Seleção das variáveis de processo mais relevantes para predição dos níveis de sucata em um processo do setor metal-mecânico

Marcela Stein (UFRGS)
Michel José Anzanello (PPGEP/UFRGS)
Victor Cervo (PPGEP/UFRGS)
Alessandro Kahmann (PPGEP/UFRGS)

Resumo
Esse artigo tem como objetivo definir os indicadores (variáveis) mais relevantes para a predição dos níveis de formação de sucata em uma empresa do ramo metal mecânico. Um modelo de regressão linear múltipla é inicialmente ajustado aos dados. 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 do através do Critério de Informação Akaike (AIC) e Soma dos Quadrados dos Erros (SQE).
Palavras chave: Seleção de variáveis, Sucata, Empresa do ramo metal mecânico.

1 Introdução

Visando acompanhar o desenvolvimento industrial mundial e apresentarem-se competitivas em um mercado altamente disputado, as empresas buscam garantir a satisfação dos clientes por meio de redução de custos produtivos e de otimizações em seus processos e sistemas. O apropriado entendimento do funcionamento da organização, bem como a correta definição dos aspectos passíveis de melhorias, é de extrema importância à obtenção do sucesso das empresas. Assim sendo, essas necessitam conhecer com precisão 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 que seja desenvolvido um planejamento estratégico eficaz com vistas à melhoria do processo, entre outros benefícios (Kober, 2006).

Para melhorar e controlar um processo, diversas sistemáticas de coleta e armazenagem de dados são utilizadas. Entretanto, o modo como os dados são empregados influencia diretamente na determinação do quanto o sistema como um todo agregou valor ao produto e ao serviço (Harrington, 1993 *apud* Müller, 2003). 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. Quando o gerenciamento das informações provenientes de um processo não é realizado de forma adequada, podem não só ocorrer 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 não qualidade e do produto, prejudicando diretamente 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. Não se verifica, contudo, disponibilidade de ferramental estruturado com vistas à identificação das variáveis que contribuem ativamente para a formação de sucata no final do processo e para a consequente perda de produtividade conectada à sucata gerada.

Este artigo propõe uma sistemática 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. As variáveis independentes foram, então, sistematicamente eliminadas com base na magnitude do coeficiente de regressão, e a precisão de predição do modelo resultante avaliada através de indicadores apropriados. 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 assuntos 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

Diversos processos produtivos envolvem duas ou mais variáveis que se relacionam entre si. Tal interação pode ser modelada por meio de regressões que permitem compreender a relação existente entre as variáveis e viabilizar ações no processo em análise. Em modelos de regressão existe uma variável dependente (*Y*), conhecida como variável de 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 e 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. Essas funções são frequentemente baseadas em estudo 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 regressoras (*X*) são responsáveis pela determinação do nível de uma variável de resposta (*Y*), conforme ilustrada na 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 um erro aleatório εi oriundo da diferença entre os *Y* observados e os *Y* gerados pela equação. Porém, os erros são independentes e apresentam distribuição normal com média zero e variância σ^2 desconhecida (Barros et al, 2008). A regressão linear múltipla pode estar relacionada a *k* variáveis e os parâmetros βi, i=0,1,...,k, são definidos como coeficientes de regressão. Tais coeficientes são desconhecidos, devendo ser estimados por 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 equação (2). A solução desse método pode ser trabalhosa, pois o número de equações a serem resolvidas é proporcional ao número de parâmetros a serem estimados. Segundo Barros et al. (2008), o método baseia-se na 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 mais conveniente utilizar notação matricial, pois os dados, modelos e resultados são exibidos compactados. 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 Medidores de Precisão

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 à regressão em estudo definir um subconjunto de variáveis com os quadrados médios residuais iguais ou próximos ao mínimo. Desta forma, espera-se obter uma regressão na qual a adição de novas variáveis não acresça precisão na predição (Werkema e Aguiar, 1996). O QMR pode ser calculado conforme a equação (7).

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

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 (R2)

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 R2 é considerado problemático, pois o coeficiente aumenta quando variáveis regressoras são acrescentadas ao modelo. O R2 ajustado, por sua vez, penaliza a adição de variáveis desnecessárias na regressão (Downing e Clark, 2006, Montgomery e 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 total;

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

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 isso, a busca por variáveis relevantes é imprescindível.

