Modelos econométricos aplicados à previsão de demanda por transporte interestadual de passageiros de ônibus no Brasil

Mirian Buss Gonçalves[1]; Edson Tadeu Bez[2]; Antônio Galvão Novaes[3]

Resumo: O transporte rodoviário de passageiros, no Brasil, é um serviço público essencial. O grau de importância desse serviço pode ser avaliado quando se observa que o transporte rodoviário por ônibus é a principal modalidade na movimentação coletiva de usuários no território nacional. A dinâmica desse setor tem demandado a implantação de uma sistemática ágil de planejamento dos serviços a serem oferecidos aos usuários. Nesse contexto, a geração de cenários com estimativas do volume de movimentações é um dos aspectos mais relevantes a serem considerados no planejamento de novos serviços.

Neste artigo são apresentados modelos econométricos de previsão de demanda, que incorporam diversas variáveis sócio-econômicas e as variáveis do modelo gravitacional clássico. No procedimento de calibração fez-se uso da regressão linear múltipla e da regressão de erros absolutos utilizando os softwares Minitab e EGPA-FR (versão Evolucionária do Gradiente com Perturbações Aleatórias – com Fórmula de Representação) no processo de ajuste. O conjunto de dados utilizado na calibração, correspondente ao ano de 2003, foi obtido por meio de um processo de análise e sistematização de um banco de dados, constante dos Anuários (1999 – 2003) disponibilizados pela Agência Nacional de Transportes Terrestres – ANTT.

Os resultados finais obtidos neste experimento poderão ser usados pela ANTT para a geração de cenários com estimativas do volume de movimentações, o que vem a ser uma contribuição importante no processo de conhecimento da demanda por transporte interestadual por ônibus.

Abstract: The road transport of passengers in Brazil is an essential public service. Just how important this service is can be seen when we note that road transport by coach is the main means of collective traveling of users in Brazil. The dynamics of this sector has led to the implementing of an agile system of planning services to be offered to users. In this context, the generation of scenarios with estimates of passenger flows (of the amount of movement) is one of the most relevant aspects to be taken into account when planning new services.

In this article econometric models of forecasted demand which incorporate several socioeconomic variables and the variables of the classic gravitational model are presented. In the assessment calibration procedure multiple linear regression and the regression of absolute errors are employed, making use of the Minitab and EGPA-FR software (Evolutionary version of the Gradient with Random Perturbation – with the Representation Formula) in the adjustment process. The set of data used in the calibration assessment, corresponding to the year 2003, was obtained by means of a process of analysis and systematization of a database, appearing in the Yearly Publications (1999 – 2003) made available by the National Agency of Land Transport – ANTT.

The final results obtained from this experiment can be used by ANTT in order to generate scenarios with the objective of estimating passenger flows, which is an important contribution to the process of forecasting the demand for interstate transport by bus.

1. INTRODUÇÃO

O transporte rodoviário de passageiros, no Brasil, é um serviço público essencial. O grau de importância desse serviço pode ser avaliado quando se observa que o transporte rodoviário por ônibus é a principal modalidade na movimentação coletiva de usuários no território nacional. A dinâmica desse setor tem demandado a implantação de uma sistemática ágil de planejamento dos serviços a serem oferecidos aos usuários. Nesse contexto, a geração de cenários com estimativas do volume de movimentações é um dos aspectos mais relevantes a serem considerados no planejamento de novos serviços.

A demanda real por transporte de ônibus, em geral, não pode ser inferida diretamente dos volumes de passageiros transportados, pois parte dessa demanda pode utilizar outra modalidade de transporte ou, simplesmente, não viajar, em decorrência das características da oferta disponível (freqüência, tarifa, conforto, etc.) ou de suas próprias preferências.

Além disso, no caso dos países em desenvolvimento, muitas famílias não têm renda que possibilite viagens para lazer, turismo, visitar parentes etc. Dessa forma, os fluxos observados são apenas parte da demanda, traduzindo o ponto de equilíbrio entre a oferta e a demanda. No entanto, pode-se afirmar que a oferta de serviços de transporte por ônibus, observada hoje no Brasil, atende a demanda de forma aproximada, oferecendo freqüências, níveis de conforto e demais condições (nível de serviço) até certo ponto ajustadas às condições locais de demanda. Ademais, todo o sis-

[1] **Miriam Buss Gonçalves,** Departamento de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina. Florianópolis, SC, Brasil. (e-mail: mirianbuss@deps.ufsc.br).

[2] **Edson Tadeu Bez,** Universidade do Vale do Itajaí, Campus São José. São José, SC, Brasil. (e-mail: edsonbez@univali.br).

[3] **Antônio Galvão Novaes,** Departamento de Engenharia de Produção e Sistemas, Universidade Federal de Santa Catarina. Florianópolis, SC, Brasil. (e-mail: novaes@deps.ufsc.br).

Manuscrito recebido em 22/3/2007 e aprovado para publicação em 4/10/2007. Este artigo é parte de TRANSPORTES, volume XV, número 1, de junho de 2007. ISSN: 1415-7713.

tema de transporte interestadual por ônibus é controlado pelo mesmo órgão, a ANTT - Agência Nacional de Transportes Terrestres, que lhe impõe uma estrutura coerente e bastante uniforme.

Dessa forma, considera-se adequado o desenvolvimento de modelos de demanda diretos a partir dos fluxos existentes, para determinado corredor, admitindo implicitamente que constituem uma aproximação da demanda. Uma vez que se trabalha com viagens produzidas (fluxos observados), o modelo de análise da demanda é focalizado na distribuição dos fluxos. Há diversos tipos de modelo para representar o processo de distribuição dos fluxos (Ortúzar e Willumsen, 2001), sendo os mais difundidos os modelos de fatores de crescimento e os modelos sintéticos, especialmente o modelo gravitacional.

Geralmente esses modelos entram na segunda fase do método clássico das quatro etapas: geração, distribuição, repartição modal e alocação. Neste estudo, no entanto, propõe-se a utilização de um modelo econométrico mais amplo, que incorpora as variáveis do modelo gravitacional clássico, como um modelo de "demanda direto", contemplando as fases de geração e distribuição num só passo.

