Universidade Tecnológica Federal do Paraná - UTFPR
Campus Ponta Grossa - Paraná - Brasil
ISSN 1808-0448 / v. 10, n. 01: p. 120-135, 2014
D.O.I: 10.3895/gi.v10i1.1576

Revista Gestão Industrial

SISTEMÁTICA PARA IDENTIFICAÇÃO DAS VARIÁVEIS PREDITIVAS MAIS RELEVANTES EM UM PROCESSO DO SETOR METAL-MECÂNICO

APPROACH FOR PREDICTIVE VARIABLE SELECTION IN A PROCESS OF THE METAL-MECHANICAL INDUSTRY

Marcela Stein[1]; Michel José Anzanello[2]; Alessandro Kahmann[3]
[1]Universidade Federal do Rio Grande do Sul – UFRGS – RS – Brasil
marcela.stein_ctbm@hotmail.com
[2]Universidade Federal do Rio Grande do Sul – UFRGS – RS – Brasil
michel.anzanello@gmail.com
[3]Universidade Federal do Rio Grande do Sul – UFRGS – RS – Brasil
alessandro.kahmann@ufrgs.br

Resumo

As avançadas tecnologias de monitoramento e coleta de dados de processos industriais têm gerado um grande volume de informações que viabilizam a análise do desempenho destes processos. Neste contexto, procedimentos de seleção de variáveis constituem-se em importante recurso para aprimorar o monitoramento e entendimento de tais informações, tipicamente apoiadas em bancos formados por variáveis altamente correlacionadas, ruidosas ou com informação pouco relevante. Esse artigo propõe uma sistemática para identificar as variáveis (indicadores) mais relevantes com vistas à predição dos níveis de formação de sucata em uma empresa do ramo metal mecânico. Para tanto, um modelo de regressão linear múltipla é inicialmente ajustado aos dados normalizados. As variáveis são então sistematicamente removidas com base no valor absoluto do coeficiente de regressão. Após cada eliminação de variável, a capacidade preditiva do modelo é avaliada através das medidas de desempenho Critério de Informação Akaike (AIC) e Soma dos Quadrados dos Erros (SQE). A capacidade preditiva dos modelos resultantes foi considerada adequada por especialistas de processo.

Palavras-chave: seleção de variáveis; sucata; empresa do ramo metal mecânico.

1. Introdução

Com o objetivo de acompanhar o desenvolvimento industrial mundial e mostrarem-se competitivas em um mercado altamente disputado, as empresas buscam garantir a satisfação dos clientes através de reduções de custos produtivos e de otimizações em seus processos e sistemas. A correta compreensão do funcionamento da organização, bem como uma definição apropriada dos aspectos passíveis de melhorias, é de extrema importância à obtenção do sucesso. Assim sendo, as empresas necessitam conhecer minuciosamente a relação entre os dados de entrada e saída do processo, pois é nessa relação que se deflagra a agregação de valor ao produto ou serviço

(BERNARDI et al, 2010). Tal conhecimento permite o desenvolvimento de um planejamento estratégico eficaz com vistas à melhoria do processo, entre outros benefícios (KOBER, 2006).

Com os avanços de tecnologias para criação, monitoramento e análise de dados, os processos produtivos modernos geram grande volume de informações, as quais podem ser utilizadas como base para a análise de seu desempenho. Todavia, o banco de dados resultante da coleta dos mesmos pode apresentar informações ruidosas e pouco relevantes para a empresa (LIU e YU, 2005). Quando o gerenciamento das informações provenientes de um processo não é realizado de forma adequada, podem ocorrer não apenas perdas financeiras, mas também desperdício de recursos nas fases de coleta e análise dos dados menos importantes.

As empresas do ramo metal-mecânico apresentam elevado nível de sucata proveniente dos diversos processos relacionados à sua atividade produtiva. Esse desperdício impacta no custo da qualidade e do produto, reduzindo significativamente a produtividade (WENSING, 2010). Há muitos indicadores e variáveis associados à formação da sucata; no entanto, apenas alguns deles possuem relevância na configuração de um nível elevado de resíduo. Verifica-se, no entanto, limitações acerca da disponibilidade de ferramentas estruturadas com o objetivo de identificar as variáveis que contribuem ativamente para a formação de sucata no processo, e para a consequente perda de produtividade conectada à sucata gerada.

Este artigo propõe um método para identificar as variáveis mais relevantes na formação de sucata em um processo produtivo de uma empresa do ramo metal-mecânico. O processo foi modelado através de regressão múltipla linear, sendo que as variáveis independentes do modelo descrevem indicadores de processo, enquanto o volume de sucata é quantificado por uma variável de resposta. Nas proposições deste artigo, as variáveis independentes foram sistematicamente eliminadas com base na magnitude do coeficiente de regressão, e a precisão de predição do modelo resultante avaliada através de métricas apropriadas. Os modelos gerados apresentaram satisfatória capacidade preditiva, tendo significativamente reduzido o número de variáveis necessárias para predição dos níveis de sucata produzidos.

O artigo está estruturado em cinco seções, além desta introdução. A segunda seção trata dos fundamentos de regressão linear e seleção de variáveis, cujo entendimento é fundamental para a compreensão da metodologia proposta na seção seguinte. A quarta seção traz os resultados obtidos. Na última seção, o estudo traz as considerações finais.

2. Referencial teórico

2.1 Regressão linear

Inúmeros processos produtivos são dependentes de duas ou mais variáveis que se relacionam entre si. Tal interação pode ser modelada por meio de regressões que possibilitam a compreensão da

relação existente entre as variáveis, viabilizando assim ações no processo em análise. Em modelos de regressão existe uma variável dependente (*Y*), também chamada de variável resposta, e *k* variáveis independentes (X_1, X_2,..., X_k), vistas como variáveis regressoras. A interação entre *Y* e *X* pode ser modelada por uma equação matemática, definida como modelo de regressão (RIBEIRO; TEN CATEN, 2000).

