Seleção de Variáveis para Predição utilizando Regressão Linear em Processos Logísticos de Distribuição

Miguel Slomp Masiero, UFRGS, miguelmasiero@hotmail.com
Michel José Anzanello, UFRGS, anzanello@producao.ufrgs.br

Resumo

Este artigo tem como objetivo selecionar os principais indicadores que influenciam o desempenho de um processo de distribuição logístico. Métodos de seleção de variáveis são analisados de maneira a eleger aquele que melhor se adéqua às necessidades do estudo, definindo as variáveis independentes que influenciam diretamente uma variável de resposta denominada “tempo em rota”, a qual quantifica o tempo despendido pelas equipes para efetuar as entregas. O modelo gerado permite prever tempos de rota para diversos cenários, bem como quantificar a influência de cada variável independente em relação à variável de resposta.

Palavras Chave: Seleção de variáveis, Predição, Logística de distribuição.

Abstract

This article aims to select the main indicators that influence the performance of a logistic distribution process. Variable selection methods are analyzed to select the one that better defines the independent variables that directly influence a response variable named "time route", which quantifies the time required for the delivering process. Our model can predict the time course for different scenarios, and quantify the influence of each independent variable on the response variable.

Key Words: Variable selection, Prediction, Distribution logistics.

1. INTRODUÇÃO

Visando ao aumento de competitividade e diminuição de custos, as organizações vêm constantemente desenvolvendo formas de aumentar a eficiência dos processos produtivos e serviços prestados. Tal busca tem gerado diversos métodos para reduzir perdas, controlar os processos e maximizar a produção.

Nesse contexto, torna-se importante o acompanhamento dos diversos fatores que alteram os níveis de produtividade da organização, os quais são representados por variáveis específicas. Algumas variáveis possuem maior representatividade sobre o desempenho global da organização, ou seja, uma pequena variação desses fatores reflete em mudanças significativas, como faturamento ou nível de serviço. A análise dessas variáveis pode demonstrar a eficiência de programas de melhoria e de alterações realizadas no processo. Para tanto, é necessária a compreensão e entendimento das correlações entre as variáveis utilizadas pela organização no gerenciamento de suas operações. Em diversas organizações, as variáveis mais importantes são classificadas como indicadores de desempenho, os quais possuem metas específicas para cada variável. Martins e Neto (1998) afirmam que os indicadores de desempenho sinalizam necessidade de ação para restaurar uma causa especial crônica ou atingir um desempenho nunca antes atingido.

Em empresas de logística, verifica-se uma carência no entendimento do indicador de desempenho “tempo em rota”, que quantifica o tempo despendido pelas equipes de entregas para cumprir sua rotina. A quantidade de horas trabalhadas pelas equipes de entrega durante o processo de distribuição deve ser inferior a nove horas e vinte minutos, com intervalo de uma hora para almoço. Ultrapassar essa carga de trabalho implica em diversas consequências, como problemas sindicais, queda do rendimento dos funcionários, insatisfação com a empresa e absenteísmo (falta ao trabalho). A procura por melhor desempenho das equipes de entrega deve estar atrelada à preocupação do bem estar das pessoas, que pode ser medida, parcialmente, pela quantidade de horas trabalhadas por dia.

Frequentemente, informações utilizadas para a análise do indicador “tempo em rota” são armazenadas em bancos de dados com variáveis altamente correlacionadas e com elevados níveis de ruído. A interpretação errônea das informações, bem como o desconhecimento das correlações entre as variáveis, pode levar a organização a tomar decisões equivocadas, causando a melhoria de um indicador de desempenho em detrimento de outro que, em alguns casos, pode levar à redução da produtividade global (HUGE, 1990). De tal forma, justifica-se a utilização de sistemáticas que permitam a identificação das variáveis mais relevantes para análise do processo.

Este artigo tem como objetivo selecionar as variáveis mais relevantes para predição de uma variável de resposta (variável dependente) denominada tempo em rota, item importante na avaliação de desempenho de processos logísticos de distribuição de uma organização. As variáveis de processo (ou independentes) a serem incluídas na regressão consistem de indicadores secundários associados ao processo de distribuição. A análise da regressão é uma técnica estatística que permite investigar e modelar a relação entre variáveis (Montgomery e Peck, 1992). O artigo compara 4 métodos tradicionais de seleção de variáveis de processo, além de propor um indicador de eficiência para a identificação do melhor método de seleção de variáveis.

Este artigo está estruturado como segue, além desta introdução. Uma revisão teórica sobre seleção de variáveis, logística e regressão linear é apresentada na seção 2. Na seção 3 são descritos os procedimentos metodológicos, e na seção 4 apresentam-se os resultados. Na seção 5, são apresentadas as conclusões do estudo.

2. REFERENCIAL TEÓRICO

2.1 Regressão Linear

A regressão é uma técnica estatística que visa à compreensão das relações entre diversas variáveis. Tal ferramenta auxilia no processo de tomada de decisão, pois fornece informações consistentes sobre a relação dos indicadores com a variável de resposta monitorada (WERKEMA e AGUIAR, 1996). Os mesmos autores citam que a análise de regressão é utilizada para investigar e modelar o relacionamento entre diversas variáveis de um processo, baseando-se na idéia de empregar uma equação para expressar tal relacionamento. Obtida a equação de regressão, evidenciam-se os fatores que necessitam de melhorias.

Gráficos relacionando as variáveis de regressão ($x_i \times x_j$) permitem verificar a existência de multicolinearidade entre as variáveis. A multicolinearidade informa que duas ou mais variáveis de regressão são dependentes entre si, o que dispensa a inclusão de ambas no modelo de regressão. A inclusão de variáveis altamente correlacionadas pode tornar o modelo pouco preciso (MONTGOMERY e PECK, 1992). Dois modelos clássicos de regressão linear são descritos na sequência.