É importante ressaltar que os modelos agregados, categoria na qual se encaixa o modelo proposto neste estudo, foram amplamente utilizados nas décadas de 50 a 70 passadas, num grande número de estudos desenvolvidos principalmente na Inglaterra e nos Estados Unidos, tendo sido criticados no período que se seguiu (final da década de 70 até início da década de 90), por não fornecerem estimativas precisas. A partir da década de 90, muitos estudos retomaram a modelagem agregada (Horovitz e Farmer, 1998; Wirasinghe e Kumarage, 1998; Zhao et al, 2004).

No contexto brasileiro, a modelagem da demanda por transporte rodoviário interestadual de passageiros sempre teve grande importância para nortear as ações dos órgãos públicos encarregados da normatização e planejamento desses serviços. Os primeiros estudos foram realizados na década de 70, destacando-se o trabalho pioneiro de Luiz Eugênio D. Gomes. Recentemente, diversos outros estudos foram desenvolvidos, buscando aprimorar as estimativas fornecidas pelos modelos anteriormente usados (Consórcio STE-ENEFER, 2001; Fundação Universitária José Bonifácio, 2003). Nos diversos estudos, a modelagem tem em comum a adoção das variáveis contempladas no modelo gravitacional clássico e da técnica de regressão linear múltipla para a calibração dos modelos.

No presente trabalho essa base comum é mantida, buscando-se aprimorar os modelos de demanda através da incorporação de variáveis sócio-econômicas relevantes na formação dos fluxos e com a utilização de um conjunto de dados maior para a calibração dos modelos. No procedimento de calibração faz-se uso da regressão linear múltipla e da regressão de erros absolutos utilizando os softwares Minitab e EGPA-FR (Bez et al, 2005) no processo de ajuste. O conjunto de dados utilizado na calibração é bastante amplo, conforme será descrito na seção 3, e foi obtido por meio de um processo de análise e sistematização de um banco de dados, constante dos Anuários (1999 – 2003) disponibilizados pela ANTT.

O trabalho é estruturado como segue. Na seção 2 apresenta-se a especificação preliminar do modelo. Na seção 3 apresenta-se a análise e sistematização do banco de dados. Os procedimentos adotados na calibração do modelo e os resultados obtidos são apresentados na seção 4. Na seção 5 apresenta-se uma avaliação desses resultados e, finalmente, na seção 6, são feitas algumas considerações finais.

2. ESPECIFICAÇÃO DO MODELO

Ortúzar e Willumsen (2001) apresentam uma série de fatores que devem ser levados em conta na escolha da abordagem para modelar um problema de transporte. Entre eles destacam-se o contexto da tomada de decisão, a precisão requerida, a disponibilidade de dados adequados, o estado da arte da modelagem, os recursos disponíveis para o estudo e a necessidade de processamento de dados. Todos estes aspectos norteiam a escolha da modelagem adotada, em especial a disponibilidade de dados adequados. Como Ortúzar e Willumsen (2001) salientam, em alguns casos existem poucos dados disponíveis, em outros, podem existir razões para suspeitar da informação.

Uma análise preliminar foi desenvolvida pelos autores em relação aos dados existentes, tendo como fonte os Anuários Estatísticos da ANTT e os dados do senso demográfico do IBGE. Desta análise resultou um conjunto amplo de variáveis, que, na visão dos autores, poderiam explicar a demanda do sistema de transporte objeto do estudo. Também foi feita uma ampla revisão da modelagem que poderia ser adequada, resultando na seleção de um modelo agregado de demanda direto, levando-se em conta seu potencial de diminuição dos erros que normalmente ocorrem na modelagem seqüencial, quando não se têm modelos de geração precisos (Ortúzar e Willumsen, 2001). A opção foi por um modelo multiplicativo, do tipo gravitacional, modificado de forma a incorporar diversas variáveis sócio-econômicas, sem, no entanto, aumentar demasiadamente a complexidade do mesmo. A especificação inicial concebida é dada por:

$$T_{i,j} = \alpha_0 \frac{P_i^{\alpha_1} P_j^{\alpha_2} mig_i^{\alpha_3} mig_j^{\alpha_4} m_i^{\alpha_5} m_j^{\alpha_6} (r_i . r_j)^{\alpha_7} dummy_j^{\alpha_8}}{d_{ij}^{\alpha_9}} \quad (1)$$

onde, T_{ij} = fluxo anual total de passageiros entre os

municípios *i* e *j*; P_i, P_j = população dos municípios *i* e *j*, respectivamente; r_i, r_j = renda média per capita dos municípios *i* e *j*, respectivamente; m_i, m_j = número médio anual per capita de viagens interestaduais por ônibus dos municípios *i* e *j*, respectivamente; $dummy_j$ = variável que determina se o município é ou não um pólo turístico (valores adotados: $dummy_j = 2$, se *j* é um pólo turístico e $dummy_j = 1{,}2$ em caso contrário); mig_i e mig_j = índice de habitantes que não são naturais dos estados dos municípios de origem e destino, respectivamente, definido como o quociente entre o número de habitantes naturais de outros estados pelo número total de habitantes do município; d_{ij} = uma medida da distância rodoviária entre os municípios *i e j*; α_0, α_1, α_2, α_3, α_4, α_5 , α_6, α_7, α_8, α_9 = coeficientes a serem determinados.

Na equação (1) a zona de origem i corresponde ao município de menor população. Parte-se do pressuposto de que os habitantes dos municípios menores tendem a buscar bens e serviços em municípios maiores, de acordo com a Teoria do Lugar Central de Christaller (Gonçalves, 1992).

As variáveis mig_i e mig_j que correspondem ao índice de habitantes que não são naturais dos estados dos municípios de origem e destino, respectivamente, foram introduzidas para captar a influência de movimentos migratórios na formação dos fluxos, tendo em vista que houve uma migração intensa no Brasil, nas últimas décadas (Nordeste rumo a Sudeste; Sul rumo a Centro-Oeste, etc.) (Censo Demográfico 2000, 2003). Além disso, o motivo de viagem "visitar parentes" representa um peso percentual considerável (entre 35 e 40 por cento) do universo das viagens interestaduais de passageiros, conforme constatado em estudos recentes (Datamétrica, 2005; LabTrans/ UFSC, 2005). A variável *dummy*, que determina se o município é ou não um pólo turístico, é usada para captar a atratividade própria desses pólos. A renda média per capita foi adicionada ao modelo, pois pode explicar a geração de demanda ocasionada por fatores sócio-econômicos diversos.