Para Montgomery et al (2006), a análise de regressão é uma técnica estatística que busca aproximar da forma mais realista a relação entre variáveis de interesse. As funções resultantes desta técnica são frequentemente baseadas em estudos de física, química, engenharia ou teorias científicas e, por auxiliar tais áreas do conhecimento, são definidas como modelos mecânicos. Além disso, um bom critério de interação entre variáveis depende da coleta adequada dos dados, pois a precisão do modelo gerado apoia-se na qualidade e da confiabilidade dos mesmos. Os principais modos de coleta de dados são dados históricos, estudo de observação e experimento planejado.

Há dois modelos clássicos de regressão linear: regressão linear simples e regressão linear múltipla – sendo esta uma extensão da primeira. Para o enfoque do trabalho, apenas o segundo modelo de regressão será abordado.

2.1.1 Regressão linear múltipla

Na regressão linear múltipla, diversas variáveis independentes (*X*) são responsáveis pela determinação do nível de uma variável de resposta (*Y*), conforme Equação 1 (SARTORIS, 2003).

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \ldots + \beta_k X_{ki} + \varepsilon_i \quad (1)$$

O modelo possui atrelado a si um erro aleatório ε_i, oriundo da diferença entre os *Y* observados e os *Y* gerados pela equação (1); tais erros são independentes e apresentam distribuição normal com média zero e variância σ^2 desconhecida (BARROS et al, 2008). A regressão linear múltipla está relacionada à *k* variáveis e os parâmetros β_i, i=0,1,...,*k*, são denominados coeficientes de regressão. Tais coeficientes são desconhecidos, devendo ser estimados pelos dados amostrais (WERKEMA e AGUIAR, 1996).

Para Simon e Freud (1997), os coeficientes $\beta_0, \beta_1, \ldots, \beta_k$, podem ser estimados através do método dos mínimos quadrados, conforme mostra a Equação 2. A solução desse método pode ser trabalhosa, pois o número de equações a serem resolvidas cresce proporcionalmente ao número de parâmetros estimados. Segundo Barros et al (2008), o método tem como objetivo a minimização da função S com respeito aos coeficientes de regressão. Além disso, o estimador do mínimo quadrado deve satisfazer à equação igualada a zero, conforme as Equações 3 e 4.

$$\Sigma(y - \hat{y})^2 \quad (2)$$

$$S(\beta_i) = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j \, x_{ij}\right)^2 \quad (3)$$

$$\left.\frac{\partial S}{\partial \beta_i}\right|_{\beta_i} = -2 \sum_{i=1}^{n} \left(y_i - \beta_0 - \sum_{j=1}^{k} \beta_j \, x_{ij}\right) x_{ij} \quad (4)$$

Através da utilização do método dos mínimos quadrados, pode-se estimar o valor do vetor dos parâmetros β, como representado na Equação 5.

$$\beta = (x'x)^{-1} x'y \quad (5)$$

Conforme Ribeiro e Ten Caten (2000), para facilitar o ajuste do modelo de regressão linear múltipla, é conveniente utilizar notação matricial, pois os dados, modelos e resultados são exibidos de forma compactada. Tal representação é ilustrada na Equação 6.

$$y = x\beta + \varepsilon \quad (6)$$

$$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ Y_K \end{bmatrix}; \; x = \begin{bmatrix} 1 & x_{11} & x_{12} & \cdots & x_{k1} \\ 1 & x_{21} & x_{22} & \cdots & x_{k2} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_{n1} & x_{n2} & \cdots & x_{kn} \end{bmatrix}; \; \beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{bmatrix}; \; \varepsilon = \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

2.2 Índices de precisão

A utilização de índices de precisão é justificada pela necessidade de quantificar a diferença entre os valores estimados por um modelo e seus valores reais. Uma vez que os modelos de regressão não são perfeitos, torna-se necessário utilizar tais ferramentas para avaliar a adaptabilidade do modelo de regressão. Para o estudo da acurácia de predição da regressão linear múltipla são apresentados três medidores de precisão, descritos a seguir.

2.2.1 Quadrado médio residual (QMR)

Esse indicador tem como finalidade definir um subconjunto de variáveis que minimizem os quadrados médios residuais. Desta forma, espera-se obter uma regressão na qual a adição de novas variáveis não introduza precisão na predição (WERKEM; AGUIAR, 1996). O QMR pode ser calculado conforme a Equação 7.

$$QMR(p) = \frac{SQE(p)}{n-p} \quad (7)$$

SQE - Soma dos Quadrados dos Erros;

p - Número de parâmetros do modelo (*k*+1);

n - Número de observações;

k – Número de variáveis.

2.2.2 Coeficiente de determinação múltipla (R^2)

O coeficiente de determinação global múltipla, definido na Equação 8, é usado como uma estatística global para verificar a qualidade de ajuste do modelo aos dados, medindo a porcentagem

da variação da variável dependente explicada pela regressão. Para modelos de regressão linear múltipla, o R^2 é considerado problemático, pois o coeficiente aumenta quando variáveis regressoras são acrescentadas ao modelo. O R^2 ajustado, por sua vez, penaliza a adição de variáveis desnecessárias na regressão (DOWNING; CLARK, 2006; MONTGOMERY; RUNGER, 2007).

$$R^2 = \frac{SQREG}{SQT} = 1 - \frac{SQE}{SQT} \quad (8)$$

$$R^2_{ajustado} = 1 - \frac{SQE/(n-p)}{SQT/(n-1)} \quad (9)$$

R^2 - Coeficiente de determinação múltipla: varia de 0 a 1;

SQRE - Soma dos quadrados de regressão;

SQT - Soma dos quadrados totais;

p - Número de parâmetros do modelo (*k*+1);

n - Número de observações.