2.1.2 Regressão Linear Simples

A regressão linear simples baseia-se na tentativa de estabelecer uma equação matemática linear que relacione uma variável independente e uma variável de resposta. Tal método pode ser aplicado para conhecer a relação em um processo simples, formado por apenas uma variável aleatória que influência a variável resposta. Como exemplo, pode-se citar o consumo de água tratada nas grandes cidades que cresce linearmente com o aumento da temperatura nos meses de verão. No exemplo, apenas a variável temperatura influencia a variável resposta consumo de água tratada (WERKEMA e AGUIAR, 1996).

A regressão linear simples apresenta dois parâmetros: o coeficiente angular ou declividade a (derivada primeira da reta), e o coeficiente linear b, que simboliza o valor de y (variável de resposta) quando $x=0$.

$$Y = aX + b \qquad (1)$$

Ressalta-se que, para cada valor de x_i, podem existir um ou mais valores de y_i, e que para cada valor de x_i existirá um desvio d_i (ou erro ε_i) dos valores de y_i.

2.1.3 Regressão Linear Múltipla

A regressão linear múltipla possui o mesmo objetivo e características da regressão linear simples, diferenciado-se pelo número de variáveis independentes *x* envolvidas no problema (ou seja, a variável de resposta *y* depende de duas ou mais variáveis independentes $x_,$ dos parâmetros desconhecidos além de um erro (ε) estimado), conforme ilustrado na equação 2.

$$Y = \beta_o + \beta_1 x_1 + \beta_2 x_2 + ... + \beta_{k-1} x_{k-1} + \beta_k x_k + \varepsilon \qquad (2)$$

Cabe salientar que o adjetivo "linear" é empregado porque o modelo é função linear dos parâmetros desconhecidos $\beta_o, \beta_1 ... \beta_k$ (WERKEMA e AGUIAR, 1996).

O parâmetro de erro ε representa o desvio entre os valores observados e os estimados para cada observação e não deve possuir informação sistemática para a determinação de *y* que não tenha sido capturada pelas variáveis x_j. O parâmetro de erro é pressuposto como independente e normalmente distribuído (CHATTERJEE e PRICE, 1991). Para estimar os valores de $\beta_o, \beta_1 ... \beta_k,$ necessita-se de uma base de dados que represente a realidade do processo em questão. Para obter os estimadores de mínimos quadrados de $\beta_o, \beta_1 ... \beta_k$, deve-se minimizar a função erro *L*:

$$L = \sum_{i=1}^{n} \varepsilon_i^2 = \sum_{i=1}^{n} (y_i - \beta_o - \sum_{j=1}^{k} \beta_1 x_{ij})^2 \qquad (3)$$

Pode-se resolver o sistema de equações expressando-as em notação matricial, conforme apresentado:

$$Y = X\beta + \varepsilon \qquad (4)$$

Manipulações matemáticas sobre os vetores *y*, β e ε e sobre a matriz *X* permitem estimar o vetor de coeficientes de regressão β que minimiza o valor dos desvios quadrados na equação (3) (JOBSON, 1991). A equação (5) apresenta a forma matricial resultante das manipulações.

$$\beta = (X'X)^{-1} X'y \qquad (5)$$

Onde os termos podem ser representados da seguinte forma:

$y = \begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix}$ vetor das observações $\beta = \begin{bmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_n \end{bmatrix}$ vetor dos coeficientes de regressão

$$X = \begin{bmatrix} 1 & x_{11} & x_{12} & x_{1k} \\ 1 & x_{21} & x_{22} & x_{2k} \\ \vdots & \ddots & \vdots & \vdots \\ 1 & x_{n1} & \dots & x_{nk} \end{bmatrix}$$ matriz dos níveis das variáveis

2.2 Seleção de Variáveis

A seleção de variáveis aplicada à modelagem de sistemas objetiva a redução da quantidade de variáveis de entrada de um modelo. Esta análise justifica-se quando o número de variáveis chega à ordem de dezenas, centenas ou milhares, ou o conjunto de dados representado por elas alcança grandes dimensões (JUNIOR, 2006).

A seleção de variáveis em modelos torna-se importante por diversas características. Um modelo composto por elevado número de variáveis pode apresentar aderência satisfatória aos dados modelados, todavia não oferece garantias em termos de predição (devido ao demasiado número de variáveis independentes) e classificação (atribuído ao ruído inserido por variáveis com pouca relevância). A identificação de variáveis com base no conhecimento empírico de especialistas, por sua vez, é frequentemente sujeita a equívocos. Por fim, há preferência por modelos reduzidos, por demandarem menor tempo de análise e coleta, além de serem menos complexos (GUYSON e ELISSEEFF, 2003).

Para Gauchi e Chagnon (2001), modelos de predição apresentam melhores resultados após a aplicação de métodos de seleção de variáveis. Atributos irrelevantes, com pouco impacto sobre o modelo, acabam por prejudicar a real interação dos dados e diminuir a exatidão do modelo de predição. Allen (1974) reforça o uso de seleção de variáveis ao afirmar que a adição de variáveis não significativas aumenta a variabilidade do modelo e diminui a qualidade da predição. Thanassoulis (1996) ressalta que a modificação do conjunto de variáveis selecionadas poderá ter grande impacto no resultado da avaliação. Logo, torna-se relevante comparar diferentes métodos de seleção de variáveis. Há inúmeros métodos para seleção de variáveis na literatura, dentre os quais destacam-se o *Forward Selection, Backward Elimination* e *Stepwise Regression*. Tais modelos são descritos a seguir, conforme Montgomery e Peck (1992) e Hocking (1976):

Forward Selection: o modelo inicia sem variáveis na equação; variáveis estatisticamente significativas são incorporadas, uma a uma, ao modelo. A primeira variável a entrar no modelo (x_1) é aquela que possui a maior significância (normalmente medida através do teste *F*); as demais inclusões seguem o mesmo padrão.