A Equação 1 é usada como ponto de partida para a calibração, ou seja, é feita uma primeira regressão linear múltipla com todas as variáveis e, a seguir, com o auxílio de um software adequado (neste trabalho utilizou-se o software Minitab), identifica-se o melhor conjunto de variáveis para a especificação final do modelo. Para isso é usado um algoritmo para seleção de variáveis passo a passo (*stepwise regression)*, que testa, a cada passo, se a nova variável apresenta contribuição adicional. No conjunto final de variáveis selecionadas, não se espera que fiquem variáveis independentes que sejam correlacionadas.

Para que o modelo seja eficaz na determinação da variável preditora alguns pressupostos devem ser verificados, quais sejam: a) a normalidade dos resíduos, ou seja, os erros devem ser normalmente distribuídos; b) os erros devem possuir média zero; c) a homocedasticidade deve ser verificada, isto é, os erros devem possuir variância constante e; d) as variáveis explicativas não podem ser correlacionadas.

3. ANÁLISE E SISTEMATIZAÇÃO DO BANCO DE DADOS

Utilizou-se um amplo banco de dados, constante dos Anuários (1999 – 2003) disponibilizados pela ANTT. Esse conjunto de dados é composto de 236 empresas com 3.041 linhas o que corresponde a um total de 67.561 seções e uma movimentação anual (ano 2003) de passageiros de 70.514.322.

Uma análise preliminar do banco de dados mostrou que o mesmo continha registros que não faziam parte do escopo desta pesquisa, bem como ausência de informação em alguns casos. Decidiu-se, então, pela busca de critérios para eliminar esses registros.

No procedimento inicial adotado buscou-se obter uma maneira de avaliar as séries históricas, de modo a retirar uma amostra confiável que pudesse ser utilizada na calibração dos modelos de previsão da demanda e, para isso, foram retiradas do banco de dados as empresas que não operaram em 2003, tendo em vista que seriam utilizados os dados de 2003 para a calibração dos modelos. Do total de 236 empresas restaram 191. Em seguida, foram removidas as empresas que não tinham dados em alguns dos anos anteriores. Do total de 191 empresas restaram 142.

Foi, então, executado um processo de depuração desses dados, tendo como base as séries históricas (1999-2003) agregadas por empresa, realizando o cruzamento de informações através dos seguintes indicadores:

- Ocupação de veículo *(OV):*

$$OV = \frac{número\ de\ passageiros}{número\ de\ viagens} \tag{2}$$

- Quilometragem percorrida *(KP):*

$$KP = extensão \cdot número\ de\ viagens \tag{3}$$

- Índice de passageiros por quilômetro *(IPK):*

$$IPK = \frac{número\ de\ passageiros}{KP} \cdot 100 \tag{4}$$

O número de passageiros (*n.passageiros)* usado para determinar os índices *OV* e *IPK* foi obtido considerando-se o critério de proporcionalidade, como descrito a seguir.

Suponha que uma linha *i* tenha *N* seções. Denota-se: PAS_n = número de passageiros na seção *n*; *EXT* = extensão da linha *i*; EXT_n = extensão da seção *n*.

Uma medida do número de passageiros na linha *i*, $NPAS_i$, é dada por:

$$NPAS_i = \sum_{n=1}^{N} \frac{EXT_n}{EXT} \cdot PAS_n \quad (5)$$

Para analisar a regularidade desses indicadores no período de 1999 a 2003, foi calculado o coeficiente de variação (*CV*) (Spiegel, 1978) que é uma medida estatística que permite avaliar comparativamente empresas pequenas, médias e grandes. Este coeficiente é dado pela seguinte expressão:

$$CV = \frac{\sigma}{\overline{Y}} \quad (6)$$

onde $(\overline{Y})$ é a média e (σ) é o desvio padrão.

Buscou-se utilizar na calibração dos modelos os dados oriundos de empresas que apresentassem maior estabilidade temporal dos indicadores citados. Estabeleceu-se, então, o seguinte critério: caso uma empresa estivesse entre os 20% maiores valores do coeficiente de variação (*CV*) em 2 ou mais indicadores, os dados da mesma seriam descartados.

Do total de 142 empresas 26 empresas foram descartadas usando o critério anteriormente descrito, restando 116 e, com o objetivo de verificar a consistência dos dados dessas empresas, decidiu-se realizar uma análise multivariada (regressão múltipla).

Para isso foi usada a seguinte expressão:

$$P_j = \alpha_0 \left(C_j OV_j KP_j\right)^{\alpha_1} NM_j^{\alpha_2} \quad (7)$$

onde, P_j = número de passageiros transportados pela empresa *j*; C_j = capacidade do veículo (46 lugares) da empresa *j*; OV_j = ocupação média do veículo (em porcentagem) da empresa *j*; KP_j = quilometragem percorrida pela empresa *j*; NM_j = índice de motoristas da empresa *j* por quilômetro percorrido pela mesma, definido como o quociente entre o número total de motoristas da empresa e a quilometragem total percorrida; α_0, α_1 e α_2 são coeficientes a serem determinados.

Para esta análise foi utilizado um conjunto de dados adicional, referente às empresas que operaram no ano de 2003, disponibilizado pela ANTT. Em virtude da inexistência de informações relativas ao número de motoristas, mais duas empresas foram descartadas, restando um total de 114 empresas.

Para efetuar a regressão múltipla, foi aplicado à Equação 7 o logaritmo, resultando na expressão:

$$\log(P_j) = \log(\alpha_0) + \alpha_1 \log\left(C_j OV_j KP_j\right) + \alpha_2 \log\left(NM_j\right) \quad (8)$$

Na Tabela 1 apresentam-se os resultados estatísticos da regressão. Observa-se que o coeficiente de determinação obtido garante que aproximadamente 70% da variabilidade da variável independente (número de passageiros transportados) é explicada pelo modelo de regressão e que os parâmetros, exceto α_0 apresentam significância estatística.

Tabela 1. Resultados da regressão múltipla utilizando-se a Equação 8

Coeficientes	*Stat t*	*valor-p*
α_0 = 0,353734	0,493882	0,622365
α_1 = 0,721392	15,85074	2,8E-30
α_2 = 0,449665	3,150764	0,002093
R^2 = 0,694184		

Nos testes que verificam a eficácia da variável preditora do modelo, a homocedasticidade foi verificada, os erros se apresentaram normalmente distribuídos e os demais pressupostos, que garantem a eficiência do modelo, foram satisfeitos.

