Campus de São Carlos

**GERAÇÃO DE VIAGENS DE PASSAGEIROS URBANOS:
MODELO INDIVIDUAL COM A INCLUSÃO DA INFLUÊNCIA
DAS CARACTERÍSTICAS DO DOMICÍLIO**

MARIA DA CONCEIÇÃO FARIAS FREITAS TANDEL

ORIENTADOR: PROF. DR. EIJI KAWAMOTO

UNIVERSIDADE DE SÃO PAULO

ESCOLA DE ENGENHARIA DE SÃO CARLOS

**GERAÇÃO DE VIAGENS DE PASSAGEIROS URBANOS:
MODELO INDIVIDUAL COM A INCLUSÃO
DA INFLUÊNCIA DAS CARACTERÍSTICAS DO DOMICÍLIO**

MARIA DA CONCEIÇÃO FARIAS FREITAS TANDEL

Tese apresentada à Escola de Engenharia de
São Carlos, da Universidade de São Paulo,
como parte dos requisitos para a obtenção do
Título de Doutor em Engenharia de
Transportes.

DEDALUS - Acervo - EESC

31100016399

ORIENTADOR: Prof. Dr. Eiji Kawamoto

São Carlos
1997

Clas.	TESE - EESC
Cutt.	5952
Tombo	T158/97

Área Transportes

st 0927483

Ficha catalográfica preparada pela Seção de Tratamento
da Informação do Serviço de Biblioteca - EESC-USP

T163g Tandel, Maria da Conceição Farias Freitas
Geração de viagens de passageiros urbanos :
modelo individual com a inclusão da influência das
características do domicílio / Maria da Conceição
Farias Freitas Tandel. -- São Carlos, 1997.

Tese (Doutorado). -- Escola de Engenharia
de São Carlos-Universidade de São Paulo, 1997.

Área: Transportes

Orientador: Prof. Dr. Eiji Kawamoto.

1. Geração de viagens. 2. Produção de viagens.
3. Taxas de viagens individuais. 4. Modelos
individuais. 5. Modelos desagregados. 6. Demanda
por viagens. 7. Demanda por transportes.

I. Título.

FOLHA DE APROVAÇÃO

Candidata: Bacharela **MARIA DA CONCEIÇÃO FARIAS FREITAS TANDEL**

Tese defendida e aprovada em 18-8-1997
pela Comissão Julgadora

Prof. Associado **EIJI KAWAMOTO (Orientador)**
(Escola de Engenharia de São Carlos - Universidade de São Paulo)

Prof. Doutor **JOSÉ REYNALDO ANSELMO SETTI**
(Escola de Engenharia de São Carlos - Universidade de São Paulo)

Prof. Titular **CARLOS ALBERTO FARIA**
(Universidade Federal de Uberlândia - UFU)

Profª. Doutora **SILVIA NAGIB ELIAN**
(Instituto de Matemática e Estatística - Universidade de São Paulo)

Profª. Doutora **YAEKO YAMASHITA**
(Universidade de Brasília - UnB)

Prof. Doutor **ANTÔNIO NELSON R DA SILVA**
Coordenador da Área de Transportes

Prof. Titular **JOÃO VITOR MOCCELLIN**
Vice-Presidente da Comissão de Pós-Graduação
em exercício

*A Luiz Alves de Freitas e
Roque Yuri Tandel*

*meu pai, que me possibilitou iniciar a carreira docente;
meu esposo, que sempre incentiva a prosseguir.*

AGRADECIMENTOS

Profundos e sinceros agradecimentos

- Ao Departamento de Transportes da EESC-USP pela agradável acolhida e pelo avanço que proporcionou a minha carreira acadêmica;
- Ao Prof. Dr. Eiji Kawamoto (EESC - USP - São Carlos, SP) pela orientação e pela oportunidade de viver experiências ímpares, que sem dúvida contribuíram para meu crescimento pessoal e profissional;
- Ao Prof. Dr. Manoel Alba Soria (EESC - USP - São Carlos, SP) pela inestimável contribuição da disciplina Metodologia de Pesquisa;
- A Cia do Metropolitano-METRÔ pelo fornecimento dos dados da pesquisa Origem/Destino-1987, bem como pelo excelente atendimento do Departamento de Planejamento dos Transportes Metropolitanos, representado por Sr. Epaminondas Duarte Coelho, Maria Cecília Laiza, Emília, Eliza e pelo Sr. João Carlos Scatena;
- A CAPES- Coordenação de Aperfeiçoamento de Pessoal de Nível Superior, pela bolsa do PICD- Programa Institucional de Capacitação de Docentes;
- Aos colegas: docentes e funcionários do DEMAC-IGCE UNESP, Rio Claro, SP, pela amizade e sobretudo pela idealização e operacionalização do LAI- Laboratório Avançado de Informática, cujos recursos computacionais contribuíram decisivamente para a elaboração desta tese;
- A Profª Dra. Sílvia Nagib Elian (IME - USP - São Paulo, SP), pelo acompanhamento, sugestões, apoio e confiança;
- Ao Prof. Dr. José R. A Setti (EESC - USP - São Carlos, SP) pela minuciosa leitura, e valiosas correções, sugestões, apoio e confiança;
- Ao Sr. Marcelo Corrêa Alves (CIAGRI - USP - Piracicaba, SP) pelas preciosas sugestões no uso do SAS;
- A Cristiane Piza Raiado e Kléber Wolney Raiado pelo profissionalismo e qualidade dos excelentes serviços de digitação e editoração;
- E ao apoio incomensurável de minha família:
 - Joana Anita de Farias Freitas, minha mãe pelo auxílio e amor;
 - Luiz Alves de Freitas, meu pai e meus irmãos pela colaboração;
 - Jovina Maria Tandel, minha sogra, pelas incansáveis orações;
 - Maria do Carmo de Arruda e Maria Adalgisa P. do Nascimento, fiéis e dedicadas ajudantes;
 - Roque Yuri Tandel, meu esposo, pela compreensão, resignação e incentivo;
 - Yria Freitas Tandel, minha filha de 2 anos, pelo descanso em seu sorriso.