Backward Elimination: ao contrário da *Forward Selection*, o modelo inicial conta com todas as variáveis. Na sequência, a variável tida como menos significativa (medida através do teste *F*) é extraída do modelo, e assim sucessivamente. Esse método é particularmente utilizado por analistas que querem observar os efeitos de todas as variáveis no modelo, e então de modelos parciais gerados pela eliminação sistemática de variáveis.

Stepwise Regression: integra as sistemáticas Forward e Backward; ao adicionar uma variável (x_{n+1}), uma variável (x_n) adicionada anteriormente pode tornar-se redundante por causa de sua relação com a variável (x_{n+1}) recém incluída na equação. Se provada a não significância da variável (x_n), ela é retirada do modelo.

Essa família de métodos apresenta vantagens como simplicidade e ampla difusão. Todavia, deve-se ter ciência que os procedimentos podem indicar um modelo que, não necessariamente, é tido como modelo ótimo. Tal colocação é corroborada por Mantel (1970), segundo o qual um excelente modelo pode passar despercebido devido à restrição de adicionar ou remover apenas uma variável por vez.

Outro método para seleção de variáveis é o S*imple Regression (SR)*. O SR inicialmente realiza regressões lineares simples entre a variável resposta (y) e cada variável de processo (x_i), adicionando-se à modelagem aquelas cujo coeficiente for estatisticamente significativo. O conjunto de variáveis resultantes é então utilizado no modelo final. Esse método assume que as variáveis selecionadas, mesmo que altamente correlacionadas entre si, influenciam significativamente a resposta de y (GAUCHI e CHAGNON, 2001).

2.3 Medidas de precisão e ajuste do modelo

O *Mean Squared Error (MSE)* ou média dos erros quadrados de uma estimativa permite quantificar a diferença entre o valor real e o valor estimado por um modelo. Na análise de regressão, o *MSE* mensura a média dos quadrados dos erros, descrevendo a variância do erro (soma residual dos quadrados dividido pelo número de graus de liberdade). O *MSE* pode ainda se referir à média dos valores dos desvios quadrados das predições comparadas aos valores reais, através da comparação de uma amostra gerada pelo modelo com valores reais naquele mesmo intervalo (WERDEMA e AGUIAR, 1996).

O Coeficiente de Determinação (R^2), apresentado na equação (6), mede a proporção da variação total da variável dependente *y* que é explicada pela variação da variável independente (*x*) (ou seja, quanto que a variação de *y* pode ser explicada pela variação de *x*). O coeficiente R^2 é sempre um número positivo no intervalo [0;1]; quanto mais próximo de 1 maior, a relação entre as variáveis. Todavia, um elevado valor de R^2 não significa necessariamente uma boa adequação do modelo aos dados observados, dado que a inclusão de variáveis pouco significativas no modelo podem elevar o valor do coeficiente (JOBSON, 1991; MONTGOMERY e PECK, 1992).

$$R^2 = \frac{SS_R}{S_{yy}} = 1 - \frac{SS_E}{S_{yy}} \qquad (6)$$

Onde:

R^2 - coeficiente de determinação ($0 \leq R^2 \leq 1$);

SS_R - soma dos quadrados da diferença entre a média da variável dependente e os valores estimados para todas as observações;

SS_E - soma dos quadrados dos resíduos para todas as observações;

S_{yy} - soma dos quadrados totais de *y*.

2.4 Logística de Distribuição e Indicadores de Desempenho

Segundo o Conselho em Gerenciamento Logístico (1998), logística é a parte do processo da cadeia de abastecimento que planeja, implanta e controla o fluxo eficiente e eficaz de matérias-primas, estoque em processo, produtos acabados e informações relacionadas, desde seu ponto de origem até o ponto de consumo, com o propósito de atender os requisitos dos clientes.

A operação de distribuição logística é responsável pela operacionalização da entrega física de produtos, gerenciamento e controle de custos e produtividade. As entregas físicas devem ser realizadas de maneira eficaz e eficiente, garantindo que os produtos sejam entregues em tempo hábil, utilizando os recursos disponíveis e reduzindo custos associados. As variáveis estudadas nesse trabalho provêm de bancos de dados do departamento de produtividade operacional de uma empresa de logística que, através de diversos indicadores de desempenho, planeja e executa melhorias nos processos de entrega. Figueiredo (2006) conceitua a logística enxuta como ampla e envolvendo ações que visam à criação de valor para os clientes mediante um nível de serviço logístico com o menor custo global.

Nos processos logísticos de distribuição, diversas variáveis exercem influência sobre o custo e qualidade do serviço prestado. As variáveis críticas, que exercem maior influência no serviço, são “promovidas” a indicadores de desempenho, pois a manutenção dessas variáveis em limites pré-definidos garante as características de qualidade do processo. Logo, nas organizações, os indicadores de desempenho são parte constituinte de atividades e procedimentos, fornecendo informações necessárias para o controle e melhoria.