Dessa forma, as variáveis usadas na regressão foram validadas. Foi estabelecido, então, como último critério de descarte, retirar as empresas em que esta relação foi muito fraca (resíduo padronizado > 2). Como conseqüência foram descartadas mais 6 empresas, restando um total de 108 empresas.

Esse processo de limpeza do banco de dados acarretou a seleção de um conjunto de empresas com suas respectivas linhas e seções (pares OD). Foi realizada, a seguir, uma análise da representação espacial da movimentação de passageiros da amostra resultante. A representação, que não é apresentada neste texto por delimitação de espaço, levou em conta a movimentação de passageiros em cada região (Sul, Sudeste, Centro-oeste, Norte e Nordeste) e a movimentação de passageiros com origem e destino em regiões diferentes, formando um total de 15 ligações intra e inter-regionais.

Tendo em vista que a amostra resultante contemplou um número de pares OD de todas as ligações intra e inter-regionais aproximadamente proporcional à movimentação de passageiros dessas ligações, optou-se por utilizar o banco de dados resultante da depuração realizada para a calibração dos modelos de demanda, que é relatada na seção que segue.

4. PROCEDIMENTOS ADOTADOS NA CALIBRAÇÃO DO MODELO E RESULTADOS OBTIDOS

Utilizando-se a equação (1) realizou-se uma primeira regressão linear múltipla usando todas as variáveis contempladas nessa equação. A seguir, com auxílio do *software* Minitab, uma série de combinações entre essas variáveis foi desenvolvida, buscando-se resultados melhores.

Devido à extensão do território brasileiro e suas acentuadas diferenças regionais, segmentou-se o universo de análise por agrupamentos das ligações regionais e inter-regionais. O agrupamento das ligações

regionais e inter-regionais foi feito baseando-se na análise de Pareto (curva ABC), usando como variável a movimentação anual de passageiros das seções.

O princípio de Pareto é também conhecido como regra 80-20 (aproximadamente 80% dos efeitos são decorrentes de 20% das causas). Por exemplo, 80% do faturamento de uma empresa provêm de aproximadamente 20% de seus produtos; 80% do suprimento é comprado de mais ou menos 20% dos seus fornecedores, etc. Em Administração, a forma mais popular da análise ABC propõe a divisão dos itens objeto de estudo em 3 classes: A ($\sim 10\%, \sim 65\%$), B ($\sim 20\%$, $\sim 25\%$) e C ($\sim 70\%$, $\sim 10\%$), onde o símbolo ~ representa aproximadamente (Wild, 1997).

No presente estudo, a análise ABC demonstra que a maior parte da movimentação anual de passageiros concentra-se num pequeno grupo de ligações regionais e/ou inter-regionais. Busca-se, então, concentrar o esforço nos grupos de ligações com maior movimentação de passageiros, pois bons resultados nesses grupos têm grande alcance no universo de estudo.

Da segmentação realizada com base na análise de Pareto, resultaram três grupos. GRUPO A: (correspondendo a 67,43% de movimentação total): sudeste (29.841.934), sul (9.773.281), nordeste (7.934.130); GRUPO B: (correspondendo à porcentagem acumulada de 67,43% até 94,41%): sul – sudeste (6.115.716), sudeste – centro-oeste (4.709.540), centro-oeste (4.548.952), sudeste – nordeste (3.654.311); e GRUPO C: (correspondendo à porcentagem acumulada de 94,41% até 100%): norte (1.006.904), centro-oeste – norte (800.904), sul – centro-oeste (678.875), centro-oeste – nordeste (579.598), norte – nordeste (554.755), sudeste – norte (232.170), sul – nordeste (52.366), sul – norte (30.886).

A seguir, cada um dos grupos A, B e C, foi segmentado por distância, considerando-se as seguintes faixas: viagens curtas: até 400 km; viagens médias: de 400 a 800 km; e viagens longas: com mais de 800 km. Essa classificação por distância vem sendo adotada em diversos estudos (Consórcio STE-ENEFER, 2001), e é justificada devido ao grande intervalo representado por cada faixa e à percepção empírica de que, nas diferentes faixas, as viagens têm motivação e aspectos comportamentais diferenciados.

Gerou-se, dessa forma, um total de 9 (nove) grupos segmentados por agrupamentos regionais e inter-regionais e por distância. Dividiu-se, a seguir, o conjunto de ODs de cada grupo em dois subconjuntos de mesma magnitude: um para ser usado na calibração do modelo (Amostra 1) e outro na validação dos coeficientes obtidos (Amostra 2).

Com a segmentação realizada obtiveram-se, no processo de calibração, os resultados apresentados nas Tabelas 2, 3 e 4, onde são identificadas as variáveis, todas com significância estatística, que explicam a movimentação de passageiros nos diferentes grupos.

Para os grupos A, B (todas as faixas de distância) e C (de 0 a 400 km), os valores de R^2 obtidos encontram-se entre 50 e 60%. Esses valores, bom como a

Tabela 1. Resultados do Grupo A (S, SE, NE) segmentado por distância considerando a movimentação de passageiros (67,43 %)

Faixa de distância (km)	Até 400	400 – 800	Mais de 800
N. total de observações (seções)	2204	1421	476
Amostra (50 % das observações)	1102	711	238
Variáveis	$P_i, P_j, m_i, m_j,$ $d_{ij}, dummy_j$	$P_i, P_j, m_i, m_j,$ $r_i r_j, d_{ij}, dummy_j$	P_i, P_j, m_i, m_j d_{ij}
R^2	58,4	55,3	52,0

Tabela 2. Resultados do Grupo B (S-SE, SE-CO, CO, SE-NE) segmentado por distância considerando a movimentação de passageiros (67,43% até 94,41%)

Faixa de distância (Km)	Até 400	400 – 800	Mais de 800
N. total de observações (seções)	660	737	1596
Amostra (50 % das observações)	330	369	798
Variáveis	$P_i, P_j, m_i, m_j,$ d_{ij}, mig_i	$P_i, P_j, m_i, m_j,$ d_{ij}, mig_i, mig_j	$P_i, P_j, m_i, m_j,$ d_{ij}, mig_i, mig_j
R^2	55,7	59,1	53,8