2.2.3 Critério de Informação de Akaike (IAC)

Para Sartoris (2003) e Gujarati (2003), o critério de Akaike é um teste que pode ser usado para comparar modelos, penalizando aqueles que retêm maior número de regressores [ver Equação (10)]. Quanto menor o valor calculado de IAC, melhor será o ajuste do modelo. Por sua vez, Biasoli (2005) afirma que o IAC é a distância entre um modelo verdadeiro e um modelo candidato, fazendo com que quanto menor o critério de informação, mais próximo estará o modelo escolhido do verdadeiro modelo. O IAC pode ser estimado através na Equação 11.

$$IAC = 1 + \ln 2\pi + \ln \frac{SQE}{n} + \frac{2p}{n} \quad (10)$$

$$IAC = -2 \log L + 2p \quad (11)$$

onde log L consiste no logaritmo do máximo da função de verossimilhança.

2.3. Seleção de variáveis

Em processos industriais tipicamente verificam-se centenas de variáveis ruidosas ou correlatas, incluindo temperaturas, pressão, concentração de componentes e tempos de reação, entre outros (ANZANELLO et al, 2009). A seleção de variáveis tem como finalidade identificar as variáveis regressoras que melhor se correlacionam para predizer a variável de resposta (NUNES, 2008). Choi et al (2002) acrescenta que um modelo de regressão pode ser comprometido pela baixa qualidade do banco de dados ou por problemas no processo de coleta dos mesmos. Por este motivo, a busca por variáveis relevantes se torna imprescindível.

Conforme Hara e Sillanpää (2009), em um banco de dados pode haver uma ampla quantidade de variáveis explicativas (contínuas ou discretas) que dificultam a geração de um

modelo de regressão confiável. O principal desafio é definir o menor conjunto de variáveis que melhor simbolize a predição da variável de resposta. Quando não se dispõe de um banco de dados de tamanho considerável, a modelagem torna-se mais complexa (CONZ, 2005).

Métodos de seleção de variáveis podem ser aplicados em diversas áreas com várias finalidades. Martins et al (2010) relataram a utilização de seleção de variáveis em uma nova técnica de calibração. Foi desenvolvida uma regressão linear múltipla robusta em relação às diferenças entre dois instrumentos de calibração - primário e secundário. Para realizar a regressão, foi utilizado o Algoritmo das Projeções Sucessivas (APS) para seleção de variáveis robustas com a técnica de sub-amostragem e agregação de modelos conhecida como *subagging*. O estudo gerou uma melhor predição do erro sistêmico para o instrumento secundário.

Utilizar a seleção de variáveis pode trazer uma série de benefícios, os quais incluem: facilitar visualização e o entendimento dos dados, reduzir a mensuração e o armazenamento de dados, e realizar o dimensionamento a fim de melhorar e desempenho de predição da variável de resposta. Para Guyon e Elisseeff (2003) diversos métodos podem ser utilizados para realizar a seleção de variáveis, destacando-se o Método Wrapper. Entretanto, algumas abordagens apresentam melhor ênfase em certos aspectos do que outros. Assim, é importante verificar o tamanho e características da base de dados para poder utilizar o melhor método de seleção de variáveis.

Para pesquisa em modelos de sistemas complexos, a seleção de variáveis é considerada uma etapa essencial, visto que o conjunto de dados apresenta alta dimensão e elevado número de variáveis redundantes (JUNIOR, 2006). Segundo George (2000), a solução de problemas de seleção de variáveis está muitas vezes ligada a modelos lineares, utilizando como base a regressão linear. Conforme Ghani e Ahmad (2010) há três métodos tradicionais para seleção de variáveis no contexto de regressão linear mútipla: *Forward Selection*, *Backward Elimination*, e *Stepwise*. Os três métodos são descritos a seguir.

Forward Selection: variáveis são acrescentadas uma a uma. A primeira variável a entrar no modelo é a variável que possui maior correlação com a variável de resposta, repetindo-se para as seguintes variáveis o mesmo critério. A correlação é calculada através do teste estatístico *F*, onde quanto maior o valor de *F* melhor é a correlação (MORAES e HAERTEL, 2007; MONTGOMERY et al, 2006).

Backward Elimination: O algoritmo inicia com todas as X variáveis no modelo. A variável que possui o menor *F* é eliminada da regressão, resultando em um modelo com X-1 variáveis. As variáveis subsequentes são retiradas do modelo empregando a mesma sistemática de eliminação (MONTGOMERY e RUNGER, 2007).

Stepwise: É a técnica mais utilizada de seleção de variáveis, a qual consiste em adicionar ou remover uma variável a cada passo com base no teste estatístico F (MONTGOMERY e RUNGER, 2007).

3. Metodologia

O estudo apresenta natureza aplicada e enfoque quantitativo, visto que utiliza bancos de dados proveniente dos processos de uma empresa. É uma pesquisa exploratória, pois busca conhecer com maior profundidade o assunto, e utiliza procedimentos bibliográficos, empregando referências da literatura para desenvolver o método. Desta forma, é possível replicar o estudo de caso em qualquer empresa em que um ou mais processos geram sucata.

O planejamento para a aplicação de seleção de variáveis é realizado em cinco etapas: (1) coleta e tratamento de dados; (2) separação do conjunto de dados em duas porções; modelagem dos dados e eliminação da variável com o menor módulo do coeficiente de regressão |"β"|; (3) construção de gráficos relacionando desempenho de predição e variáveis retidas; (4) teste do modelo composto pelas variáveis selecionadas na porção de teste; e (5) comparação dos resultados com os gerados pelo método *Stepwise*. As etapas são detalhadas a seguir.

3.1. Coleta e tratamento dos dados

A coleta de dados consiste no levantamento de informações relevantes para a empresa sobre os processos produtivos. Essa coleta é executada através da análise do banco de dados pré-existente ou por meio do controle periódico de novos dados.