PREFÁCIO

Pode-se provar cientificamente uma hipótese ?

“Observei toda a obra de Deus, e vi que o homem não é capaz de descobrir toda a obra que se realiza debaixo do sol, por mais que o homem trabalhe pesquisando, não a descobrirá. E mesmo que um sábio diga que conhece, nem por isso é capaz de descobrir” (Eclesiastes 8:17¹).

“Em Deus se acham escondidos todos os tesouros da sabedoria e do conhecimento” (Colossenses 2:3¹).

Kawamoto (1987 p. 89 a 93) destaca a importância da apresentação das principais linhas de pensamento quanto à verificação de uma nova hipótese, assunto que historicamente têm influenciado os pesquisadores.

Lê-se:

“Uma das primeiras linhas de pensamento sobre o assunto é creditada a Francis Bacon que delineou, em sua obra “Novun Organun” publicada em 1620, a doutrina que foi aceita durante muito tempo como uma definição correta dos procedimentos a serem usados para o progresso do conhecimento científico. A visão baconiana do método científico, segundo Hegenberg (1976), consiste na idéia de que a ciência se desenvolve por meio de processos indutivos: o investigador principia suas pesquisas pela observação, registrando-as; em seguida, com cautela, passa a fazer generalizações

¹ A Bíblia Sagrada na tradução: A Bíblia de Jerusalém, edições Paulinas, 1985

gradualmente mais amplas, adotando para isso um processo indutivo, que o conduz, fase após fase, do particular para o geral; a confirmação das leis obtidas depende apenas de constante verificação, em casos concretos, de certo “acordo” entre o que a generalização afirma e o que efetivamente acontece.

Outra visão do método científico é sustentada por Popper. Segundo esse pensador, não se pode jamais atingir uma generalização a partir dos dados da observação. Propôs, então, para resolver esta questão, um sistema hipotético-dedutivo baseado no raciocínio de que por maior que seja o número de observações a confirmação de uma hipótese é impossível; no entanto ela pode ser conclusivamente refutada. Para Popper (1975), portanto, a ciência avança propondo teorias audaciosas, desenvolvendo todos os esforços para refutá-las e só aceitando provisoriamente aquelas teorias que sobrevivem a esse processo. De acordo com sua opinião, como a finalidade da ciência é a refutação, não a confirmação, só se tem a ganhar se os pontos fracos de uma conjectura forem expostos, em lugar de escondê-los. Com efeito, Popper afirma que embora não se deva abandonar as generalizações levianamente, diante de qualquer aparente refutação, é preciso que as generalizações sejam refutáveis em princípio. Essa visão tem recebido várias críticas. Beveridge (1981), por exemplo, afirma existir quatro pontos seriamente vulneráveis no esquema de Popper: a) Popper não se ocupa da origem da hipótese - o cerne da descoberta científica; b) a política de concentrar-se na refutação de hipótese, de demolir teorias, é negativa em muitos aspectos; c) a tentativa de refutação da hipótese, tema central da filosofia poperiana, depara na prática, com a dificuldade de projetar um experimento realmente crucial; e d) a ênfase na afirmação de que as únicas hipóteses válidas são aquelas suscetíveis de teste e de possível refutação é exagerada, visto que muitas hipóteses valiosas são impossíveis de serem testadas.

Numa posição intermediária entre pensamentos baconiano e poperiano, em relação à presença da indução na marcha do desenvolvimento científico, encontra-se Hempel. Para esse autor, o “indutivismo estrito” de Bacon é insustentável porquanto dados empíricos não são relevantes para um problema, a menos que se tenha uma hipótese; em última análise, é a hipótese que conduz às coletas de dados e análise dos fatos, e procura esclarecer como os fenômenos se entrelaçam. O conhecimento científico, de acordo com Hempel (1981) é obtido pela invenção de hipóteses como tentativas de resposta ao problema em estudo e submissão dessas hipóteses à verificação empírica.

Com relação à avaliação da aceitabilidade de uma hipótese, existem, atualmente, algumas propostas de medidas da aceitabilidade de hipóteses, entre as quais figuram a idéia da verossimilhança. Popper (1972) afirma que embora nunca se possa saber com certeza se uma teoria científica é verdadeira ou provável, porque ela pode ser sempre refutada pelo próximo teste, é possível a uma série de teorias, mesmo quando refutadas, acercarem-se sucessivamente da verdade.

As dificuldades observadas na estimativa da verossimilhança nos casos de redução teórica tornam-se maiores quando se trata de duas teorias concorrentes, igualmente bem apoiadas até agora pela experiência, sem que se possa dizer portanto que a evidência disponível favoreça mais a uma que a outra. De acordo com Hempel (1981), a decisão entre as duas pode ser obtida mediante “verificação crucial”. Esse tipo de verificação consiste em conceber uma situação para a qual essas hipóteses predigam resultados incompatíveis. Presume-se que o experimento refute uma das hipóteses e sustente a outra. Porém, em nossa opinião, nem mesmo a verificação crucial é decisiva, porquanto nada garante que a “situação crucial” seja única e que o resultado sempre favoreça a mesma teoria.

A discussão acima leva a concluir que tentativas de provar a superioridade de uma teoria em relação a outra esbarram em sérias dificuldades, ainda mais quando uma das teorias é nova, com poucas evidências desfavoráveis mas também com poucas evidências favoráveis” (Kawamoto, 1987 p.89-93).

Ora, dentro deste contexto de pensamento científico sobre teorias bem estabelecidas (ou não), meu objetivo é levar o leitor a refletir que afirmações como estas a seguir, mesmo que veiculadas freqüentemente pelos meios de comunicação de massa, não tornam verdade o que é teoria, mesmo partindo de cientistas de renome como é o caso:

“Para especialistas, teoria de Darwin sofreu alterações mas a essência continua correta...

... O primeiro e mais consistente estudo sobre a evolução, de Charles Darwin, data de 1856 e já passou por várias adaptações...

... A descoberta de Darwin é um fato já cristalizado no meio científico e na sociedade geral...

