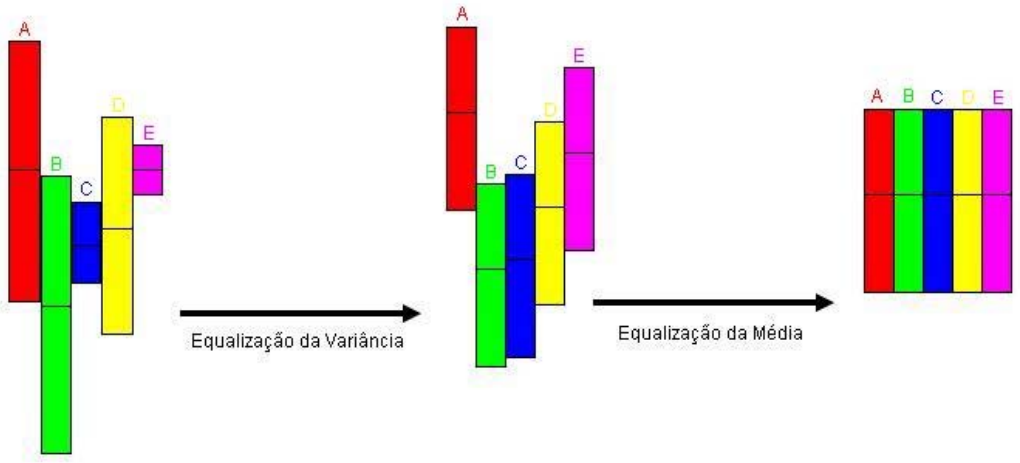

Figura 4 – Normalização de Dados

Construção e análise do modelo

Em seguida, as componentes principais dos dados são calculadas, resultando na construção de um modelo composto por 5 novas variáveis, das quais 2 serão descartadas, a custo de 20% de informações perdidas.

Em uma análise inicial, será utilizado o Loading Plot, que ilustra a relação entre todas as variáveis do processo.

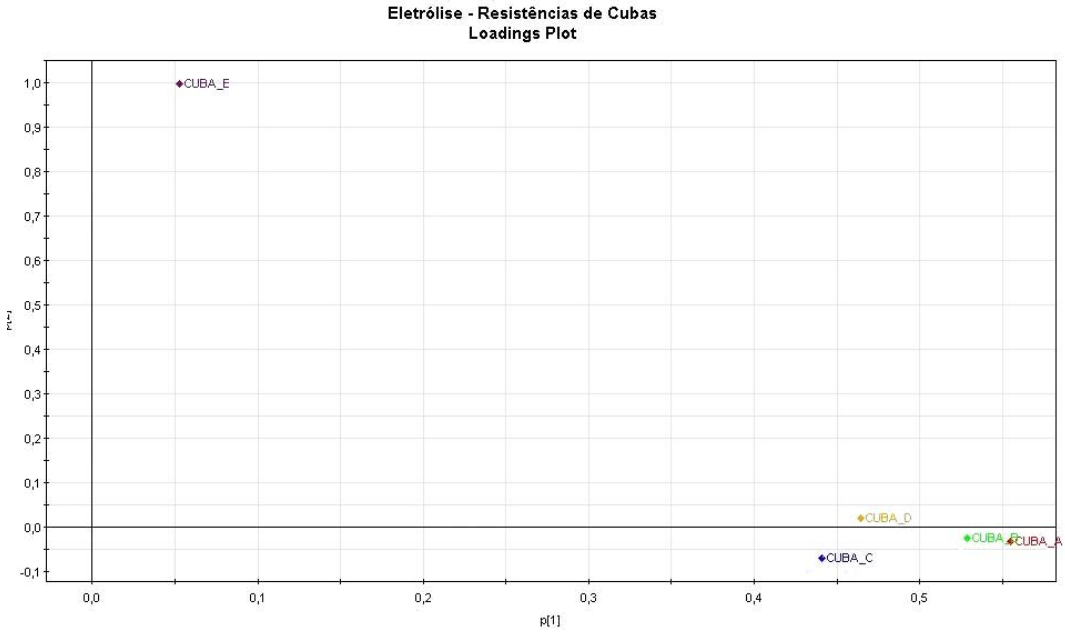

Figura 5 – Loading Plot

Neste gráfico, é possível perceber que as cubas A, B, C e D, por estarem muito próximas, contribuem com informações similares sobre o modelo. Por outro lado, a Cuba E apresenta um comportamento bastante particular, que a diferencia fortemente das demais cubas, o que não é esperado para este tipo de processo.