Tabela 3. Resultados do Grupo C (N, CO-N, S-CO, CO-NE, N-NE,SE-N, S-NE, S-N) segmentado por distância considerando a movimentação de passageiros (94,41% até 100%)

Faixa de distância (Km)	Até 400	400 – 800	Mais de 800
N. total de observações (seções)	166	201	1324
Amostra (50 % das observações)	83	101	662
Variáveis	P_i, P_j, m_j, d_{ij}	P_i, P_j, m_i, d_{ij}	P_i, P_j, m_i, d_{ij}
R^2	55,9	35,9	30,0

normalidade dos resíduos (Figura 1), se mostraram um pouco insatisfatórios. Decidiu-se, então, retirar as observações discrepantes (*outliers*) e validar a sua retirada. Em seguida calibrou-se o modelo sem estas observações e verificou-se uma melhoria dos valores de R^2 (Tabelas 5, 6 e 7). A melhoria na normalidade dos resíduos após a retirada dos *outliers* é exemplificada na Figura 2.

Para validar o modelo obtido em cada grupo com a retirada dos *outliers*, foi utilizada a Amostra 2 (metade das observações reservadas para validação dos modelos). Para isso foram obtidas as estimativas fornecidas pelo modelo, para os pares OD dessa amostra, usando

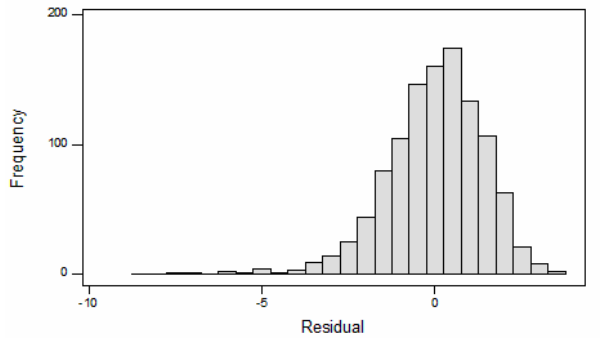

Figura 1. Histograma dos resíduos do Grupo A – sem retirada dos *outliers*

Tabela 5. Valores de R^2 no GRUPO A, com e sem a retirada dos *outliers*

Faixa de distância (km)	*Até 400*	*400 – 800*	*Mais de 800*
R^2 com os *outliers*	58,4	55,3	52,0
R^2 sem os *outliers*	67,2	65,1	59,2

Tabela 6. Valores de R^2 no GRUPO B, com e sem a retirada dos *outliers*

Faixa de distância (km)	*Até 400*	*400 – 800*	*Mais de 800*
R^2 com os *outliers*	55,7	59,1	53,8
R^2 sem os *outliers*	63,1	70,0	53,3

Tabela 7. Valores de R^2 no GRUPO C, com e sem a retirada dos *outliers*

Faixa de distância (km)	*Até 400*
R^2 com os *outliers*	55,9
R^2 sem os *outliers*	62,0

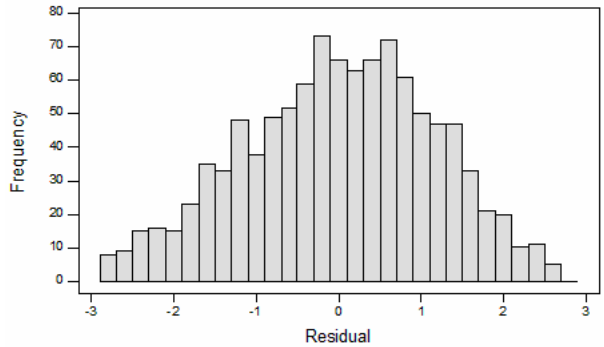

Figura 2. Histograma dos resíduos do Grupo A – com retirada dos *outliers*

os coeficientes obtidos com e sem as observações discrepantes. A seguir, foram calculadas as estatísticas PHI-normalizada (PHI), Erro Médio Absoluto Normalizado (EMAN), Índice de Dissimilaridade (ID), que são estatísticas clássicas para avaliar o ajuste de matrizes de distribuição de viagens. Quanto menor o valor de PHI, EMAN ou ID, melhor é o ajuste entre as matrizes de viagens observadas e estimadas. Uma descrição das mesmas pode ser encontrada em Gonçalves, 1992 ou em Bez, 2000.

Na Tabela 8, são exemplificados os resultados para o grupo A, na faixa de distância de 0 a 400 km. Nos demais grupos os modelos resultantes da calibração sem os *outliers* também apresentaram uma melhor performance quando aplicados na amostra 2, o que justifica a sua adoção.

Tabela 8. Resultados do ajuste do modelo com e sem a retirada dos *outliers* (Grupo A, Até 400 km)

	PHI	*ID*	*EMAN*
Com *outliers*	1,22	34,79	757,18
Sem *outliers*	1,07	33,30	724,62

Em razão dos baixos valores obtidos para R^2 na calibração do Grupo C, para viagens médias e longas, utilizou-se o modelo em sua forma multiplicativa. A calibração foi feita usando a estatística Índice de Dissimilaridade (Equação 9), sendo utilizado o método EGPA-FR (Bez *et al*, 2005) para a minimização dessa estatística.

$$ID = \frac{50}{T^*} \cdot \sum_{ij} \left| T_{ij}^* - T_{ij} \right| \qquad (9)$$

onde, T_{ij}^* = número de viagens observadas da origem i para o destino j; T_{ij} = número de viagens estimadas da origem i para o destino j; T^* = Total de viagens observadas.

Salienta-se que a minimização de uma estatística que considera os módulos das diferenças (ID) entre os valores observados e estimados é considerada mais adequada em diversas situações, sendo menos sensível a valores aberrantes (Narula e Stangenhaus, 1988).