... Todo o cientista aceita o Darwinismo...

... A teoria da evolução ainda é um dos pilares da ciência...

... Os cientistas sempre defenderam a teoria evolucionista”².

² Segatto, Cristiane Fala do papa sobre a evolução é tardia, diz cientista O Estado de São Paulo, São Paulo, 25 de outubro de 1996 ano 117 nº 37627 caderno geral seção Igreja p.A14.

Ora, diante da hipótese de que Darwin possa ter errado, ao menos no tocante às origens da vida na terra, o que você me diz disto:

“No princípio criou Deus os céus e a terra.

E a terra era sem forma e vazia; e *havia* trevas sobre a face do abismo; e o Espírito de Deus se movia sobre a face das águas.

E disse Deus: Haja luz. E houve luz.

E viu Deus que era boa a luz; e fez Deus separação entre a luz e as trevas.

E Deus chamou à luz Dia; e às trevas chamou Noite. E foi a tarde e a manhã o dia primeiro.

E disse Deus: Haja uma expansão no meio das águas, e haja separação entre águas e águas.

E fez Deus a expansão, e fez separação entre as águas que *estavam* debaixo da expansão e as águas que *estavam* sobre a expansão. E assim foi.

E chamou Deus à expansão Céus, e foi a tarde e a manhã o dia segundo.

E disse Deus: Ajuntem-se as águas debaixo dos céus num lugar; e apareça a *porção* seca. E assim foi.

E chamou Deus à *porção* seca Terra; e ao ajuntamento das águas chamou Mares. E viu Deus que era bom.

E disse Deus: Produza a terra erva verde, erva que dê semente, árvore frutífera que dê fruto segundo a sua espécie, cuja semente esteja nela sobre a terra. E assim foi.

E a terra produziu erva, erva dando semente conforme a sua espécie, e a árvore frutífera, cuja semente *está* nela conforme à sua espécie. E viu Deus que era bom.

E foi a tarde e a manhã o dia terceiro.

E disse Deus: Haja luminares na expansão dos céus, para haver separação entre o dia e a noite; e sejam eles para sinais e para tempos determinados e para dias e anos.

E sejam para luminares na expansão dos céus, para alumiar a terra. E assim foi.

E fez Deus os dois grandes luminares: o luminar maior para governar o dia, e o luminar menor para governar a noite; e fez as estrelas.

E Deus os pôs na expansão dos céus para alumiar a terra,

E para governar o dia e a noite, e para fazer separação entre a luz e as trevas. E viu Deus que era bom.

E foi a tarde e a manhã o dia quarto.

E disse Deus: Produzam as águas abundantemente répteis de alma vivente; e voem as aves sobre a face da expansão dos céus.

E Deus criou as grandes baleias, e todo o réptil de alma vivente que as águas abundantemente produziram conforme as suas espécies; e toda a ave de asas conforme a sua espécie. E viu Deus que era bom.

E Deus os abençoou, dizendo: Frutificai e multiplicai-vos, e enchei as águas nos mares; e as aves se multipliquem na terra.

E foi a tarde e a manhã o dia quinto.

E disse Deus: Produza a terra alma vivente conforme a sua espécie; gado e répteis, e bestas-feras da terra conforme a sua espécie. E assim foi.

E Deus fez as bestas-feras da terra conforme a sua espécie, e o gado conforme a sua espécie, e todo o réptil da terra conforme a sua espécie. E viu Deus que era bom.

E disse Deus: Façamos o homem à nossa imagem, conforme à nossa semelhança; e domine sobre os peixes do mar, e sobre as aves dos céus, e sobre o gado, e sobre toda a terra, e sobre todo o réptil que se move sobre a terra.

E criou Deus o homem à sua imagem; à imagem de Deus o criou; macho e fêmea os criou.

E Deus os abençoou, e Deus lhes disse: Frutificai e multiplicai-vos, e enchei a terra, e sujeitai-a; e dominai sobre os peixes do mar, e sobre as aves dos céus, e sobre todo o animal que se move sobre a terra.

E disse Deus: Eis que vos tenho dado toda a erva que dá semente, que *está* sobre a face de toda a terra; e toda a árvore, em que há fruto de árvore que dá semente, ser-vos-á para mantimento.

E a todo o animal da terra, e a toda a ave dos céus, e a todo o réptil da terra, em que há alma vivente, toda a erva verde *será* para mantimento. E assim foi.

E viu Deus tudo quanto tinha feito, e eis que era muito bom: e foi a tarde e a manhã o dia sexto.

ASSIM os céus, e a terra e todo o seu exército foram acabados.

E havendo Deus acabado no dia sétimo a sua obra, que tinha feito, descansou no sétimo dia de toda a sua obra, que tinha feito.

E abençoou Deus o dia sétimo, e o santificou; porque nele descansou de toda a sua obra, que Deus criara e fizera” (Gênesis 1 a 2:1-3³).

“Lembra-te do teu Criador nos dias da tua mocidade, antes... que o pó volte a terra, como o era, e o espírito volte a Deus, que o deu” (Eclesiastes 12:1 e 7³)

Pense nisto !!!

³ Bíblia Sagrada - Traduzida em português por João Ferreira de Almeida, Edição revista e corrigida pela Sociedade Bíblica do Brasil em 1969.