Os indicadores de desempenho são essenciais para gerenciar o processo de melhoria contínua de uma empresa. Possíveis alterações no processo só poderão demonstrar real eficácia comparando-se os resultados obtidos com um histórico. Bandeira (1997) afirma que medir o desempenho de processo apenas se justifica quando existe o objetivo de aperfeiçoá-lo. Kaydos (1991) apresenta as diversas atribuições dos indicadores: (i) comunicar a estratégia e valores, (ii) identificar problemas e oportunidades, (iii) definir responsabilidade, (iv) guiar e mudar comportamentos, (v) tornar o trabalho realizado visível, (vi) favorecer o envolvimento das pessoas, (vii) servir de base para um sistema de remuneração. Para Ñauri (1998) os indicadores de desempenho permitem ainda oferecer tanto uma visão vertical como horizontal do desempenho organizacional. A visão vertical refere-se à gestão dos recursos da organização, enquanto que a visão horizontal refere-se à gestão de resultados.

Neste estudo, indicadores de desempenho logístico são utilizados como variáveis independentes em um modelo de regressão para predição de tempo em rota. O objetivo principal consiste na seleção dos indicadores mais relevantes para tal propósito. Os processos de distribuição envolvem indicadores que medem o desempenho da operação, sendo os seguintes indicadores abordados neste estudo (HIJJAR, 2005):

-Percentual de Insatisfação: Quantidade de clientes não atendidos.

-*Drop-Size*: Quantidade média descarregada por parada.

-Capacidade: Capacidade máxima suportada pelo veículo.

-Tempo em rota: Horas trabalhadas por equipe de entrega.

-Nível de serviço: Qualidade do serviço prestado percebido pelo cliente.

3. PROCEDIMENTOS METODOLÓGICOS

Este estudo caracteriza-se por sua natureza aplicada com abordagem quantitativa (YIN, 2003), dado que é realizado com bases de dados quantitativos sobre processos logísticos existentes e previamente registrados. Possui um objetivo exploratório e utiliza uma base bibliográfica que criará método para realizar um estudo de caso em empresa que realiza operações logísticas de distribuição. Na sequência são detalhadas as etapas do método para identificação dos indicadores mais relevantes para medição do desempenho do processo de distribuição.

3.1 Etapa 1 - Definição de Indicadores

Primeiramente, devem ser classificados os indicadores (também vistos como variáveis de processo ou independentes) capazes de descrever o processo de distribuição (caracterizado por uma variável de resposta apropriada). Tais indicadores podem ser segregados como: (i) temporais: variáveis que mensuram espaço de tempo; (ii) produtividade: mensuram eficiência dos processos; (iii) custo: atrelado a alocação de recursos financeiros e investimentos; (iv) distância: quantificam percurso percorrido; (v) nível de serviço: mensura aspectos de satisfação dos clientes.

3.2 Etapa 2 - Tratamento dos indicadores

As bases de dados disponíveis possuem diversas informações incorretas ou irrelevantes. Exemplos de informações incorretas incluem registros nos quais caminhões, após retornar de uma viagem, expressam kilometragem idêntica ou inferior à de saída em decorrência de avarias no odômetro do veículo ou de marcação equivocada. Como exemplo de informação irrelevante para o estudo, cita-se cadastros de clientes atendidos nas viagens (nome, endereço, CNPJ, telefone e preço dos produtos). Por fim, são identificados amostras que não são realizados com frequência e que podem distorcer o modelo, sendo caracterizados como fontes de ruído e, portanto, eliminados da base de dados.

3.3 Etapa 3 - Procedimento de Seleção de Variáveis e Avaliação do Modelo Gerado

Realizado o tratamento dos indicadores, os quatro métodos de seleção de variáveis descritos na seção 2.1 são utilizados para gerar uma lista de variáveis selecionadas. Tais variáveis são inseridas em modelos de regressão linear múltipla com vistas à predição da variável de resposta. Os métodos testados são: (i) $M_{S:}$ Método *Stepwise*; (ii) M_B: Método

Backward; (iii) M_F: Método *Forward* e (iv): M_{SR}: Método *Simple Regression*. Tais modelos são operacionalizados valendo-se de software estatístico.

No presente estudo, o modelo é gerado utilizando uma amostra parcial do banco de dados composto por 70% dos dados (porção treino). As variáveis são selecionadas com base nessa porção. O modelo resultando é aplicado aos 30% restantes (porção teste), as quais ilustram novas observações a serem preditas, como se esses dados não estivessem disponíveis quando realizado o estudo. A Figura 1 ilustra a divisão sugerida para o banco de dados.

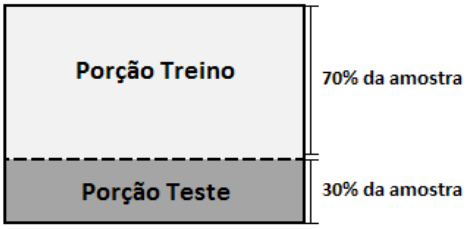

Figura 1- Representação porção treino/teste

3.4 Etapa 4 - Escolha do Melhor Método de Seleção

Para selecionar o melhor método de seleção, deve-se (*i*) avaliar a qualidade da aderência do modelo aos dados da porção de treino (através do Coeficiente de Determinação R^2), (*ii*) avaliar a qualidade da predição gerada pelo modelo proveniente de cada método de seleção de variáveis, o que é viabilizado pelo MSE de teste, e (*iii*) avaliar a parcimônia do modelo (ou seja, priorizar modelos retendo o menor número possível de variáveis, o que reduz a complexidade e o tempo despendido para coleta de dados e modelagem).