Conforme Hara e Sillanpää (2009), em um banco de dados há ampla quantidade de variáveis explicativas (contínuas ou discretas) que dificultam os problemas de regressão. 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 tona-se mais complexa e podendo, muitas vezes, ser impossível a sua realização (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.

Há muitos benefícios em utilizar seleção de variáveis: 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, entre os métodos citados pelos autores está o Método Wrapper. Entretanto, alguns apresentam melhor ênfase em certos aspectos do que outros. Assim, é importante verificar o tamanho e características da base de dados.

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 Regression. 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 Fstatistic, 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 Regression: 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, devido à utilização de 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: coleta e tratamento de dados; separação do conjunto de dados em duas porções; modelagem dos dados e eliminação da variável com o menor módulo do coeficiente de regressão |“β”|; construção de gráficos relacionando desempenho de predição e

variáveis retidas; e teste do modelo composto pelas variáveis selecionadas na porção de teste. Por fim, comparam-se os resultados com os do método de 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 dados espúrios.

O gráfico de controle estatístico é a ferramenta designada 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 controleilustra o gráfico de controle para uma variável *n*, sinalizando a necessidade de eliminação de uma observação.

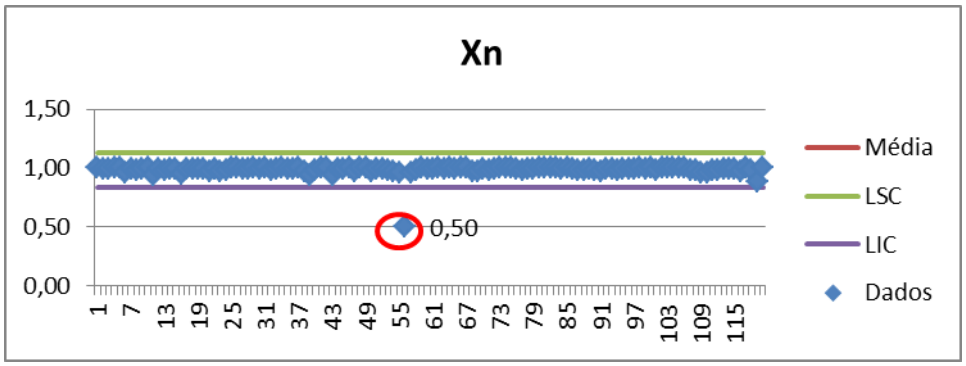
$$LSC = \mu_0 + 3\sigma_0 \quad (12)$$

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

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

Onde, μ0 representa a média e σ0 o desvio padrão.

Figura 1 - Gráfico genérico de controle



Na sequência, é realizada a normalização dos dados, a qual reduz efeitos de escala das variáveis coletadas. Além disso, a normalização possibilita utilizar a magnitude do coeficiente de regressão como índice de importância das variáveis.

3.2 Separar Banco de Dados em duas Porções

O banco de dados deve ser 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.

3.3 Ajustar Regressão à porção de treino dos dados e Eliminar Variáveis com Menor |“β”|

Nessa etapa, é realizado o ajuste da regressão linear múltipla utilizando todas as variáveis. Para estimar os coeficientes do modelo de regressão, utiliza-se o método de mínimos quadrados, conforme apresentado anteriormente. A qualidade de 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, os módulos dos coeficientes β_i, i=0,1,...,k são redistribuídos em ordem decrescente. A variável independente X associada ao menor módulo do coeficiente de regressão deve ser eliminada, pois variações em seu valor causam as menores alterações em 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 predição - 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 – Gráfico relação Akaike e variáveis no modelo2. 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

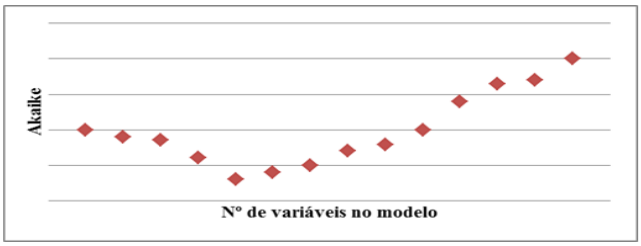
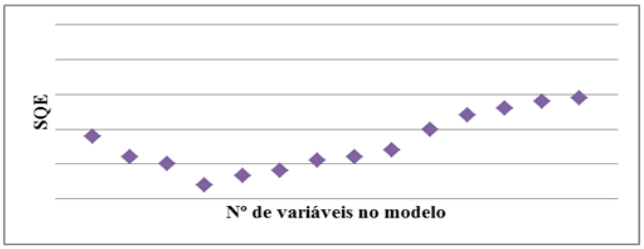