SUMÁRIO

Dedicatória	ii
Agradecimentos	iii
Prefácio.....	iv
Lista de tabelas	x
Lista de gráficos.....	xii
Lista de figuras	xiii
Resumo.....	xiv
Abstract	xv
1. INTRODUÇÃO	1
1.1 Objetivos.....	5
1.2 Justificativa para a escolha do tema.....	6
1.2.1 Complementação da justificativa do tema: breve estudo das transições importantes no padrão de atividades das famílias da Região Metropolitana de São Paulo - SP.....	8
1.3 Métodos de pesquisa	19
1.4 Apresentação dos capítulos	20
2. A ANÁLISE DA DEMANDA POR TRANSPORTES	22
2.1 Os modelos de demanda por transportes.....	24

3. GERAÇÃO DE VIAGENS	30
3.1 Enfoques atuais para a geração de viagens	32
3.1.1 Modelos agregados.....	34
3.1.2 Modelos desagregados	39
3.1.2.1 Modelos baseados em domicílios	39
3.1.2.2 Modelos baseados em indivíduos	43
3.1.3 Considerações sobre modelos desagregados	48
3.2 Discussão dos métodos analíticos usuais para a modelagem de geração de viagens	49
3.2.1 A análise de categoria	50
3.2.2 A análise de regressão linear	54
3.2.3 O modelo linear geral.....	56
3.2.4 Comparações entre os métodos analíticos.....	58
4. INFLUÊNCIA DAS CARACTERÍSTICAS DO DOMICÍLIO SOBRE O INDIVÍDUO	64
4.1 Inserção deste tema no contexto geral de análise de demanda por transportes	64
4.2 A necessidade de maior compreensão do homem em seu meio ambiente	68
4.3 A importância das características domiciliares e do grupo familiar inserido no contexto da sociologia e psicologia social.....	69
4.4 A importância do tema na etapa de geração de viagens.....	71
4.5 Cuidados a serem tomados na obtenção de resultados quantitativos	79

5. O ESTUDO DO CASO DA REGIÃO METROPOLITANA DE SÃO PAULO: ANÁLISES ESTATÍSTICAS DESCRITIVAS.....	83
5.1 Descrição sumária da pesquisa origem/destino - Metrô - SP - 1987.....	83
5.1.1 O plano de amostragem da O/D 1987- Metrô	84
5.1.2 Dados coletados pela O/D - 1987 - Metrô	87
5.1.3 Características gerais dos dados coletados.....	89
5.2 Escolha das variáveis a serem analisadas.....	89
5.2.1 Estrutura familiar e índice do ciclo de vida das famílias: ICV	91
5.3 Análises estatísticas preliminares dos dados da Região Metropolitana de São Paulo	96
5.3.1 Resultados da análise estatística descritiva dos dados individualizados.....	99
5.3.1.1 Correlação de Pearson.....	100
5.3.1.2 Classificação dos indivíduos segundo as variáveis de interesse: distribuição de frequências.....	101
5.4 Número de viagens diárias individuais - análise estatística descritiva	113
5.4.1 Caracterização dos indivíduos com número de viagens diárias igual a zero.....	115
5.4.2 Número médio de viagens diárias individuais segundo os níveis das variáveis de interesse	117
5.4.3 Análise da distribuição do número de viagens diárias individuais segundo alguns cruzamentos dos níveis das variáveis domiciliares.....	122
5.4.3.1 Descrição do <i>box plot</i>	123

6. VERIFICAÇÃO ESTATÍSTICA DA TESE E ANÁLISE DE SUA VIABILIDADE.....	130
6.1 Regressão linear múltipla: notas metodológicas.....	130
6.1.1 Métodos de seleção de variáveis e testes da contribuição de uma ou mais variáveis	135
6.2 Significância estatística da contribuição ao modelo das variáveis ligadas à influência das características domiciliares sobre o indivíduo	141
6.2.1 Alterações mediante a inclusão das variáveis: número de automóveis e ocupação principal	152
6.3 Detalhamento da estimação de um modelo individual	161
6.4 Discussão dos resultados obtidos com os modelos gerados a partir dos dados da O/D-1987 - Metrô	167
7. A CONTRIBUIÇÃO EFETIVA DESTE TRABALHO	172
8. CONCLUSÕES, RECOMENDAÇÕES, SUGESTÕES.....	175
ANEXO A - ANÁLISE DE REGRESSÃO LINEAR - UM RESUMO DESCRITIVO E INTERPRETATIVO.....	182
A.1 O modelo de regressão linear.....	182
A.2 Desvios das hipóteses do modelo de regressão: causas e diagnóstico (detecção)	184
A.2.1 Causas dos desvios das suposições	184
A.2.2 Diagnóstico: detecção dos desvios das suposições	185

A.2.2.1 A análise descritiva dos dados	186
A.2.2.2 Análise de resíduos.....	187
A.2.2.3 Testes de hipóteses	191
A.3 Multicolinearidade: definição, fontes, efeitos e detecção	192
A.3.1 O conceito de multicolinearidade	192
A.3.2 As fontes de multicolinearidade.....	193
A.3.3 Efeitos da presença de multicolinearidade.....	194
A.3.4 Detecção e medidas do grau de multicolinearidade.....	195
A.4 <i>Outliers</i> : definição, causas, detecção	200
A.5 Comparação e escolha entre modelos com e sem intercepto	203
A.6 Intervalo de confiança para $E(Y/X=x)$ e intervalo de previsão de Y.....	206
 ANEXO B - BOX PLOT	 209
 REFERÊNCIAS BIBLIOGRÁFICAS	 229

LISTA DE TABELAS

Tabela 1 - Correlação de Pearson calculada com 109.838 indivíduos da O/D-87, Metrô - SP.....	100
Tabela 2 - Nível de instrução - distribuição de frequências	103
Tabela 3 - Papel ocupado dentro da estrutura familiar- dist de freq	104
Tabela 4 - Estrutura familiar-distribuição de frequências	105
Tabela 5 - Distribuição de idade em anos completos	107
Tabela 6 - Índice de ciclo de vida (ICV) para os dados coletados na O/D-87 - Metrô - SP	108
Tabela 7 - Índice de ciclo de vida (ICV) segundo dados coletados pelo SEADE na Pesquisa de Condições de Vida de 1990.....	108
Tabela 8 - Idade média dos chefes, cônjuges e filhos das famílias nucleares completas da RMSP, segundo as etapas do ciclo da vida familiar.....	111
Tabela 9 - Distribuição de frequências do nº de viagens diárias individuais.	114
Tabela 10 - Número médio de viagens diárias individuais por nível de instrução	122
Tabela 11 - Estimação dos parâmetros dos modelos incluindo as variáveis domiciliares	146
Tabela 12 - Estimação dos parâmetros dos modelos excluindo as variáveis domiciliares	147
Tabela 13 - Descrição das variáveis	148
Tabela 14 - Alterações no ajuste geral dos modelos com e sem a inclusão das variáveis: número de automóveis da família e ocupação principal do indivíduo. Teste F-Snedecor e coeficiente de determinação (r^2).....	158