O tratamento dos dados é realizado a fim de eliminar inconsistências no banco, tornando-o mais confiável. Uma base de dados pode conter informações atípicas oriundas de diversas fontes: erros humanos, situações incomuns explicáveis ou não, falhas de equipamentos e máquinas, entre outros. Esses casos especiais podem comprometer e distorcer o modelo de predição, justificando a eliminação de observações que contêm dados espúrios.

O gráfico de controle estatístico é a ferramenta indicada para realizar o tratamento dos dados. Calcula-se a média e o desvio padrão para cada variável do banco de dados, valendo-se das equações (12), (13) e (14) para definir o limite superior de controle (LSC), a média (LM) e o limite inferior de controle (LIC). Na sequência, os gráficos de controle são gerados e as observações inseridas fora dos limites de controle são removidas do banco de dados. A Figura 1 - Gráfico genérico de controle - ilustra o gráfico de controle para uma variável *n*, sinalizando a necessidade de eliminação de uma observação.

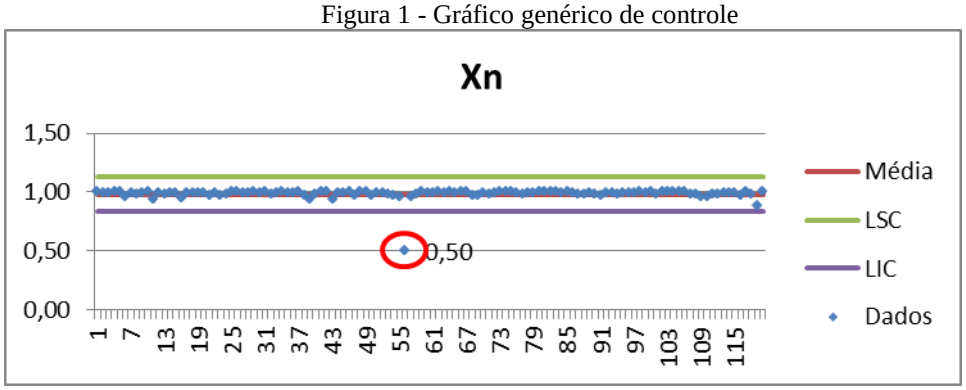
$$LSC = \mu_0 + 3\sigma_0 \quad (12)$$

$$LM = \mu_0 \quad (13)$$

$$LIC = \mu_X - 3\sigma_0 \ (14)$$

onde μ_0 representa a média e σ_0 o desvio padrão.

Figura 1 - Gráfico genérico de controle



Na sequência, é realizada a normalização dos dados para evitar distorções causadas pelas diferentes escala das variáveis coletadas. Além disso, a normalização possibilita utilizar a magnitude do coeficiente de regressão como indicador de importância das variáveis independentes.

3.2. Separar banco de dados em duas porções

O banco de dados deve ser aleatoriamente segmentado em duas porções: a primeira é definida como porção de treino, composta com 70% do banco de dados. Essa porção é utilizada para criar o modelo e selecionar as variáveis mais importantes. A segunda porção de teste é formada pelos 30% restantes, e possui como finalidade testar a capacidade de predição do modelo. Essa divisão é feita de tal forma que a junção dos dois bancos é igual ao conjunto original de dados.

3.3. Ajustar o modelo de regressão à porção de treino dos dados e eliminar as variáveis com menor |"β"|

Com a divisão dos dados, ajusta-se a regressão linear múltipla à porção de treino consistindo de todas as variáveis. Para estimar os coeficientes do modelo de regressão, utiliza-se o método de mínimos quadrados anteriormente apresentado. A aderência do modelo aos dados é estimada através do Critério de Informação Akaike (AIC) e Soma dos Quadrados dos Erros (SQE).

Após ajustar a regressão aos dados e calcular os índices de precisão (AIC e SQE), analisa-se a importância das variáveis independentes na predição da variável dependente. Como os dados foram normalizados na seção 3.1, seus coeficientes β_i, i=0,1,...,k, em valores absolutos, são redistribuídos em ordem decrescente. A variável independente X associada ao coeficiente de regressão com menor valor absoluto deve ser eliminada, pois variações em seu valor causam as menores alterações no valor de Y.

Na sequência, uma nova regressão deve ser ajustada com base nas variáveis remanescentes, e os índices de precisão devem ser novamente avaliados. Esse processo de predição/eliminação é repetido até sobrar apenas uma variável de processo no modelo.

3.4. Construir gráficos relacionando desempenho de predição e variáveis retidas

Concluído o processo de eliminação descrito na seção anterior, são construídos gráficos relacionando os índices de precisão - Akaike e SQE – ao número de variáveis remanescentes. Tais gráficos visam verificar a relevância das variáveis nos índices de precisão à medida que variáveis são eliminadas. O primeiro relaciona o Critério de Informação Akaike com o número de variáveis restantes no modelo, conforme ilustrado na Figura 2.

O segundo gráfico relaciona a Soma dos Quadrados do Erro com o número de variáveis restantes no modelo, Figura 3. Por fim, seleciona-se o conjunto de variáveis retidas que, simultaneamente, conduz ao menor valor de Akaike e SQE. Caso os conjuntos apontados pelos índices sejam diferentes, opta-se pelo índice que reteve o menor número de variáveis.