Como a nova regressão realizada leva em consideração os erros absolutos (Regressão *L1*), para análise da significância estatística dos coeficientes encontrados, tornou-se necessário a adoção de uma nova estatística que substituísse a estatítisca *t de student*, utilizada na regressão de mínimos quadrados, já que esta se baseia na soma de quadrados. De acordo com Narula e Stangenhaus (1988), o teste *t de student* não é adequado em uma regressão de erros absolutos. Os autores apresentam a seguinte estatística, que pode ser usada nesse caso:

$$Z_* = \left| \frac{b_i}{\lambda \sqrt{(X'X)_{ii}^{-1}}} \right| \quad (10)$$

onde b_i é o i-ésimo coeficiente obtido na regressão, $(X'X)_{ii}^{-1}$ é o i-ésimo elemento da diagonal de $(X'X)^{-1}$, onde X é a matriz formada pelas variáveis independentes e λ é um valor desconhecido. Um estimador sugerido para λ é dado por:

$$\hat{\lambda} = \frac{e(t) - e(s)}{2(t-s)/n} \quad (11)$$

onde $e(1) \leq e(2) \leq ... \leq e(n)$ representam os resíduos ordenados. É recomendado que os valores de t e s sejam escolhidos de tal maneira que estejam posicionados simetricamente em torno do índice da mediana amostral, ou seja,

$$\begin{cases} t = [(n+1)/2] + v \\ s = [(n+1)/2] - v \end{cases}$$ no caso em que n é ímpar, e

$$\begin{cases} t = [n/2] + 1 + v \\ s = [n/2] - v \end{cases}$$ no caso de n par.

onde [•] indica a parte inteira do número e v é escolhido de maneira que t e s sejam simétricos em relação ao índice da mediana amostral. A diferença entre t e s deve ser pequena e $e(t)$ e $e(s)$ não podem ser nulos. Salienta-se, também, que $\hat{\lambda}$ é um estimador consistente de λ (Narula e Stangenhaus, 1988). Maiores detalhes sobre a utilização da regressão *L1* na calibração de modelos de distribuição de viagens podem ser vistos no trabalho de Bez e Gonçalves (2006).

A calibração do modelo foi bem sucedida para as duas faixas de distância consideradas (Grupo C, 400 a 800 km e mais que 800 km), resultando na equação (12), que representa a forma clássica do modelo gravitacional:

$$T_{i,j} = \alpha_0 \frac{P_i^{\alpha_1} P_j^{\alpha_2}}{d_{i,j}^{\alpha_3}} \quad (12)$$

Na Tabela 9 apresenta-se uma análise comparativa dos dois modelos (regressão de mínimos quadrados e regressão de erros absolutos) para a faixa de distância de 400 a 800 km, evidenciando-se um melhor ajuste do modelo obtido pela regressão de erros absolutos, o que justifica a sua adoção.

Tabela 9. Análise comparativa dos modelos obtidos pela regressão de mínimos quadrados e regressão de erros absolutos, para o grupo C, faixa de distância de 400 a 800 km

Estatísticas	*Regressão de mínimos quadrados*	*Regressão de erros absolutos*
ID	41,04	19,61
EMAN	82,91	39,62
PHI	1,57	0,74

A seguir são apresentados os resultados finais para os 9 (nove) grupos (Tabelas 10, 11, 12 e 13). A Equação 12 é usada no Grupo C com distâncias de 400 a 800 km e maior que 800 km e nos demais grupos a Equação 13, que segue:

$$T_{ij} = \exp\left[\alpha_0 + \alpha_1 \ln\left(\frac{P_i}{1000}\right) + \alpha_2 \ln\left(\frac{P_j}{1000}\right) + \alpha_3 \ln\left(mig_i\right) + \alpha_4 \ln\left(mig_j\right) + \alpha_5 \ln\left(m_i\right) + \alpha_6 \ln\left(mig_j\right) + \alpha_7 \ln\left(\frac{r_i \cdot r_j}{10000}\right) + \alpha_8 \ln\left(d_{ij}\right)\right] \quad (13)$$

Tabela 10. Parâmetros do GRUPO A: Sudeste, Sul e Nordeste

Parâmetros	*Até 400*	*400 – 800*	*Mais de 800*
α_0	8,301365873873120	8,087857093549980	8,060233479808390
α_1	0,796942591205408	0,746394122270084	0,768417612202321
α_2	0,635112753161796	0,460060954344267	0,749215460295464
α_5	0,392490583052169	0,389003078693068	0,324061268619131
α_6	0,350311335187946	0,284395978815848	0,304697160914189
α_7	0	0,187844704235045	0
α_8	0,768546997382765	1,583012375455010	0
α_9	-1,317742790589580	-1,386710533881670	-1,309897388857810
$\alpha_3 = \alpha_4 = 0$			

Tabela 11. Parâmetros do GRUPO B: Sul - Sudeste, Sudeste – Centro-Oeste, Centro-Oeste, Sudeste – Nordeste

Parâmetros	*Até 400*	*400 – 800*	*Mais de 800*
α_0	9,669452781533640	15,547490974495700	2,463275144398500
α_1	0,938684186940491	0,878143234255628	0,765201860744870
α_2	0,729338172445292	0,703106804331187	0,726892058542628
α_3	0,386203240623749	0,459425537749364	0,224353812843686
α_4	0	0,430361198729195	0,679006362917358
α_5	0,383598091603631	0,238797735493225	0,176601251936519
α_6	0,147531198903630	0,443097286036165	0,191253171835200
α_9	-1,565279941665680	-2,287268593352390	-0,343487785426711
$\alpha_7 = \alpha_8 = 0$			

Tabela 12. Parâmetros do GRUPO C: Norte, Centro-Oeste – Norte, Sul – Centro-Oeste, Centro-Oeste – Nordeste, Norte – Nordeste, Sudeste – Norte, Sul – Nordeste, Sul – Norte

Parâmetros	*Até 400*
α_0	12,461258153384400
α_1	0,575401236367286
α_2	0,896521019315957
α_6	0,328571443851490
α_9	-2,043542076826620
α_3, α_4, α_5, α_7, α_8	0

Tabela 13. Parâmetros do GRUPO C: Norte, Centro-Oeste - Norte, Sul - Centro-Oeste, Centro-Oeste - Nordeste, Norte - Nordeste, Sudeste - Norte, Sul - Nordeste, Sul - Norte

Parâmetros	*400 – 800*	*Mais de 800*
α_0	5,2761875456E-04	0,16185606032509
α_1	0,12160295556711	1,01032526542223
α_2	1,18590486757524	0,79690438393203
α_3	0,38992882494686	1,93144653444439

5. AVALIAÇÃO DO RESULTADO

Inicialmente é importante destacar que não houve uma busca por um modelo único para todo o território nacional, pois se partiu do pressuposto que seria interessante ter um modelo para cada subconjunto. Uma das vantagens de calibrar um modelo único é a utilização de uma amostra única. No entanto, quando o banco de dados contém um número significativo de registros, não há prejuízo em usar segmentações, desde que em cada segmento se tenha um número de observações adequado.