A escolha do melhor modelo de seleção é feita através da equação (7), a qual gera o Indicador para Seleção de Modelo (ISM). Os resultados de R^2 da porção treino apresentam caráter informativo da precisão do modelo utilizando dados históricos, sendo importante para atividades gerenciais e tomadas de decisão. O MSE da porção de teste tem caráter preditivo, informando o poder do modelo para quantificar o desempenho futuro com base em novos indicadores coletados do processo enquanto que o percentual de variáveis retidas informa o caráter de parcimônia do modelo, que se refere à característica de mitigar a quantidade de variáveis de um modelo no intuito de torná-lo mais simples nas suas análises e facilitar futuras regressões dos dados. Os coeficientes gerencial (ω_1), preditivo (ω_2) e de parcimônia (ω_3), sendo $\omega_1 + \omega_2 + \omega_3 = 1$, são definidos pelo usuário e permitem uma avaliação subjetiva na escolha do melhor método.

Quanto maior o valor de gerencial (ω_1) maior o caráter gerencial do modelo, dando maior relevância para a compreensão do comportamento das variáveis dos dados históricos. Já

o coeficiente (ω_2) define um caráter preditivo ao modelo, pois aumentará a participação dos indicadores que avaliam a qualidade de predição do modelo; o coeficiente (ω_3) define a relevância dada a modelos de menor complexidade. Ao avaliarem-se os quatro métodos, devem-se usar os mesmos valores para todos os coeficientes para não distorcer a avaliação, ou seja, os valores dos coeficientes devem ser escolhidos previamente conforme o interesse do usuário em informações gerenciais, preditivas e de parcimônia. O método com o menor ISM deve ser escolhido, pois conjuga a melhor modelagem à melhor predição.

$$ISMn = \left(\omega_1 . \frac{1}{R^2}\right) + \left(\omega_2 . \frac{MSE_{Teste}}{Ymédio_{Teste}}\right) + \left(\omega_3 . \frac{\#\ Variáveis\ Retidas}{\#\ Variáveis\ Total}\right) \quad (7)$$

Onde:

ISMn - Indicador para Seleção de Modelo para n modelos;

MSE_{Teste} - Valor médio dos resíduos obtido na porção teste;

R^2 - Coeficiente de Determinação;

ω_1 - Coeficiente gerencial;

ω_2 - Coeficiente preditivo;

ω_3 - Coeficiente de parcimônia;

Tais informações podem ser arranjadas conforme a Figura 2.

	$\omega_1 . \frac{1}{R^2}$	$\omega_2 . \frac{MSE_{Teste}}{Ymédio_{Teste}}$	$\omega_3 . \frac{\#\ Variáveis\ Retidas}{\#\ Variáveis\ Total}$	$Total$
$\mathbf{M_S}$	–	–	–	–
$\mathbf{M_B}$	–	–	–	–
$\mathbf{M_F}$	–	–	–	–
$\mathbf{M_{SR}}$	–	–	–	–

Figura 2 – Critérios avaliados no processo de seleção de variáveis

Além da análise do *ISM*, recomenda-se a construção de gráficos de resíduo para avaliar tendências na modelagem dos dados de treino. Os valores residuais de cada modelo devem estar aleatoriamente distribuídos em torno do eixo $\varepsilon=0$, demonstrando que o modelo é adequado. Valores residuais com sequência de dados crescentes, decrescentes ou contínuos apontam modelos tendenciosos que denotam uma modelagem inadequada. Se um modelo com elevado valor de ISM apresentar gráfico de resíduos tendencioso, o mesmo não deve ser

utilizado (MONTGOMERY e PECK, 1992). As Figuras 3 e 4 ilustram resíduos não-tendenciosos e tendenciosos, respectivamente.

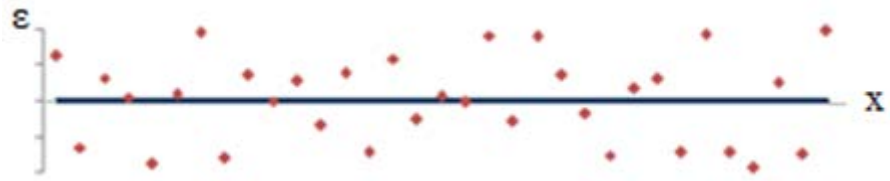

Figura 3 - Não Tendencioso - Erros aleatoriamente distribuídos

Figura 4 - Tendencioso - Erros seguem tendência

4. RESULTADOS

4.1 Caracterização do Ambiente

O estudo foi realizado em uma empresa de bebidas que fornece produtos para diversos pontos de venda na cidade de Porto Alegre, Rio Grande do Sul. Cada caminhão segue diariamente um roteiro pré-estabelecido e otimizado para entrega dos produtos. A rotina de entregas pode ser brevemente descrita como segue.

A rotina inicia com uma reunião, de aproximadamente 30 minutos, dos motoristas e ajudantes de entrega com seus superiores para apresentação dos resultados do dia anterior e informações do planejamento para o dia atual. Essas informações contemplam quantidade e descrição dos clientes que serão atendidos no dia, produtos que serão entregues para cada cliente e suas respectivas notas fiscais. No momento em que a equipe sai do Centro de Distribuição Direta (CDD) para iniciar as entregas, conferentes localizados na portaria marcam o tempo de saída de cada equipe, que será posteriormente utilizada para medir o tempo em rota.

A sequência de entregas é planejada previamente com vistas à minimização das distâncias percorridas e tempo despendido pelas equipes. Aproximando-se do ponto de venda, o motorista estaciona o caminhão em um local adequado, sinaliza através de cones que o veículo está parado para serviço, e inicia o processo de carga e descarga do veículo. Nesse

processo, a equipe procura o responsável pelo ponto de venda, apresenta as notas fiscais com os respectivos produtos a serem entregues e, dada a confirmação do cliente, a equipe descarrega os produtos que são levados para o depósito do cliente e recolhe os vasilhames, que são as garrafas de vidro que posteriormente serão reutilizadas. Efetuada a entrega de produtos, o cliente assina a nota fiscal, o motorista recebe o pagamento pelos produtos, enquanto os ajudantes recolhem os diversos materiais utilizados na entrega.