Figura 2 - Gráfico relação Akaike e variáveis no modelo

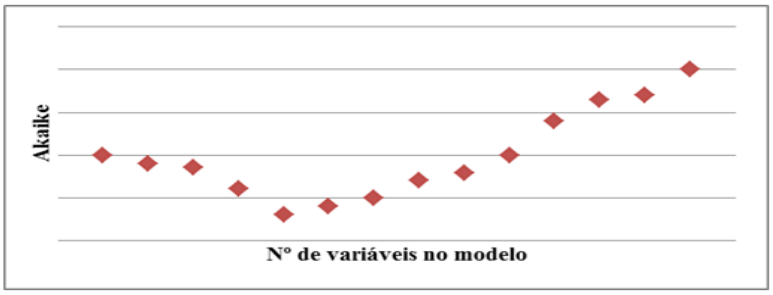
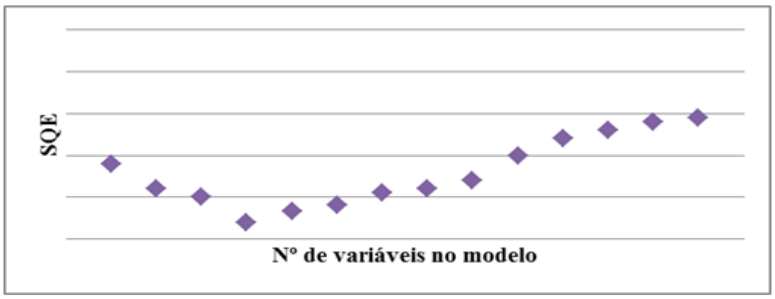


Figura 3 - Gráfico relação SQE e variáveis no modelo



3.5. Testar modelo selecionado na porção de teste e comparar com método *stepwise*

O subconjunto de variáveis que conduze ao menor índice de precisão é então utilizado nas amostras da porção de teste para que a capacidade preditiva do modelo possa ser avaliada para novos dados. A porção de teste também é importante caso os índices de precisão apontem conjuntos distintos de variáveis a serem retidas.

O desempenho do método de seleção de variáveis proposto é comparado ao tradicional método *Stepwise*, visto que ambos utilizam o mesmo princípio de eliminação sistemática das

variáveis. A comparação permite verificar se a seleção de variáveis com base na magnitude dos coeficientes de regressão (proposta neste artigo) é mais eficiente do que a seleção baseada em testes estatísticos (princípio do método Stepwise).

4. Resultados

O método proposto foi aplicado em uma empresa multinacional americana do ramo metal-mecânico, que conta com 25.000 colaboradores distribuídos em 26 países. O estudo foi realizado na unidade de Gravataí, situada no estado do Rio Grande do Sul. A empresa é fornecedora de elevada gama de produtos, fabricados através de processo de forjaria, usinagem, soldagem, entre outros processos tradicionais do setor metal-mecânico. Na unidade em estudo, o layout é organizado em células, distribuídas por famílias de produtos.

O método foi aplicado em uma célula com uma ampla família de produtos, com um alto nível de sucata formada. Para facilitar a análise, foram escolhidos os dois principais produtos processados pela célula, aqui chamados de Peça 1 e Peça 2. O banco de dados empregado no estudo é oriundo dos indicadores de processo, utilizado para avaliar e controlar o desempenho dos processos. Os dados foram coletados diariamente durante um período de seis meses, obtendo 178 registros, cada um contendo 10 variáveis: eficiência, produtividade, obtenção de peça conforme no primeiro processamento, número de paradas, número de pessoas, horas aplicadas, disponibilidade, MTBF, MDT e defeito de fornecedor.

Após coleta, os dados foram tratados através do gráfico de controle (conforme descrito na seção de método), sendo que 4% dos dados apresentaram inconsistência por se tratarem de causas especiais. Esses foram eliminados para prevenir problemas de aderência do modelo gerado. Além disso, a variável “defeito de fornecedor” foi retirada do banco de dados para ambas as peças, uma vez que diversos valores para essa variável estavam localizados fora dos limites de controle do gráfico. Na sequência, o banco de dados foi normalizado e dividido em porção de treino (70% dos dados), e porção de teste (30% restante).

Na sequência, iniciou-se o processo de seleção das variáveis para geração do modelo de predição de quantidade de sucata gerada pelo processo. Os resultados para cada peça são apresentados simultaneamente. Através da porção de treino, foram calculados os coeficientes de regressão para cada variável e, assim, definidos os índices de qualidade de predição (Akaike e SQE) para os modelos consistindo de todas as variáveis. Na sequência, a variável com menor módulo de β (variável X_2 para o modelo preditivo da Peça 1 e X_4 para a Peça 2) foram eliminadas. Tal procedimento foi repetido até que apenas uma variável restasse em cada modelo (X_6 para Peça 1 e X_8 para a Peça 2). Os coeficientes e índices SQE e IAC gerados após cada eliminação de variável são apresentados nas Tabelas 1 e 2.

Tabela 1- Coeficientes e índices para Peça 1

Nº de Iterações	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	SQE	IAC
1	0,0807	0,0085	0,1543	0,0244	0,1746	0,2376	0,1072	0,1215	0,1131	84,055	2,6362
2	0,0781		0,1552	0,0244	0,1707	0,2319	0,1068	0,1217	0,1131	84,058	2,6242
3	0,0956			0,1541	0,1681	0,2290	0,1109	0,1238	0,1166	84,071	2,6123
4			0,1216		0,2004	0,3281	0,0233	0,0855	0,0342	84,298	2,6030
5			0,1208		0,1980	0,3283		0,0764	0,0076	84,306	2,5790
6			0,1192		0,1991	0,3275		0,0717		84,801	2,5728
7			0,1194		0,1919	0,3531				84,306	2,5790
8					0,1847	0,3361				84,996	2,5630
9						0,2235				87,585	2,5810

Tabela 2 - Coeficientes e índices para Peça 2

Nº de Iterações	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9	SQE	IAC
1	0,3758	0,0222	0,1018	0,0077	0,1722	0,6065	0,3220	0,0765	0,2796	83,558	2,6303
2	0,3703	0,0222	0,1014		0,1714	0,6056	0,3206	0,0771	0,2785	83,560	2,6650
3	0,3772		0,1038		0,1613	0,5910	0,3219	0,0777	0,2785	83,577	2,6480
4	0,4059		0,0869		0,1626	0,6263	0,4388		0,4265	83,714	2,5960
5	0,4175				0,1628	0,6253	0,4417		0,4292	83,812	2,6163
6	0,3083					0,4475	0,4579		0,4495	85,134	2,6147
7						0,1914	0,2824		0,3810	89,061	2,6426
8						0,2742			0,4450	92,247	2,6605
9									0,1930	93,560	2,6574