Figura 3 - Gráfico relação SQE e variáveis no modelo



3.5 Testar Modelo na Porção de Teste e Comparar com Método Stepwise

O modelo composto pelas variáveis selecionadas é então aplicado na 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 um alto nível de sucata formada, com uma ampla família de produtos. Para facilitar a análise, foram escolhidos os dois principais produtos processados pela célula, definidos como 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. São elas: 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 o 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, o indicador "defeito de fornecedor" foi retirado do banco de dados para as duas 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_n (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 do Akaike e do 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 valor mínimo 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.

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

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

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 Ypeça 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 refere-se 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

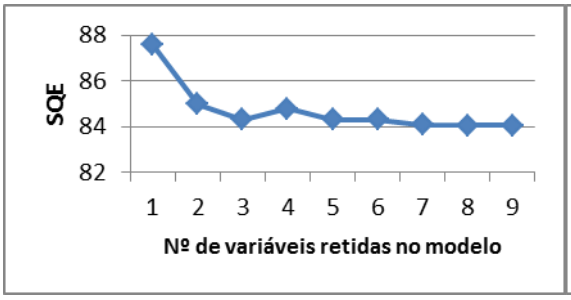


Figura 5- SQE para a peça 2

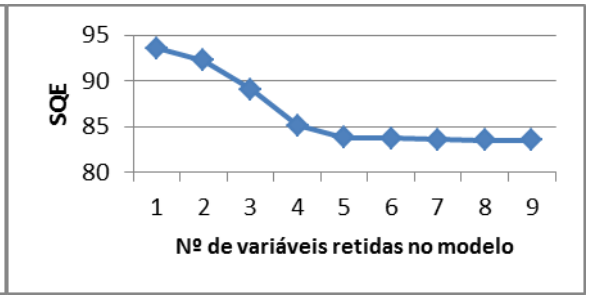


Figura 6 – Akaike para a peça 1

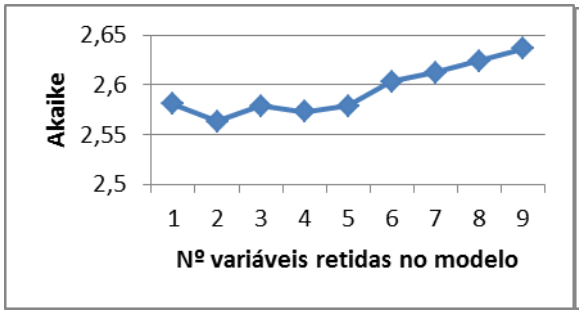
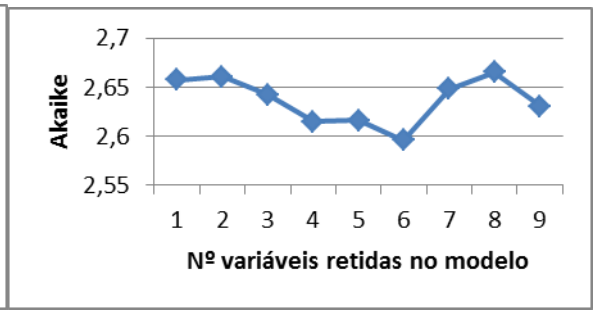


Figura 7 – Akaike para a peça 2



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 em uma indústria do ramo metal mecânica é avaliada 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.

Referências Bibliográficas

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. Dissertação (Mestrado em Engenharia de Produção). Modelagem Conjunta de Média e Variância em Experimentos Fracionados sem Repetição Utilizando GLM. 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.

CONZ, V. 2005. Dissertação (Mestrado em Engenharia Química). Desenvolvimento de Analisadores Virtuais Aplicados a Colunas de Destilação Industriais. 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.

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.

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.

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

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

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

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.

MÜLLER, C.J. 2003. Tese (Doutorado em Engenharia de Produção). Modelo de Gestão Integrando Planejamento Estratégico, Sistemas De Avaliação de Desempenho e Gerenciamento De Processos (MEIO – Modelo de Estratégia, Indicadores e Operações). 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. Tese (Doutorado em Química). Uma Nova Técnica Para Seleção de Variáveis em Calibração Multivariada Aplicada às Espectrometrias UV-VIS e NIR. Programa de Pós Graduação em Química, Universidade Federal da Paraíba.

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.