O processo de entrega repete-se em cada ponto de venda em que a equipe entregará produtos ao longo do dia. O planejamento prévio dessa rotina é, portanto, essencial para a otimização dos resultados, que são a quantidade de clientes atendidos bem como a satisfação do serviço prestado. Após o último cliente ser atendido, a equipe retorna ao CDD e, ao entrar na portaria, os conferentes marcam o tempo de retorno da equipe. Ao ingressar no CDD, o motorista entrega o dinheiro recebido ao longo do dia, enquanto os ajudantes auxiliam na descarga dos vasilhames e produtos que não puderam ser entregues, esses procedimentos duram aproximadamente 30 minutos.

4.2 Base de Dados

A base de dados utilizada no trabalho contempla um histórico dos indicadores usados para controle do processo de entrega nos meses de verão, onde a demanda por bebidas aumenta e os recursos para efetuar as entregas (caminhões e equipes) tornam-se escassos. A base de dados constitui-se de aproximadamente 25 Variáveis e 3000 registros. Dentre as informações presentes na base, utilizaram-se apenas aquelas relevantes para o estudo como quantidade de clientes e produtos entregues, tempo e kilometragem despendida para realização da rotina. Informações como nome, CNPJ, localização dos clientes não foram utilizadas, bem como informações referentes a valores dos produtos e remuneração das equipes de entrega. Excluindo-se essas informações restaram 13 variáveis. Toda e qualquer dado nesse estudo foi modificada de maneira a preservar informações sigilosas da empresa sem comprometer a relação entre as variáveis.

Efetuando a metodologia proposta, inicia-se a etapa de tratamento dos indicadores. Aproximadamente 59,15% dos registros não foram utilizadas, 55,78% por apresentarem inconsistências nos valores de kilometragem ou tempo despendido (ver Etapa 3.2) e os outros 3,37% restantes por representarem processos atípicos que ocorrem com menor frequência, como entrega de produtos em outras cidades ou regiões mais afastadas, que poderiam comprometer a qualidade da predição. Diversas variáveis utilizadas no banco de dados referiam-se ao mesmo tipo de indicador (por exemplo, há registro do horário de saída dos

caminhões do CDD e de regresso ao final do dia; o tempo despendido em rota é, portanto, o horário de regresso descontado o horário de saída). Logo, nessa etapa realizou-se uma pré-filtragem seleção subjetiva de variáveis, onde o conhecimento sobre o processo possibilitou a redução do número de variáveis, de 13 para 7 variáveis, a serem analisadas nas etapas seguintes.

4.3 Método de Seleção

As etapas seguintes valeram-se do software estatístico *IBM SPSS Statistics*, versão *student*. Os quatro métodos de seleção de variáveis foram testados com 70% dos dados disponíveis, que foram selecionados de forma aleatória para evitar distorções entre a porção treino e a teste.

A quantificação de cada porção que compõem o ISM bem como o resultado total é apresentada na figura 5. O método escolhido é o *Simple Regression*. As parcelas referentes à adequação da modelagem (R^2) e ao poder de predição do modelo (*MSE*) diferiram pouco entre os modelos, enquanto que a parcela de parcimônia do método *Simple Regression* teve o melhor resultado, garantindo, por uma diferença de 0,453% a escolha do *Simple Regression* como melhor método de seleção de variáveis.

	$\omega_1 \cdot \frac{1}{R^2}$	$\omega_2 \cdot \frac{MSE_{Teste}}{Ymédio_{Teste}}$	$\omega_3 \cdot \frac{\#\,Variáveis\ Retidas}{\#\,Variáveis\ Total}$	$Total$
$\mathbf{M_S}$	**0,95463**	**0,05117**	**0,14286**	**1,14865**
$\mathbf{M_B}$	**0,94683**	**0,05099**	**0,19047**	**1,18829**
$\mathbf{M_F}$	**0,94683**	**0,05099**	**0,19047**	**1,18829**
$\mathbf{M_{SR}}$	**0,99625**	**0,05196**	**0,09524**	**1,14344**

Figura 5 – Resultado dos critérios avaliados no processo de seleção de variáveis

A sutil diferença entre os resultados do ISM torna mais significante a escolha adequada dos valores dos coeficientes gerencial (w_1), preditivo (w_2) e de parcimônia (w_3). No estudo assumiu-se o valor de 0,333 para os três coeficientes, observa-se que, caso o estudo não tivesse interesse na simplicidade do modelo gerado e o coeficiente de parcimônia recebesse um valor suficientemente menor do que os demais, o resultado apontaria para os modelos *BackWard* e *Forward*, que possuem melhores resultados de adequação aos dados reais e poder de predição em detrimento de menor parcimônia.

O método de seleção de variáveis *Simple Regression* apontou que duas variáveis influenciam de maneira significativa a variável de resposta tempo em rota, sendo elas "número de entregas" e "quantidade de caixas carregadas". Por "número de entregas" entende-se o número de clientes atendidos durante o percurso; cada cliente atendido envolve um determinado tempo para estacionar o caminhão, apresentar nota fiscal, procedimentos de segurança entre outras tarefas. Já "quantidade de caixas carregadas" demonstra o volume de produtos que serão entregues ao longo do dia, quanto maior o volume de produtos a ser entregue, maior será o tempo despendido pela equipe para descarregar o caminhão.

É importante enfatizar que existe dependência entre as duas variáveis, não sendo possível efetuar uma entrega sem descarregar algum volume de produtos (ou recolher vasilhames), tampouco o caminhão pode sair do Centro de Distribuição Direta carregado de produtos sem efetuar nenhuma entrega. Nessas condições há restrição de ambas as variáveis serem maiores do que zero.