Este gráfico também indica que a Cuba E apresenta grande influência sobre a segunda componente principal (apresenta alta coordenada neste eixo), enquanto as demais cubas são predominantemente explicadas pela primeira componente.

Para investigação do comportamento da Cuba E, são analisadas as amostras individuais coletadas do processo, através do Scores Plot. Nestes gráficos, cada score representa uma amostra individual projetada no novo espaço de dados definido pelas componentes principais. Um processo em condições normais de operação tende a apresentar um Scores Plot no qual as amostras encontram-se relativamente condensadas, sendo que uma amostra discrepante das demais pode ser considerada uma condição de alarme.

A análise de scores permite a percepção de um padrão de comportamento nos dados e, conseqüentemente, a identificação de amostras que se distanciam de tal padrão. O contorno de uma elipse determina a região do espaço considerada como pertencente às condições normais de operação.

As figuras a seguir ilustram duas visões de um Scores Plot tridimensional:

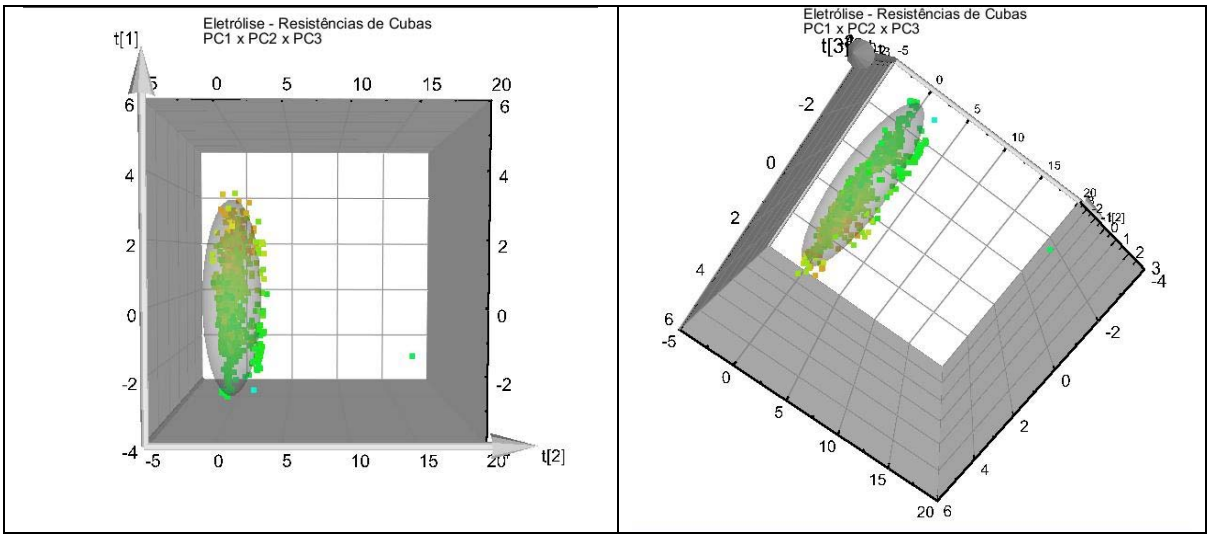

Figura 6 – Scores Plots Tridimensionais

Percebe-se a existência de um ponto excessivamente distante dos demais, o qual merece uma análise mais acentuada. A seguir, são ilustrados:

Um Score Plot bidimensional (PC1 x PC2), no qual é possível perceber com mais clareza o quanto este ponto apresenta uma coordenada alarmante na segunda componente principal.

O Contribution Plot deste ponto, indicando que a variável de maior peso sobre esta amostra é a resistência da Cuba E, conforme o esperado.