Tabela 15 - Teste F-Snedecor da contribuição estatística do grupo de variáveis domiciliares para modelos com e sem a inclusão das variáveis: número de automóveis da família e ocupação principal do indivíduo	160
Tabela 16 - Correlação de Pearson para o número de viagens diárias individuais	170

LISTA DE GRÁFICOS

Gráfico 1 - ICV médio por idade média do casal (IDMC) das famílias nucleares completas	110
Gráfico 2 - Número médio de viagens diárias individuais por renda pontual	119
Gráfico 3 - Número médio de viagens diárias individuais por faixas de renda.....	120
Gráfico 4 - Número médio de viagens diárias individuais por faixas de idade	121
Gráfico 5 - Número médio de viagens diárias individuais por idade pontual	121
Gráfico 6 - Frequências dos resíduos do modelo estimado com 232 indivíduos da RMSP	163
Gráfico 7 - Dispersão dos resíduos do modelo estimado com 232 indivíduos da RMSP	164

LISTA DE FIGURAS

Figura 1 - <i>Box Plot</i>	124
----------------------------------	-----

RESUMO

TANDEL, M. C. F. Freitas Geração de Viagens de passageiros urbanos: modelo individual com a inclusão da influência das características do domicílio. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo.

Um dos enfoques mais utilizados na modelagem de geração de viagens de passageiros urbanos é a desagregação em nível de domicílios. Esta tese oferece razões lógicas e fortes evidências estatísticas para substituir modelos baseados em domicílios, introduzindo modelos desagregados em nível de indivíduos incorporando a influência das características e da estrutura do domicílio. Estes modelos provêm um considerável avanço sobre as práticas tradicionais. Um estudo de caso foi conduzido na região metropolitana de São Paulo - SP. Utilizou-se informação em nível individual, através de algumas variáveis disponíveis na pesquisa origem/destino de 1987 da Cia. do Metropolitano de São Paulo. Foi possível verificar várias vantagens em relação aos modelos tradicionais, bem como responder várias críticas aos modelos atuais. Analisou-se também os erros obtidos com o modelo individual com e sem a influência das características e estrutura domiciliares.

Palavras-chave: Geração de viagens; produção de viagens; taxas de viagens individuais; modelos individuais; modelos desagregados; demanda por viagens; demanda por transportes.

ABSTRACT

TANDEL, M. C. F. Freitas Urban passenger trip generation: individual model incorporating the influence of household characteristics. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo.

One approach that is widely used to conduct the urban passenger trip generation modelling is the disaggregate household based model. This thesis offers logical reasons and strong statistical evidences to replace household based models, introducing individual disaggregate models, incorporating the influence of household characteristics and structure. These models represent a significant improvement over traditional practice. A case study was conducted using data from metropolitan region of São Paulo City, São Paulo State. Information at individual level was used, through some variables available in the São Paulo Metropolitan Company, 1987 origin/destination research. It was possible to verify several advantages in relation to traditional models as well as to answer several criticisms to the current models. It was also analysed errors obtained in the individual models with and without the influence of household characteristics and structure.

Keywords: trip generation; trip production; person trip rates; individual models; disaggregate models; travel demand; transportation demand.

1. INTRODUÇÃO

A importância do transporte no desenvolvimento do mundo é múltipla. Constitue o componente principal da estrutura econômica, social, cultural e política da nossa sociedade, é um fator vital na civilização contemporânea. Não existe como um fim em si mesmo, consiste num meio não em um fim. Possui funções e objetivos específicos, entre os quais Yu (1982) cita:

- tornar bens e serviços disponíveis quando deles se necessita.
- tornar efetivo o uso de recursos naturais.
- disponibilizar aos consumidores, várias mercadorias.
- possibilitar a descentralização de zonas industriais ou especializar uma dada região.
- promover a competição de mercado de bens e serviços, colaborando para aumentar o nível de qualidade e preços satisfatórios.
- promover mobilidade para o desempenho de atividades culturais, educacionais, de saúde, lazer, etc.
- promover a unidade política e social de uma nação.
- contribuir para o desenvolvimento da capacidade de defesa de uma nação.

A engenharia de transportes foi definida pelo Instituto de Engenheiros de Transportes (USA) em 1987 como sendo a aplicação de princípios científicos e tecnológicos ao planejamento, operação e gerenciamento de instalação para qualquer modo de transporte, com o fim de prover um seguro, rápido, confortável, conveniente e econômico movimento de pessoas e outros bens.

A prática da engenharia de transportes envolve uma diversidade de atividades desempenhadas por especialistas de diversas áreas: políticos, executivos,

engenheiros, planejadores, cientistas. Além disto é uma área de estudo interdisciplinar, com a qual contribuem: engenheiros, economistas, sociólogos, geógrafos, psicólogos, estatísticos, matemáticos, analistas de sistemas, entre outros (Khisty, 1990). O desejo de pessoas moverem-se e a necessidade de bens e serviços criam a demanda por transportes. A modelagem de demanda por transportes é uma das áreas de estudo da engenharia de transporte.

Especificamente tratando da modelagem da demanda de viagens de passageiros urbanos, observa-se uma progressiva evolução nos últimos 30 a 40 anos sob uma bem estabelecida metodologia, comumente chamada de aproximação clássica ou tradicional (Ortúzar & Willumsen, 1994), ou ainda chamada de modelo de demanda sequencial de 4 etapas.

A primeira etapa deste processo é a geração de viagens. Seu objetivo é estimar T_i ou T_j , isto é, o número de viagens de passageiros originados (T_i) ou finalizados (T_j) numa dada região (Oppenheim, 1995). Há dois grandes enfoques para a estimação desta etapa do processo da análise de demanda: agregado: em nível de zonas ou regiões, isto é, quando o número de viagens é expresso como uma função de características da zona; e desagregado: geralmente em nível de domicílios, isto é, quando o número de viagens é descrito como uma função de características do domicílio.