O modelo de regressão linear múltipla proveniente do método de seleção *Simple Regression* foi utilizado para compreender as relações entre as variáveis selecionadas e a variável de resposta tempo em rota, demonstrado pela equação (8).

$$TR\ (min) = 6{,}80.NE + 0{,}66667.\mathrm{QC} + 239{,}683 \qquad (8)$$

Onde:

TR - Tempo em Rota (em minutos);

NE - Número de Entregas;

QC - Quantidade de Caixas Carregadas.

A equação demonstra que a cada entrega realizada, o tempo em rota aumenta em 6,80 minutos, enquanto que cada caixa entregue pela equipe representa 0,66667 minutos despendidos, aproximadamente 40 segundos. A constante b_0, que quantifica o valor de tempo que independe dos valores das variáveis, assumiu o valor de 239,683 minutos, aproximadamente 4 horas. Nessa constante estão incluídos o tempo de almoço da equipe, com duração de uma hora, deslocamento do CDD até o primeiro ponto de venda e retorno da equipe ao final das entregas ao CDD (o tempo necessário para percorrer distância entre CDD e centro da cidade é de aproximadamente 40 minutos), volume de tráfego e o tempo despendido até encontrar local apropriado para estacionar o caminhão.

Comparando o modelo gerado com as informações reais (ver figura 6), tem-se que a média entre os valores reais e preditos pelo modelo de tempo em rota apresenta diferença apenas com digse 15 (casas após a vírgula). O 1º Quartil, que representa 25% da amostra ordenados de forma crescente, demonstra que essa parcela da amostra de dados preditos teve erro igual ou inferior a 5,00% (aproximadamente 30 minutos) em relação aos dados reais. Aumentando a parcela para 50% da amostra, representado pelo 2º Quartil (mediana), tem-se a informação que metade da amostra possui erro igual ou inferior a 10,45% (aproximadamente 1 hora e 3 minutos).

	Valores Reais	Valores Preditos	Erros	Erros (%)
Média	09:56:40	09:56:40	01:13:37	14,40%
Desvio Padrão	01:53:23	01:05:35	00:55:59	22,28%
Variância	00:08:56	00:02:59	00:02:11	4,96%
Valor Mínimo	01:39:00	05:46:00	00:00:01	0,00238%
Valor Máximo	13:00:00	12:54:42	08:34:11	451,03%
1° Quartil	08:51:00	09:18:12	00:30:28	5,00%
2° Quartil	10:07:00	09:58:33	01:03:03	10,45%
3° Quartil	11:24:00	10:43:47	01:44:09	17,55%

Figura 6 – Comparação entre valores reais e preditos

Como última etapa para avaliar a qualidade da modelagem, construiu-se o gráfico de resíduos na busca de informações que sinalizassem algum tipo de tendência dos erros. Conforme a figura 7, os erros estão aleatoriamente distribuídos ao redor do eixo zero, demonstrando que o modelo não gera informações tendenciosas.

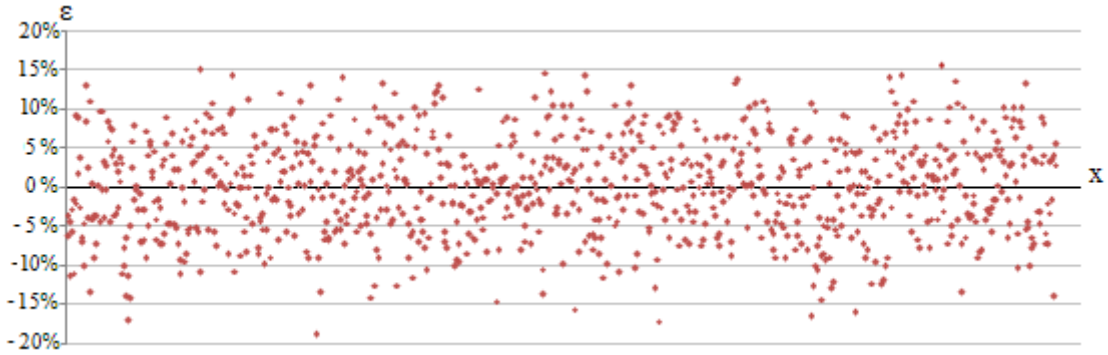

Figura 7 – Gráfico de dispersão; erros não seguem tendência

Por fim, criaram-se cenários hipotéticos pertinentes ao funcionamento do centro de distribuição com a função objetivo de maximizar, através do *Microsoft Excel Solver*, o número de caixas carregadas seguindo diferentes restrições de número de entregas respeitando

a capacidade de carga do caminhão e o limite de tempo em rota de 9h:20min. Obtiveram-se, portanto, os seguintes resultados:

- Restrições:

Quantidade de Caixas Carregadas: menor ou igual a 400 unidades.
Tempo em Rota: menor ou igual a 9h:20min.
Quantidade de Entregas: restrição conforme cenário.