Após obter os indicadores de precisão para os modelos oriundos de cada iteração, foram construídos os gráficos de Akaike e SQE, apresentados nas Figuras 4 a 7. Como se pode observar, o índice SQE para as duas peças aumenta à medida que o número de variáveis no modelo diminui. Entretanto, tal variação apresenta uma alteração pouco significativa com a eliminação das variáveis. Para a Peça 1, o SQE aumenta consideravelmente quando o modelo passa de duas para uma variável retida (indicando perda significativa de capacidade preditiva). O gráfico de Akaike (Figura 6), por sua vez, apresenta seu menor valor quando duas variáveis são retidas. Desta forma, a regressão linear que melhor prediz o volume de sucata da Peça 1 possui duas variáveis, representadas pela equação (15). Análise semelhante foi realizada para a segunda peça: o SQE aumenta significativamente quando o modelo passa de 4 para 3 variáveis retidas (Figura 5). Já o gráfico do Akaike apresenta um mínimo relativo nesse mesmo instante (Figura 7). A regressão gerada para a Peça 2 é apresentada na equação (16). A descrição das variáveis retidas é apresentada na sequência.

$$y_{peça1} = 0{,}461 - 0{,}184x_5 + 0{,}336x_6 \quad (15)$$

$$y_{peça2} = 0{,}829 - 0{,}308x_1 + 0{,}447x_6 + 457x_7 + 0{,}449x_9 \quad (16)$$

Onde X_1 representa a eficiência, X_5 a quantidade de Pessoas, X_6 a quantidade de horas aplicadas, X_7 a disponibilidade, X_9 o MDT e $y_{peça}$ o nível de sucata. A variável X_6, horas aplicadas, foi considerada significativa para a criação dos modelos de predição para as duas peças. Tal variável é

diretamente proporcional à sucata, pois se refere à produção convertida em horas. Assim, quanto maior a produção, maior será o nível de sucata formada.

Na Peça 2 foram retidas as variáveis X_1, X_7, e X_9. A primeira variável representa a eficiência, sendo esse indicador relacionado com a máquina gargalo. Como o nível de sucata formada por essa máquina foi elevado, tal variável tornou-se importante na predição de sucata. As outras duas variáveis, disponibilidade e MTD, representam indicadores de manutenção. O elevado volume de produção da Peça 2 acaba gerando desgastes na máquina, o que eleva o nível de formação de sucata.