A estimação é desenvolvida utilizando-se várias técnicas de análise estatística de dados, com diferentes graus de complexidade. Desde a simples estimação pelas taxas de viagens observadas para as várias zonas ou domicílios até modelos que incluem análises de variância, discriminante, fatorial, de regressão linear, etc.; além de métodos de classificação cruzados, ou análises de categoria, usados para classificar os domicílios ou indivíduos em alguns

grupos homogêneos e distintos, para posteriormente atribuir-se as taxas de viagens que os caracterizem.

Todas estas técnicas trazem subsídios para se obter modelos matemáticos-estatísticos de demanda.

Dentro deste contexto deve-se ter em mente alguns fatores que influenciam a geração de viagens:

- O uso do solo: regiões comerciais, educacionais, de lazer produzem diferentes características de geração de viagens.
- O domicílio: o número de pessoas no domicílio, o tipo de estrutura familiar, o estágio no ciclo de vida, etc. determinam atividades a serem desempenhadas que geram deslocamentos.
- A disponibilidade de automóvel: a demanda por viagens é afetada pela disponibilidade de meios de transporte e pela adequação do sistema viário.
- A renda familiar ou individual.

Outros fatores podem ainda ser citados (Bruton, 1993): a estrutura etária da população em estudo, características sócio-econômicas gerais, o grau de urbanização da área de interesse, a qualidade dos serviços de transportes oferecidos, etc. Tais fatores afetam o padrão de atividades dos indivíduos.

Segundo Yu (1982) a mobilidade é o suporte principal do sistema de atividades da raça humana. Um grau adequado de mobilidade tende a ampliar as perspectivas do indivíduo. A necessidade de 'ligar' atividades que se

realizam em diferentes lugares e conduzir pessoas ou bens através destes '*links*', aumenta na medida em que uma sociedade torna-se mais complexa.

Uma cidade pode ser considerada como um local de combinação ou organização de atividades. O local das atividades afeta os seres humanos que por sua vez modificam seus planos e a organização de seus locais de atividades. Esta interação se manifesta no movimento de pessoas, bens e de informação.

Um modelo pode ser definido como uma representação simplificada de parte do mundo real ou do sistema de interesse. Normalmente ela se concentra em certos elementos, considerados importantes sob um particular ponto de vista.

A literatura recente vem apontando a importância de se analisar diferentes aspectos da análise de demanda sob a ótica do comportamento do usuário de transporte. Ressalta também a necessidade de modelos desagregados que possam explicar os demais aspectos da análise de demanda ainda não estudados suficientemente (Rengaraju & Satyakumar, 1994).

Neste contexto desenvolve-se este trabalho com o intuito de estudar a importância da influência das características do domicílio, sobre o indivíduo na etapa de geração de viagens no modelo sequencial de análise de demanda.

Cabe ressaltar que a presente tese não tratará de encontrar um melhor ajuste ou um modelo ótimo para a estimação da geração de viagens. Trata-se de mostrar a relevância de variáveis que caracterizam os domicílios, incluindo a caracterização das relações de parentesco do grupo de pessoas convivente, para a construção de modelos desagregados em nível individual. Convém lembrar que estes modelos têm sido negligenciados pela literatura, excetuando-

se somente Ortúzar & Willumsen (1994) que ressaltam a importância do modelo individual desenvolvido por Januz Supernak em 1979.

1.1 Objetivos

Vários são os objetivos que se pretende atingir. Poder-se-ia dizer existirem objetivos principais e secundários.

Como objetivos principais pretende-se mostrar:

- A necessidade de desagregar o modelo até o nível individual.
- Que com o modelo individual pode-se estimar adequadamente a mobilidade total de uma região com amostras razoavelmente pequenas, sendo que isto significa redução de custos.
- Diagnosticar a importância da influência das características do domicílio sobre o indivíduo, ou seja, mostrar que o indivíduo se comporta diferentemente, no que tange ao padrão de deslocamentos, dependendo do contexto domiciliar no qual está inserido. Faz parte deste contexto, o papel que ele desempenha, o tipo de relação familiar que mantém com os demais membros, ou seja, o tipo de estrutura familiar ao qual pertence e o estágio no ciclo da vida em que se encontra, o número de membros do domicílio, etc.
- A significância estatística da inserção de características domiciliares ao modelo individual de geração de viagens de passageiros urbanos.

Entre os secundários pretende-se:

- Apresentar os enfoques e métodos usuais utilizados na etapa de geração de viagens, com ênfase aos aspectos metodológicos.
- Analisar o caráter de modelagem destes métodos, descrevendo suas vantagens e limitações.
- Analisar a adequabilidade dos métodos para gerar dados quantitativos mediante o estudo da possibilidade de realizar-se previsões, verificar-se a qualidade do ajustamento, a magnitude das margens de erros das estimativas e o tamanho de amostra necessário.

Finalmente, pretende-se oferecer aos interessados uma contribuição no momento da escolha do enfoque: agregado, desagregado domiciliar ou individual; bem como alertar para o uso adequado das técnicas de análises estatísticas apontando suas limitações e potencialidades.

1.2 Justificativa para a escolha do tema

As estimativas feitas na etapa de geração de viagens, no modelo seqüencial de análise de demanda, influenciam todas as demais etapas, que dela dependem, derivando daí sua importância, pois os erros cometidos são propagados para o processo inteiro.

Vários são os fatores considerados importantes nesta etapa de análise: o número de automóveis disponíveis no domicílio, a renda, etc.

Normalmente relega-se a segundo plano o tipo de estrutura do domicílio, o estágio no ciclo da vida e o papel que o indivíduo desempenha dentro deste contexto, embora aceite-se facilmente que estes fatores são relevantes nas mudanças de comportamento pessoal em relação às viagens e às atividades.

Exemplificando: uma mulher cônjuge de uma família composta por um casal com filhos, deve ter atividades bem diferentes de uma mulher chefe de uma família onde não está presente o cônjuge e por conseguinte o padrão de deslocamentos deve ser diferente. Contudo, ainda não foi possível incluir estas diferenças na etapa de geração de viagens (Ortúzar & Willumsen, 1994).