- Cenário Caixas:

Nº de Entregas: 5 entregas
Quantidade de Caixas Carregadas: 400 caixas
Tempo em Rota: 09h00min
Quantidade média de caixas entregues: 80,0 caixas/entregas
Tempo despendido por entrega com caixas entregues média: 1hora

- Cenário Ponderado:

Nº de Entregas: 15 entregas
Quantidade de Caixas Carregadas: 326 caixas
Tempo em Rota: 09h19min
Quantidade média de caixas entregues: 21,7 caixas/entregas
Tempo despendido por entrega com caixas entregues média: 21min 27seg

- Cenário Entregas:

Nº de Entregas: 30 entregas
Quantidade de Caixas Carregadas: 174 caixas
Tempo em Rota: 09h19min
Quantidade média de caixas entregues: 5,8 caixas/entregas
Tempo despendido por entrega com caixas entregues média: 10min 40seg

O cenário “caixas” representa a situação de maximização da quantidade de caixas carregadas com um valor mínimo de clientes atendidos no processo de entregas em estudo. Os valores do cenário ponderado referem-se ao cenário de otimização de caixas carregadas e número de entregas, enquanto que o cenário “entregas” apresenta a quantidade de caixas com um valor máximo de clientes atendidos. Outros cenários podem ser criados dando maior relevância a uma variável em detrimento da outra. Quanto maior a quantidade de caixas

carregadas, melhor será a utilização do caminhão e equipe de entregas e, quanto maior o número de entregas, maior o número de clientes atendidos por dia.

5. CONCLUSÕES

O desconhecimento das relações entre as variáveis que compõem processos logísticos de distribuição compromete a eficácia de processos de melhoria ou manutenção de indicadores importantes para a produtividade de distribuição. Nesse cenário, o estudo objetivou a identificação das variáveis de maior relevância em processos logísticos de distribuição bem como compreensão do comportamento dessas variáveis em relação à variável de resposta. Utilizando métodos de seleção de variáveis e equação para eleger o melhor método de seleção, estratificou-se aquelas de maior relevância em relação à variável de resposta tempo em rota, que mede o tempo despendido pelas equipes de entrega para cumprir a rotina de distribuição de produtos em uma rota pré-definida.

Das variáveis utilizadas para controle do processo de distribuição, apenas “número de entregas” e “quantidade de caixas carregadas” possuem relevância perante a variável de resposta tempo em rota na cidade de Porto Alegre. A cada entrega, o tempo em rota aumenta aproximadamente 6,80 minutos, enquanto que o aumento de caixas carregadas representa acréscimo de 40 segundos. No caso hipotético de uma entrega de 15 caixas, o tempo despendido seria de 16,80 minutos. A constante beta zero, que quantifica o valor de tempo que independe dos valores das variáveis, obteve o valor de aproximadamente quatro horas, sendo tempo de almoço e deslocamentos os principais valores que compõem esse número.

Por fim, o estudo demonstrou a falta de comprometimento das equipes de entregas e daqueles responsáveis pela coleta de informações, conferentes e fiscais, para com a veracidade e qualidade das informações. 55,78% das informações presentes nos bancos de dados não puderam ser utilizadas por apresentarem diversas inconsistências (por exemplo o registro de caminhões que regressaram ao CDD após rotina de entrega com kilometragem igual ou inferior àquela de quando saíram para iniciar a rotina de entregas).

6. REFERÊNCIAS

ALLEN, D. M. **The Relationship Between Variable Selection and data Argumentation and a Method for Prediction**. *Yechnometrics*,1974.

BANDEIRA, A. A. **Rede de Indicadores de Desempenho para Gestão de uma Hidrelétrica.** Dissertação de Mestrado. São Paulo, 1997.

CHATTERJEE, S.; **PRICE**, B. **Regression Analysis by Example.** John Wiley & Sons, Inc, New York, 1991.

CLM, **Council of Logistics Management**, 1998. www.clm1.org

FIGUEIREDO, K. **A Logística Enxuta**. Centro de Estudos em Logística – COPPEAD / UFRJ, 2006.

GAUCHI, J.; **CHAGNON**, P. **Comparison of Selection Methods of Exploratory Variables in PLS Regression with Application to Manufacturing Process Data**. Chemometrics and Intelligent Laboratory Systems, 2001.

GUYSON, I.; **ELISSEEFF**, A. **An Introduction to Variable and Feature Selection**. Journal of Machine Learning Research, 2003.

HIJJAR, M. F. **Diagnóstico Externo do Desempenho Logístico: Utilizando Pesquisas de Serviço ao Cliente para Identificação de Oportunidades de Melhorias**. COPPEAD/UFRJ, 2005.

HOCKING, R. R. **A Biometrics Invited Paper. The Analysis and Selection of Variables in Linear Regression.** *Biometrics*, 1976.

HUGE, E. C. **Measuring and Rewarding Performance. Total Quality - An Executive's Guide for the 1990's**. Homewood, Illilois, Business One Irwin, 1990.

MONTGOMERY, Douglas C.; PECK, Elizabeth A. **Introduction to Linear Regression Analysis**. Nova Iorque: JOHN WILEY E SONS, INC. 1992.

JOBSON, J. D. **Applied Multivariate Data Analysis – Volume I: Regression and Experimental Design**. Springer-Verlag, New York, 1991.

JUNIOR, F. P. **Seleção de Variáveis e Características como Aplicação Paralela para Cluster MPI.** Dissertação de Mestrado. Maringá, 2006.

KAYDOS, W. **Performance Measurement and Performance Management. In: Measuring Managing and Maximizing Performance.** Portland: Productivity, 1991.

MANTEL, N. **Why Stepdown Procedures in Variable Selection**. *Technometrics*,1970.

MARTINS, Roberto A.; **NETO**, Pedro L. O. C. **Indicadores de Desempenho para a Gestão pela Qualidade Total: Uma Proposta de Sistematização.** Revista Gestão e Produção, 1998.

ÑAURI, Miguel Heriberto. **As Medidas de Desempenho como Base para a Melhoria Contínua de Processos**. Dissertação de Mestrado. Santa Catarina, UFSC, 1998.

THANASSOULIS, E. **Assessing The Efficiency of Schools with Pupils of Different ability using Data Envelopment Analysis.** *Journal of the Operational Research Society*, 1996.

YIN, Robert K. **Estudo de Caso: Planejamento e Métodos**. Porto Alegre: Bookman, 2003.

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