Figura 4 - SQE para a peça 1

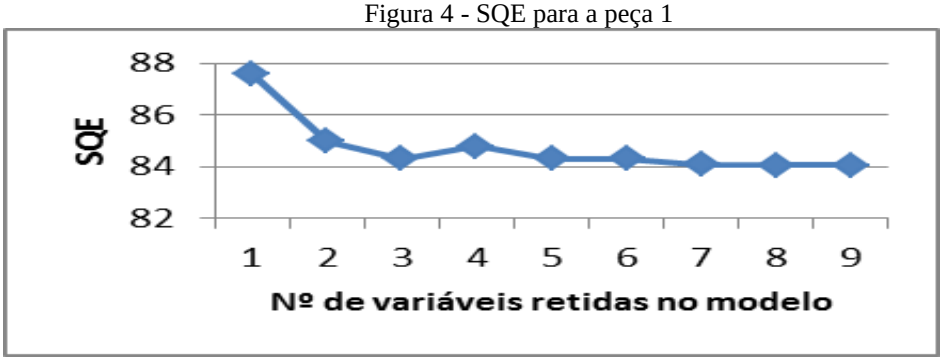


Figura 5 - SQE para a peça 2

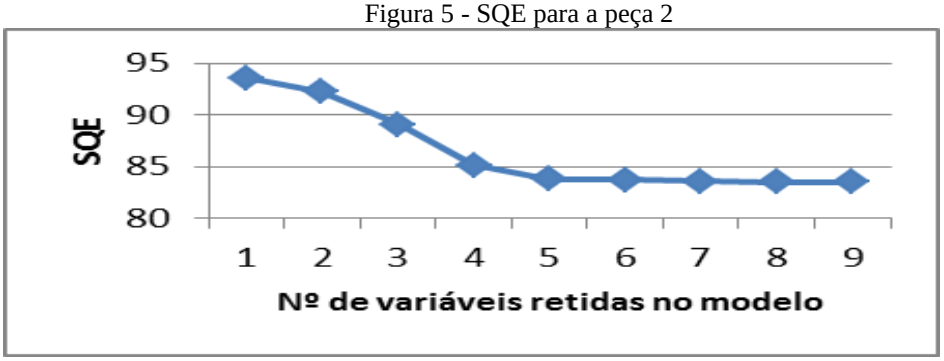


Figura 6 - Akaike para a peça 1

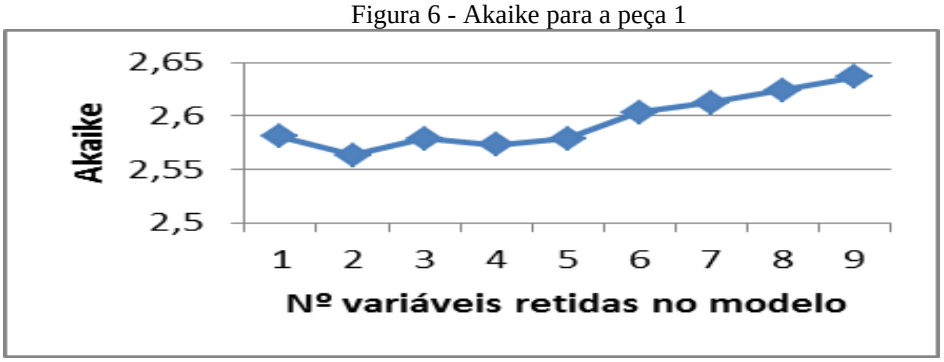
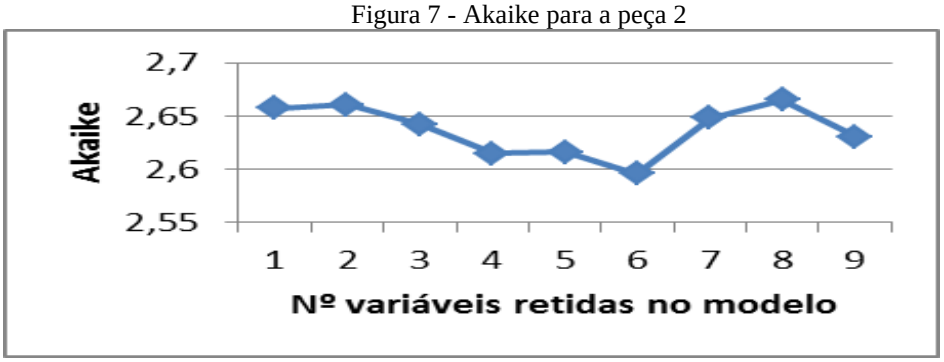


Figura 7 - Akaike para a peça 2



Os modelos gerados foram então aplicados na predição das observações inseridas na porção de teste, com vistas à avaliação de sua capacidade preditiva. Paralelamente, aplicou-se o método *Stepwise* sobre a porção de treino, obtendo-se os modelos abaixo.

$$y_{peça\ 1} = 0{,}465 + 0{,}223x_1$$

$$y_{peça\ 2} = 0{,}829 + 0{,}264x_8$$

onde X_1 representa a eficiência e X_8 o MTBF. A Tabela 3 traz o SQE gerado pelo método proposto e pelo *Stepwise*.

Tabela 3 - Comparação de SQE entre métodos

	Peça 1	Peça 2
Metodologia	24,4	38,0
Stepwise	22,8	46,4

O método proposto conduz a predições significativamente melhores para a Peça 2 (menor SQE) e SQE levemente maior para a Peça 1. Ressalta-se, no entanto, que o método proposto se apoia em preceitos mais simples que o *Stepwise*, além de dispensar um software estatístico para modelagem da regressão.

5. Conclusão

A sucata formada pelos processos de indústrias do ramo metal mecânica é considerada como perda financeira para a empresa. Tal perda é decorrente da falta de eficiência e ausência de conhecimento sobre os processos existente na organização. Existe uma grande quantidade de variáveis podem contribuir para a formação de sucata, sendo relevante identificar as mais importantes para descrição e predição do processo.

Este estudo propôs uma sistemática para seleção das variáveis mais relevantes com vistas à geração de um modelo de regressão. As variáveis menos relevantes foram eliminadas com base no módulo do coeficiente de regressão e indicadores de qualidade de predição foram gerados após cada eliminação.

O método foi aplicado em dados de produção de sucata de duas peças. Para a Peça 1, as variáveis mais significativas foram: "Número de Pessoas no processo" e "Horas aplicadas", já para a segunda peça foram mantidas as variáveis "Eficiência", "Horas aplicada", "Disponibilidade" e "MDT". Por fim, o método sugerido apresentou resultados ligeiramente superiores aos gerados pelo tradicional método *Stepwise*.

Estudos futuros incluem o desenvolvimento de outros índices de importância de variáveis para guiar o processo de eliminação das variáveis menos relevantes em termos de predição.

Abstract

Variable selection is deemed a relevant analysis in industrial process, since process usually rely on noisy and correlated databases. This article defines a minimum of variables for predicting scrap formation in a metal mechanic sector company. A multiple linear regression model is initially fitted to the data. The variables are then systematically removed based on the absolute value of the regression coefficient. After each variable is removed, the predictive ability of the model is evaluated by the Akaike Information Criteria (AIC) and Sum of Squares Errors (SSE) metrics. The resulting models were deemed coherent by process specialists.

Key-words: variable selection; scrap; metal mechanic sector company.

Referências

ANZANELLO, M.J.; ALBIN, S.L.; CHAOVALITWONGSE W.A. **Selecting the best variables for classifying production batches into two quality levels**. Chemometrics and Intelligent Laboratory Systems, 97 (2009) 111-117. crossref

BARROS, E.A.C., SIMÕES, P. A., ACHCAR, J. A., MARTINEZ, E. Z., SHIMANO, A. C. 2008. **Métodos De Estimação Em Regressão Linear Múltipla: Aplicação A Dados Clínicos**. Revista Colombiana de Estadística, v.31, n.1, p.111-129.

BERNARDI, A. C. C, RODRIGUES, A. A, MEDONÇA, F. C, TUPY, O, JUNIOR, W. B, PRIMAVESI O. 2010. **Análise E Melhoria do Processo de Avaliação dos Impactos Econômicos, Sociais e Ambientais de Tecnologias da Embrapa Pecuária Sudeste**. Revista Gest. Prod., São Carlos, v.17, n.2, p.297-316.

BIASOLI, P. K. 2005. **Modelagem Conjunta de Média e Variância em Experimentos Fracionados sem Repetição Utilizando GLM**. Dissertação (Mestrado em Engenharia de Produção). Escola de Engenharia, Programa de Pós Graduação em Engenharia de Produção, Universidade Federal do Rio Grande do Sul.