Focalizar o indivíduo dentro deste contexto é inovador, sobretudo levando-se em conta a relevância das mudanças recentes ocorridas dentro do contexto familiar e por conseguinte na sociedade como um todo.

Acredita-se que com o presente trabalho possa ser ampliado o conhecimento de fatores que influenciam a etapa de geração de viagens de passageiros urbanos.

1.2.1 Complementação da justificativa do tema: breve estudo das transições importantes no padrão de atividades das famílias da Região Metropolitana de São Paulo - SP

Sendo a demanda por transportes relacionada ou mesmo proveniente da demanda por atividades, que por sua vez depende de características sócio-econômicas, convém apresentar algumas transições recentes no padrão de atividades da população da Região Metropolitana de São Paulo, derivada também de mudanças no tipo de estrutura familiar e de comportamento nos diversos estágios do ciclo da vida advindos de restrições nas condições sócio-econômicas dos últimos anos.

Em geral os dados estatísticos apresentados referem-se ao final da década de 80 ou meados da de 90, período que coincide com os dados a serem analisados posteriormente neste trabalho.

O ciclo de desenvolvimento das famílias, ou o ciclo vital das famílias, enquanto abordagem tem sido utilizado em muitos estudos na Sociologia, Antropologia, Psicologia Social e Demografia. O conceito de ciclo de vida corresponde a uma visão dinâmica da evolução da família através de uma seqüência de fases ou estágios que correspondem a mudanças significativas em sua composição e dinâmica internas (Montali, 1990).

Tem-se encontrado uma relação clara entre os arranjos familiares possíveis de articular em cada momento do ciclo vital familiar, para enfrentar as condições de existência e as possibilidades e dificuldades que estes delineiam para a obtenção de rendimentos suficientes.

Concretizam-se distintos arranjos familiares de inserção no mercado de trabalho, a partir de situações individuais heterogêneas, como pai, filho ou cônjuge, num esforço coletivo ao nível da família para encontrar formas de enfrentamento das condições de existência e a possibilidade de provocar alterações no sentido de superar a condição de pobreza.

A condição de desempregado ou de subempregado dos provedores das famílias, aliada à aceleração da perda do poder de compra dos salários em consequência da inflação, com certeza pressionou outros componentes da família em idade ativa para a busca de trabalho, atingindo muitas mulheres/cônjuges e filhos que, em outras conjunturas, possivelmente teriam permanecido inativos.

Não existe permutabilidade dos diferentes membros da família no que concerne ao acesso ao emprego, na medida em que a posição no interior da família impõe, de diferentes maneiras, principalmente para o homem e a mulher do casal, pressões e limitações no acesso aos empregos, ou seja não há uma entrada indiferenciada no mercado de trabalho (Montali, 1990). O fato é que as famílias reagem à crise e ao desemprego levando em conta o conjunto de relações existentes em seu seio.

Mudanças significativas ocorreram e ocorrerão nas atividades econômicas e nos modos de vida da população, influenciando a organização das famílias, a condição sexual dos seus chefes e a diferenciação das oportunidades de trabalho por sexo.

Verifica-se que o assalariamento passou a predominar entre os chefes de família, ao lado do rejuvenescimento da força de trabalho e a redução no tamanho das famílias, o que indica uma diferenciação nos padrões de

reprodução e ciclos da vida familiar, que sugere uma estratégia de articulação coletiva para enfrentar as condições de vida envolvendo poucos indivíduos e apoiando-se fundamentalmente na atividade econômica das mulheres/cônjuges.

De fato, os processos de urbanização, de proletarização e de mercantilização das relações de produção podem determinar o tamanho e composição das famílias, destruiu-se a lógica da articulação entre tamanho da família e economia de subexistência, desestabilizando os níveis de fecundidade que passaram a cair rapidamente desde 1960 (de 6,28 em 1960 para 4,35 em 1980 - estimativas dos censos dos anos respectivos encontradas no Perfil Estatístico de Crianças e Mães no Brasil, IBGE, 1988).

Emerge assim, um novo modo de organização da vida familiar, sustentado em bases econômicas e culturais igualmente renovadas. O tamanho da família constitui pois um dado importante para a compreensão das possibilidades de organização, frente às condições de mercado de trabalho e do momento do ciclo da vida familiar (Guimarães, 1991).

São inúmeras as pesquisas que evidenciam que filhos jovens, na grande maioria menores de 18 anos estão inseridos no mercado de trabalho, de forma precária, bem como grande parcela das mulheres/mães. Quando as famílias são chefiadas por mulheres, todo o grupo familiar se mobiliza, trabalhando velhos, crianças e adolescentes (Bruschini, 1990).

A participação de indivíduos do sexo feminino na produção social não se define apenas pelas condições de mercado, pela estrutura do emprego, ou mais genericamente pelo nível de desenvolvimento da sociedade, mas também por sua posição na família e pela classe social à qual pertence o grupo doméstico.

Filha, esposa ou mãe, a cada uma destas etapas do ciclo vital, corresponderão determinadas necessidades e possibilidades de trabalho que, dadas as oportunidades oferecidas pelo mercado, definirão a situação da mulher em relação às atividades produtivas.

Analizando dados dos últimos censos e PNAD's: Pesquisa Nacional por Amostragem de Domicílios, IBGE, constata-se que a presença da mulher na força de trabalho brasileira vem-se ampliando consideravelmente nos últimos anos, a tendência ao crescimento é inegável (chegou quase a triplicar em menos de 15 anos), apesar disto é distante ainda do nível de regularidade com que se apresenta o trabalho masculino. Isto porque o trabalho feminino tem como característica predominante ser descontínuo, freqüentemente em tempo parcial. É um trabalho marcado pela diversidade e intermitência de entradas e saídas no mercado, marca registrada do frágil equilíbrio entre as atividades produtivas e funções reprodutivas, mantidas a todo custo pela mulher para a sobrevivência e o bem estar de todo o grupo familiar.