Em todos os grupos observou-se a significância estatística das variáveis contempladas. Além disso, os sinais obtidos para os coeficientes, em todos os grupos, são coerentes, isto é, estão de acordo com o esperado pela especificação do modelo. A retirada das observações discrepantes permitiu um melhor ajuste, tendo sido validada através da aplicação do modelo na Amostra 2 (metade das observações reservadas para a validação dos modelos).

As condições para o uso da regressão linear múltipla também foram verificadas. A correlação de variáveis foi evitada através do algoritmo de seleção de variáveis passo a passo e a homocedasticidade foi testada para todos os grupos. Isso foi feito observando-se se os resíduos têm variância relativamente constante e com distribuição próxima do normal. Nos testes realizados não houve evidência de variância não constante.

Os valores obtidos para R^2 mostram que entre 60 e 70% da variabilidade da variável preditora foi explicada pelos modelos finais adotados (Tabelas 5, 6 e 7). Embora esses valores possam ser considerados baixos em diversas situações, nesse estudo eles podem ser considerados razoáveis, tendo em vista a grande diversidade do universo de pares OD analisados.

Por fim foi feita uma avaliação dos resultados obtidos pelo modelo novo comparativamente ao modelo desenvolvido pelo consórcio STE-ENEFER (2001) (Equação 14).

$$\ln(Demanda) = b_0 + b_1 \cdot \ln(Pop\_A) + b_2 \cdot \ln(Pop\_B) + b_3 \cdot \ln(Dist) + b_4 \cdot Atrat + b_5 \cdot \mathrm{Re}g \quad (14)$$

onde, *Pop_A* = População municipal da localidade de maior demografia, em milhares de habitantes; *Pop_B* = População municipal da localidade de menor demografia, em milhares de habitantes; *Dist* = Distância rodoviária entre localidades, em quilômetros; *Atrat* = Atratividade (o valor atribuído a atratividade é 1(um) nos casos de ligações com características de polaridade espacial e 0 (zero) em caso contrário; *Reg* = Assume diretamente o valor 0 (zero) para as ligações Norte, Sudeste, Sul, Centro-Oeste, Norte – Centro-Oeste, Norte – Nordeste, Nordeste – Sudeste, Centro-Oeste – Sudeste e Centro-Oeste – Sul e assume valor 1 (um) nas ligações Nordeste, Nordeste – Centro-Oeste, Nordeste – Sul, Norte – Sul e Sudeste – Sul.

Para isso, foram retiradas, aleatoriamente, amostras com 100 pares de OD, de cada um dos nove agrupamentos realizados, isto é, para cada uma das nove equações do modelo de previsão de demanda proposto. A seguir foram calculadas as estimativas de demanda fornecidas pelo novo modelo. Também foram calculadas as estimativas de demanda dessas ODs, utilizando o outro modelo citado. Para analisar a performance dos modelos, foi utilizada a estatística Índice de Dissimilaridade (Equação 9).

Os valores dessa estatística são sempre positivos, sendo zero quando todas as estimativas coincidirem com os dados observados. No entanto, é difícil fazer uma interpretação dos valores absolutos obtidos. Quanto mais próximo de zero for seu valor, melhor o ajuste (ou seja, mais próximas dos dados observados estarão as estimativas fornecidas pelo modelo). Em geral, essa estatística é usada para comparar os resultados obtidos por modelos distintos, como é o caso do presente estudo.

Cabe ressaltar, no entanto, que na situação particular onde o número total de viagens estimadas T coincidir com o número de viagens observadas T^*, a estatística Índice de Dissimilaridade tem uma interpretação muito interessante. Neste caso seus valores variam entre 0 e 100 e ela mede a porcentagem de viagens que necessitam ser realocadas entre os pares de origem-destino (ODs), a fim de que a matriz de viagens estimada coincida com a matriz de viagens observada (Gonçalves, 1992).

Na Tabela 14 apresentam-se os resultados obtidos. Observa-se que em todos os agrupamentos os resultados fornecidos pelo novo modelo são melhores. Finalmente, foi feita uma comparação referente à porcentagem de seções onde os modelos subestimam a

Tabela 14. Comparativo entre o modelo novo (M.N.) e o e o modelo do consórcio STE-ENEFER (M.C.), usando a estatística índice de dissimilaridade (equação 9)

	GRUPO A			***GRUPO B***			***GRUPO C***		
	Até 400	400 – 800	Mais de 800	Até 400	400 – 800	Mais de 800	Até 400	400 – 800	Mais de 800
M. N.	32,41	28,73	35,18	33,26	33,59	33,59	41,92	41,25	35,02
M. C.	33,36	103,89	71,59	48,54	52,87	52,87	72,85	56,87	53,13

Tabela 15. Porcentagens de estimativas superestimadas e subestimadas pelos modelos

	GRUPO A					
	Até 400		400 – 800		Mais de 800	
MODELOS	superestima	subestima	superestima	subestima	superestima	Subestima
M. N.	44%	56%	56%	44%	54%	46%
M. C.	80%	20%	90%	10%	84%	16%
	GRUPO B					
	Até 400		400 – 800		Mais de 800	
MODELOS	superestima	subestima	superestima	subestima	superestima	Subestima
M. N.	52%	48%	46%	54%	52%	48%
M. C.	82%	18%	86%	14%	80%	20%
	GRUPO C					
	Até 400		400 – 800		Mais de 800	
MODELOS	superestima	subestima	superestima	subestima	superestima	Subestima
M. N.	65%	35%	41%	59%	40%	60%
M. C.	87%	13%	76%	24%	74%	26%

demanda observada ou superestimam (Tabela 15). Analisando essa tabela observa-se que o modelo anterior superestimou a demanda numa grande porcentagem das seções da amostra usada, enquanto para o modelo proposto isso não ocorre.

Esse fato é considerado positivo, pois diminui o risco de serem obtidos indicativos de exploração autônoma de uma linha e, posteriormente, a mesma não se mostrar viável. Uma das possíveis causas que explicam esse fato é o número significativo de ODs usados nas amostras utilizadas para a calibração dos modelos, tendo em vista que a ampla maioria dos municípios brasileiros são pequenos ou médios, e que os mesmos foram contemplados na mesma proporção que os municípios maiores.