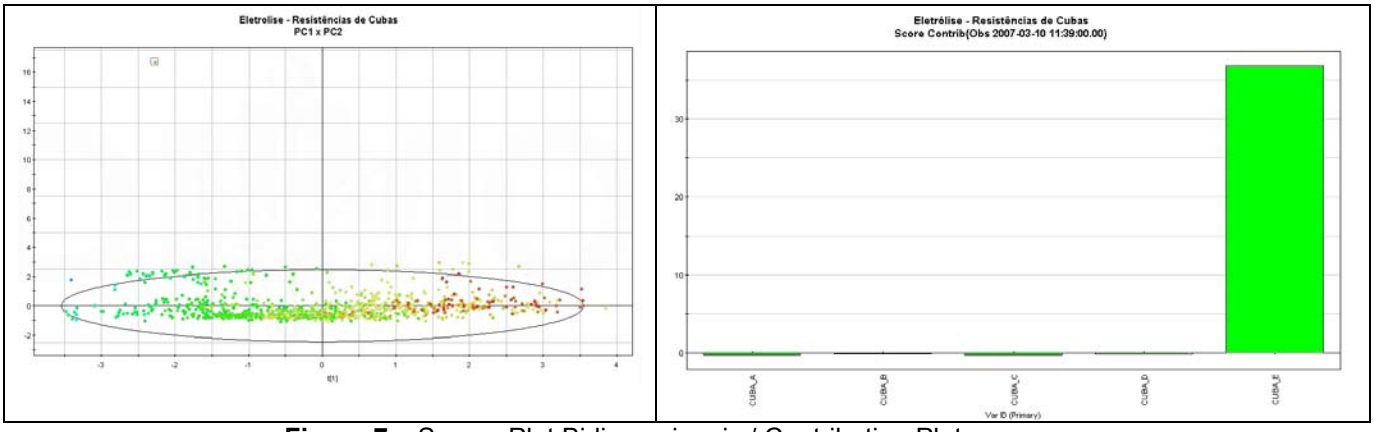

Figura 7 – Scores Plot Bidimensionais / Contribution Plot

Conclui-se que esta amostra apresenta um comportamento anormal quanto à resistência da Cuba E e, para confirmação deste resultado, é apresentado o gráfico de tendência temporal simples desta variável:

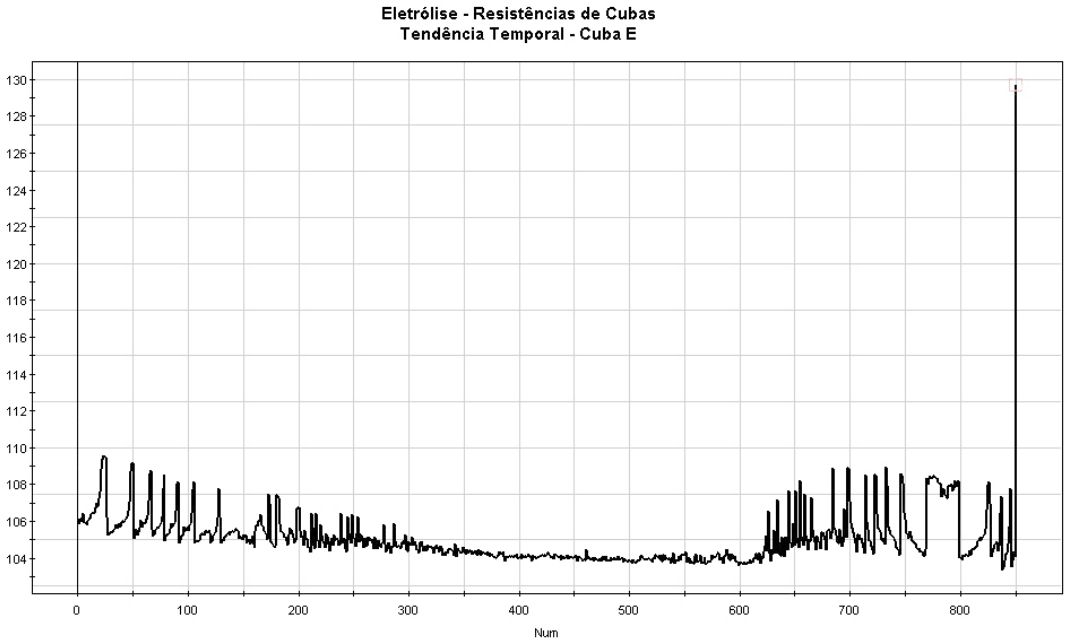

Figura 8 – Tendência Temporal – Cuba E

Claramente, esta amostra constitui um pico de medição que deve ser excluído desta massa de dados para que a confiabilidade do modelo não seja comprometida.

Monitoramento do processo a partir do modelo construído

Uma vez validado o modelo, novos dados (históricos ou em tempo real) podem ser projetados sobre o novo espaço de dados para monitoramento do processo, detecção e diagnóstico de possíveis anomalias.

Para este procedimento, foram utilizados dados posteriores, referentes a um período no qual a instabilidade do processo era sabidamente conhecida. Um novo Scores Plot é gerado:

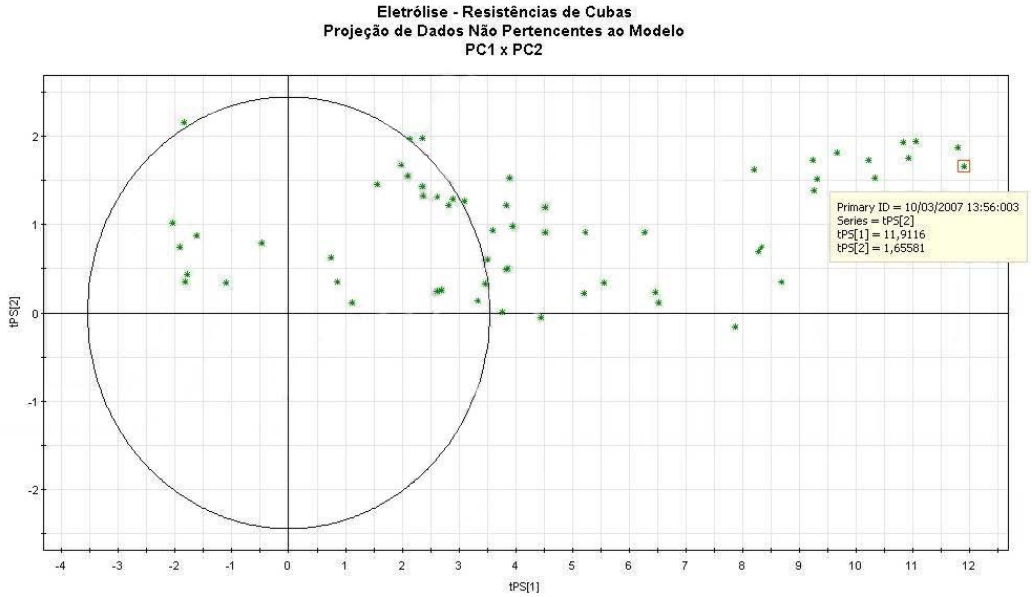

Figura 9 – Scores Plot – Dados Projetados

Percebe-se que, a partir de um certo momento, o processo sai de controle e se desloca na direção da primeira componente, indicando que providências devem ser tomadas para retomada da operação normal do processo.

Para maiores informações, uma amostra particular é analisada mais detalhadamente, através de seu Contribution Plot:

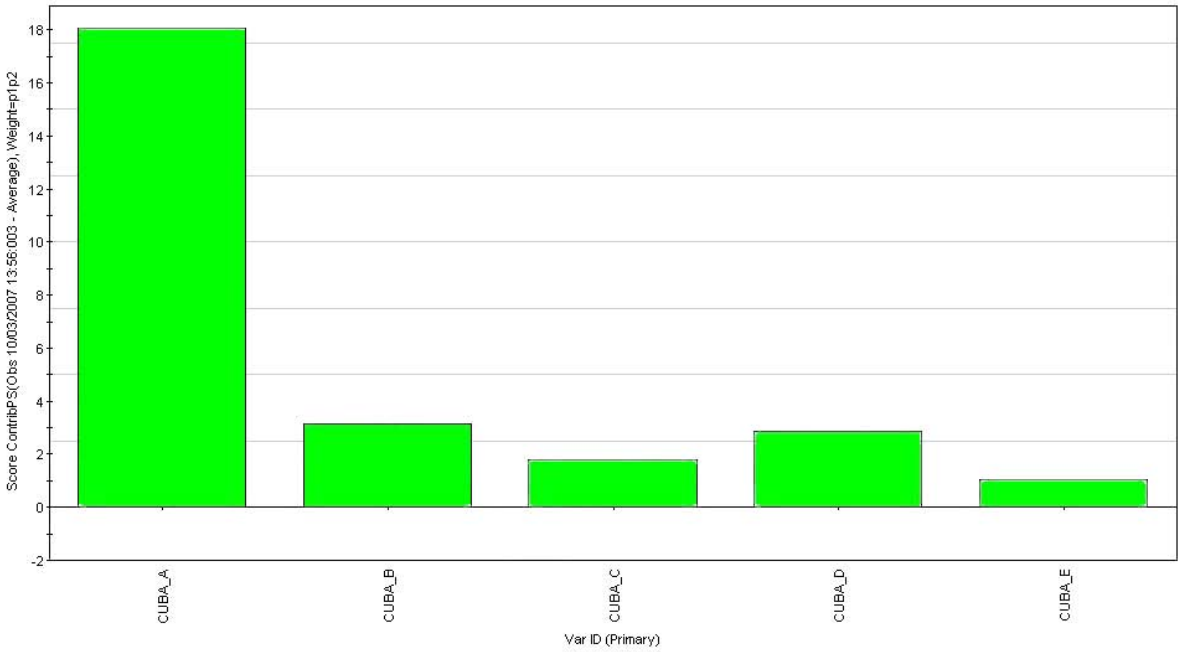

Figura 10 – Contribution Plot

Mais uma vez, analisa-se tendência temporal da variável de maior peso (Resistência da Cuba A):