CHOI, S. OH, J., CHOI, C., KIM, C. 2002. **Input Variable Selection for Feature Extraction in Classification Problems**. Signal Processing, v.92, n.3, p.636-648. crossref

CONZ, V. 2005. **Desenvolvimento de Analisadores Virtuais Aplicados a Colunas de Destilação Industriais.** Dissertação (Mestrado em Engenharia Química). Programa de Pós-Graduação em Engenharia de Química, Universidade Federal do Rio Grande do Sul.

DOWNING, D, CLARK, J. 2006. **Business Statistics**. Nova Yorque, Barron's Educational Series, inc.

FREUND, J.F., SIMON, G.A. 2000**. Estatística Aplicada Economia Administração e Contabilidade**. Porto Alegre, Bookman.

GHANI, I.M.M., AHMAD, S. 2008. **Stepwise Multiple Regression Method to Forecast Fish Landing.** Procedia Social and Behavioral Sciences, v.8, p.549–554. crossref

GUJARATI, D. 2003. **Basic Econometrics**. McGraw-Hill Companies, inc.

GUYSON, I., ELISSEEFF, A. 2003. **An Introduction to Variable and Feature Selection.** Journal of Machine Learning Research, v.3, p.1157-1182.

JEORGE, E. 2000. **The variable selection problem. Journal of the American Statistical Associatio**n, v.95, n.452, p.1-12.

JUNIOR, F.P. 2006. **Seleção de Variáveis e Características como Aplicação Paralela para Cluster MPI**. Dissertação (Mestrado em Ciência da Computação). Programa de Pós-Graduação em Ciência da Computação. Universidade Estadual de Maringá.

KOPER, R.A. 2006. **Diagnóstico Estratégico Da Produção E Operações De Uma Empresa Metalúrgica Múltiplana**. Dissertação (Mestrado em Administração). Programa de Pós-Graduação em Administração, Universidade Federal do Rio Grande do Sul.

LIU, H.; YU, L. **Toward integrating feature selection algorithms for classification and clustering**. Transactions on Knowledge and Data Engineering, 17 (2005) 491–502. crossref

MARTINS, M. N., GALVÃO, R.K.H., PIMENTEL, M.F. 2010. **Multivariate Calibration Transfer Employing Variable Selection and Subagging.** J. Braz. Chem. Soc., v.21, n.1, p.127-134. O'HARA, R. B., SILLANPÄÄ, M. J. 2009. **A Review of Bayesian Variable Selection Methods:** What, How and Which. International Society for Bayesian Analysis, v.4, n.1, p.85-118. crossref

MÜLLER, C. J. 2003. **Modelo de Gestão Integrando Planejamento Estratégico, Sistemas De Avaliação de Desempenho e Gerenciamento De Processos (MEIO – Modelo de Estratégia, Indicadores e Operações).** Tese (Doutorado em Engenharia de Produção). Escola de Engenharia, Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal do Rio Grande do Sul.

MONTGOMERY, D.C., PECK, E.A., VINING, G.G. 2006. **Introduction to Linear Regression Analysis**. Nova Iorque: A John Wiley & Sons, inc.

MONTGOMERY, D.C., RUNGER, G.C. 2007. **Applied Statistics and Probability for Engineers**. Nova Iorque: A A John Wiley & Sons, inc.

MORAES, A. O. M., HAERTEL, V. 2007. **Métodos hierárquicos para redução de dimensões e classificação de imagens** AVIRIS. Anais XIII Simpósio Brasileiro de Sensoriamento Remoto, INPE, Florianópolis, p.6481-6488.

NUNES, P. G. A. 2008. **Uma Nova Técnica Para Seleção de Variáveis em Calibração Multivariada Aplicada às Espectrometrias UV-VIS e NIR.** Tese (Doutorado em Química). Programa de Pós Graduação em Química, Universidade Federal da Paraíba.

O'HARA, R. B., SILLANPÄÄ, M. J. 2009. **A Review of Bayesian Variable Selection Methods:** What, How and Which. International Society for Bayesian Analysis, v.4, n.1, p.85-118. crossref

RIBEIRO, J. L. D., TEN CATEN. C.T. 2000. **Estatística Industrial**. Programa de pós Graduação em Engenharia de Produção, UFRGS, Porto Alegre.

SARTORIS, A. 2003. **Estatistica e Introdução à Econometria**. São Paulo: Saraiva.

WENSING, D.A. 2010. **Redução de Sucata e Retrabalho em uma Indústria Metal Mecânica**. Trabalho de graduação. Departamento de Engenharia de Produção e Sistemas, UDESC, Joinville.

WERKEMA, M.C.C., AGUIAR, S. 1996. **Análise de Regressão: Como Entender o Relacionamento entre as Diversas Variáveis de um Processo**. Fundação Christiano Ottoni, Minas Gerais.

Dados dos autores

Nome completo: **Marcela Stein**

Filiação institucional: Universidade Federal do Rio Grande do Sul – UFRGS – RS - Brasil

Departamento: Departamento de engenharia de Produção e Transportes
Endereço completo para correspondência: Av. Osvaldo Aranha, 99, Porto Alegre, RS
Telefone: (51) 3308 4423
e-mail: marcela.stein_ctbm@hotmail.com

Nome completo: **Michel José Anzanello**
Filiação institucional: Universidade Federal do Rio Grande do Sul – UFRGS – RS - Brasil
Departamento: Departamento de engenharia de Produção e Transportes
Endereço completo para correspondência: Av. Osvaldo Aranha, 99, Porto Alegre, RS
Telefone: (51) 3308 4423
e-mail: michel.anzanello@gmail.com

Nome completo: **Alessandro Kahmann**
Filiação institucional: Universidade Federal do Rio Grande do Sul – UFRGS – RS - Brasil
Endereço completo para correspondência: Av. Osvaldo Aranha, 99, Porto Alegre, RS
Telefone: (51) 3308 4423
e-mail: alessandro.kahmann@ufrgs.br

Submetido em: 02/07/2013
Aceito em: 08/10/2014