6. CONSIDERAÇÕES FINAIS

O presente trabalho atendeu ao seu objetivo inicial, obtendo modelos de previsão de demanda para o transporte rodoviário interestadual de passageiros, que podem ser usados para avaliar a demanda potencial de uma nova linha interestadual de ônibus.

A utilização de um modelo de demanda direto mostrou-se interessante, tendo em vista a não existência de dados de geração de viagens disponíveis e o alto custo para obtê-los. Um estudo exploratório do banco de dados dos municípios do IBGE permitiu uma análise das variáveis que afetam os padrões da demanda, identificando-se as principais variáveis que influenciam na formação dos fluxos e ao mesmo tempo podem ser usadas na modelagem por estarem disponíveis no bando de dados.

A proposta de agrupar as regiões e ligações interregionais, em grupos, baseando-se na análise de Pareto, demonstrou ser um procedimento adequado. Este fato fez com que se obtivesse um resultado satisfatório, para a maioria dos casos, identificando-se as variáveis que explicam a movimentação de passageiros e definindo-se as equações que representam o modelo de previsão demanda.

Além das variáveis clássicas do modelo gravitacional, que foram estatisticamente significantes em todos os grupos, destacaram-se: o número médio anual per capita de viagens interestaduais por ônibus, que foi representativo em todos os grupos, as variáveis que captam a influência dos fluxos migratórios (grupo B) e a atratividade dos pólos turísticos (grupo A, faixas de distância até 400 km e entre 400 e 800 km).

Desenvolveu-se uma análise comparativa das estimativas fornecidas pelo modelo proposto e um modelo anterior, utilizado pela ANTT, na qual se pode verificar que o modelo proposto apresentou melhor performance.

Os resultados obtidos poderão ser usados pela ANTT para a geração de cenários com estimativas do volume de movimentações, o que é um dos aspectos mais relevantes a serem considerados no planejamento de novos serviços.

Com o desenvolvimento das novas tecnologias de informação, cada vez mais é possível obter melhores informações dos sistemas reais. Cabe à Agência Reguladora e às empresas aprimorarem seus procedimentos de coleta de dados. Com isso melhores resultados poderão ser obtidos através dos modelos de previsão de demanda, que devem ser revistos e atualizados periodicamente.

AGRADECIMENTOS

Os autores agradecem à ANTT (Agência Nacional de Transportes terrestres) e ao LabTrans - UFSC (Laboratório de Transportes da Universidade Federal de Santa Catarina pela cooperação no trabalho desenvolvido. Agradecem, também, ao CNPq e à FAPESC por patrocinarem parcialmente o presente trabalho.

REFERÊNCIAS BIBLIOGRÁFICAS

Bez, E. T. (2000) *Um estudo sobre os procedimentos de calibração de alguns modelos de distribuição de viagens*. Dissertação (Mestrado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis.

Bez, E. T., Souza de Cursi, J. E., Gonçalves, M. B. (2005) A Hybrid Method for Continuous Global Optimization Involving the Representation of the Solution. *6th World Congress on Structural and Multidisciplinary Optimization – WCSMO6*. Rio de Janeiro, RJ, Brazil.

Bez, E. T., Gonçalves, M. B. (2006) Utilização da regressão de erros absolutos na calibração de modelos de distribuição de viagens.. In: XX Congresso de pesquisa e ensino em transportes, 2006, Brasília. Panorama Nacional da Pesquisa em Transportes 2006. Rio de Janeiro : ANPET, v. I. p. 455-461.

Censo Demográfico 2000 – Migração e Deslocamento (2003), IBGE, Rio de Janeiro.

Consórcio STE-ENEFER (2001) "*Relatório Técnico Reavaliação do modelo de estimativa da demanda de passageiros em ligações de transporte rodoviário*" – Estudo realizado por GISTRAN para ANTT.

Datamétrica – Consultoria, Pesquisa e Telemarketing (2005). "*Relatório de Transporte Rodoviário de Passageiros – Pesquisa de avaliação da satisfação dos usuários dos serviços das empresas de transporte terrestre*" para ANTT.

Fundação Universitária José Bonifácio (2003) "*Relatório 03 – Modelos de Demanda*" – Consultoria Técnica, para a ANTT, visando o desenvolvimento do transporte rodoviário interestadual e internacional de passageiros.

Gonçalves, M. B. (1992) *Desenvolvimento e Teste de um Novo Modelo Gravitacional – de Oportunidades de Distribuição de Viagens*. Tese (Doutorado em Engenharia de Produção) – Programa de Pós-Graduação em Engenharia de Produção, Universidade Federal de Santa Catarina, Florianópolis.

Horowitz, A.J, e Farmer, D.D (1998). *A Critical Review of Statewide Travel Forecasting Practice*, University of Wisconsin. Disponível em: <http://my.execpc.com/~ajh/Statewid.pdf>.

Labtrans/UFSC (2005) "Relatório 5 – Estudo Piloto" Convênio 018/2004 – Modelo de viabilidade, monitoramento e representação dos indicadores de desempenho das linhas de transporte rodoviário de passageiros, entre a UFSC e a ANTT.

Narula, S.C., Stangenhaus, G. (1988) *Análise de Regressão* L_1 . 8º Sinape. Campinas.

Novaes, A. G. (1982) *Modelos em planejamento urbano, regional e de transportes*. Edgard Blucher, São Paulo.

Ortúzar, J. D., Willumsen, L. G. (2001) *Modelling Transport*. Third Edition. John Wiley & Sons, New York.

Spiegel, M. R. (1978) *Probabilidade e Estatística*. Makron Books Editora Ltda., São Paulo.

Wild, T. (1997) *Best Practice in Inventory Management*. John Wiley & Sons, New York.

Wirasinghe, S. C., Kumarage, A.S. (1998) An aggregate demand model for intercity passenger travel in Sri Lanka, *Transportation***,** vol. 25, 77-98.

Zhao F., Chow L. F., Li, M. T. e Gan A. (2004) Refinement of Fsutms Trip Distribution Methodology. *Final Report*, Prepared by Lehman Center for Transportation Research, Department of Civil & Environmental Engineering, Florida International University.