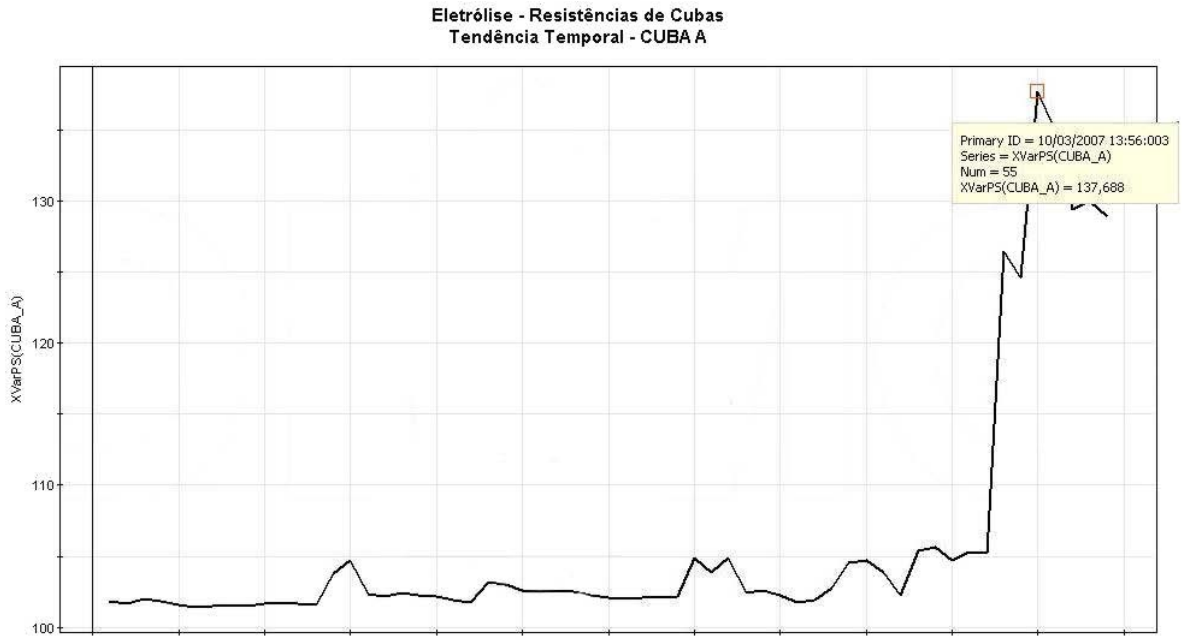

Figura 11 – Tendência Temporal – Cuba A

De fato, a resistência desta cuba apresenta um desvio nítido de comportamento, sendo que a amostra mais alarmante identificada no Scores Plot apresenta-se como um pico no gráfico temporal.

Adicionalmente, podem ser analisados também os índices estatísticos T^2 e Q, indicando, respectivamente, amostras excessivamente distantes da média do modelo e amostras não bem representadas pelo modelo em si.

Ambos apontam um grande número de amostras alarmantes, como era de se esperar.

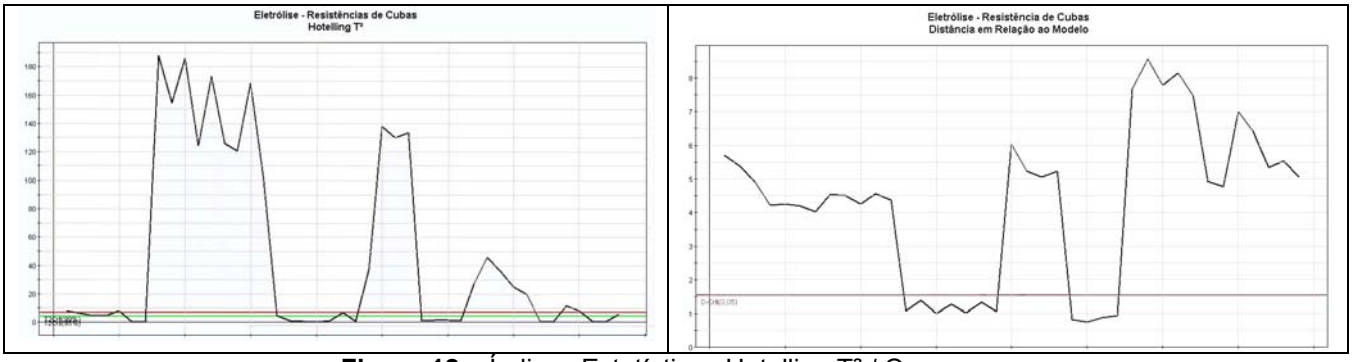

Figura 12 – Índices Estatísticos Hotelling T^2 / Q

3 DISCUSSÃO E CONCLUSÃO

Por meio deste exemplo, é perceptível o quão diversificadas e intuitivas são as ferramentas de diagnóstico oferecidas pela técnica PCA. Particularmente para aplicações de análise e monitoramento do processo metalúrgico, é extremamente útil a existência de uma ferramenta de fácil utilização, que forneça informações claras de forma rápida, possibilitando a tomada de ações corretivas que impeçam ou corrijam a ocorrência de condições indesejáveis.

A análise apresentada ilustra um caso no qual a principal aplicabilidade do PCA é a construção de um modelo para detecção de falhas. No entanto, é importante ressaltar a existência de casos de sucesso nos quais esta técnica apresenta outros focos principais, tais como:

Redução da complexidade análise dos dados

Há exemplos de siderurgia em que 24 variáveis de processo, de diferentes grandezas, são analisadas em conjunto por apenas 4 componentes principais, com aproximadamente 85% de variabilidade capturada.

Conhecimento da interação entre variáveis

Em muitos casos, a principal vantagem da aplicação PCA é o conhecimento de como as distintas variáveis de um processo interagem entre si. Torna-se possível perceber uma correlação desconhecida entre variáveis consideradas independentes e, principalmente, a forma como umas exercem influência sobre as outras.