Em qualquer análise sobre o trabalho da mulher, um aspecto crucial é o da posição na divisão social e sexual do trabalho, prioritariamente definido a partir de suas funções biológicas, o que a condiciona, de um lado, à execução de afazeres domésticos, e de outro, a ocupar posições subalternas na hierarquia produtiva.

Esta distribuição contudo, altera-se substancialmente quando há credenciais vantajosas de qualificação e preparo profissional a serem apresentadas no mercado de trabalho. É o caso das mulheres casadas, mesmo sendo mães, cujas atividades econômicas são mais elevadas quando tem instrução superior e os rendimentos obtidos passam a compensar a saída do lar (Bruschini, 1986 e 1990).

Os efeitos diferenciados da crise econômica sobre ambos os sexos, ainda não estão suficientemente esclarecidos. Em São Paulo - SP, onde o quadro se apresenta relativamente mais favorável ao trabalhador como um todo, a discriminação salarial por sexo não deixa de ocorrer (Bruschini, 1986; Haga, 1986), os dados revelam que a mulher continua sendo discriminada, que ganha salários inferiores aos do homem, tem menor acesso às garantias trabalhistas, raramente ocupa posto de chefia e é segregada em “guetos” ocupacionais, tais como a indústria têxtil.

Segundo Bruschini (1986), o caráter secundário e complementar atribuído à atividade produtiva feminina é que permite e legitima a condição de discriminação. A autora ressalta contudo a crença no fato de que mudanças significativas, ao nível da família e social, estão ocorrendo e ocorrerão devido ao impacto da presença feminina em atividades produtivas.

É indiscutível a influência da idade e presença ou não de filhos na administração do tempo da mulher entre os afazeres domésticos e as atividades de trabalho fora do lar, contudo segundo Psacharopoulos & Tzannatos (1989) existe uma tendência de decréscimo na participação da mulher em atividades produtivas fora do lar, quando a mulher se encontra no estágio do ciclo da vida caracterizado pelo cuidado de filhos em idade pré-escolar, mas não se deve generalizar esta conclusão. Há vários exemplos em seu artigo de sociedades modernas, mas especificamente entre países industrializados, onde isto não ocorre.

Tendências importantes ocorreram durante a década de 80, em todo o Brasil, mas sobretudo na região Metropolitana de São Paulo, indicando uma crescente participação da mulher-cônjuge no mercado de trabalho mesmo existindo especificidades em sua participação nos diferentes momentos do ciclo da vida

familiar. Entende-se que esta seja a resposta a uma década marcada por conjunturas de instabilidade e recessão. Não se pode negar a tendência secular, ocorrida também nos países desenvolvidos, de crescente participação da mulher no mercado de trabalho, existindo explicações múltiplas para este fenômeno (Montali, 1991).

Segundo FLACSO (1993):

- A mão de obra feminina em todo o Brasil cresceu 180% nos últimos 20 anos enquanto que a masculina cresceu 71%.

- 30% da população economicamente ativa, do Brasil, são mulheres, sendo que em zonas urbanas este número passa a 48,4%. Deve levar-se em conta ainda o fato de haver um grande sub-registro de participação feminina no trabalho, são trabalhos informais ou da zona rural em geral.

Convém traçar o perfil estatístico da RMSP no que diz respeito a alguns aspectos de interesse. O Anuário Estatístico - 1992, recentemente publicado pelo SEADE, estima que 51,25% da população da RMSP é feminina (esta é uma característica demográfica nacional).

É interessante observar alguns dados, da PNAD, IBGE de 1990, a respeito da inserção da mulher no mercado de trabalho na RMSP: Região Metropolitana de São Paulo:

- 41,17% das mulheres de 10 anos ou mais estão inseridas na população economicamente ativa: PEA;

- 37,46% da PEA é formada por mulheres;

66,85% da PEA tem 8 anos de estudo ou menos;

43,84% da PEA feminina ganha no máximo 3 salários mínimos;

23,35% da PEA masculina ganha no máximo 3 salários mínimos;

47,55% das mulheres ocupadas ganham no máximo 3 salários mínimos;

25,57% dos homens ocupados ganham no máximo 3 salários mínimos;

31,15% entre as mulheres empregadas não possuíam carteira de trabalho assinada pelo empregador;

19,30% entre os homens empregados não possuíam carteira de trabalho assinada pelo empregador;

17,96% da PEA feminina ocupam a posição de chefe da família;

67,49% da PEA masculina ocupam a posição de chefe da família.

Desde a expansão dos direitos consagrados no Estatuto da Criança e do Adolescente, ECA, sancionado em outubro de 1990, inúmeras pesquisas têm-se realizado com o intuito de compreender melhor as condições efetivas em que vivem as famílias, instância em que se concretiza o exercício de direitos da população com menos de 18 anos.

O nascimento dos filhos sempre foi mostrado como um fator decisivo para a retirada da mulher do mercado de trabalho, mas numa conjuntura onde não

existe a possibilidade de contar com o trabalho de filhos mais velhos, ficando a sobrevivência familiar dependente principalmente de seus chefes, aumentam-se os índices de pauperização. Nestas circunstâncias as exigências da sobrevivência terminam por mobilizar intensamente mulheres e filhos para o mercado de trabalho. Na ausência de direitos e políticas sociais que garantam a sobrevivência nas situações de desemprego, mas também de doença, da invalidez, da velhice, todos, homens e mulheres, crianças, adultos e velhos são virtualmente transformados em população ativa.

Segundo Telles (1992) não há regras, parâmetros ou limites de qualquer tipo que permitam preservar, ao menos potencialmente, aqueles que poderiam ou deveriam se manter fora do mercado de trabalho. A situação dos filhos é neste sentido paradigmática: sua entrada precoce no mercado de trabalho representa um ônus evidente para sua escolarização e qualificação profissional. Telles ressalta ainda que as pesquisas têm indicado as mesmas práticas de discriminação do trabalho infantil que atinge as mulheres, as crianças e adolescentes são absorvidos como mão-de-obra desqualificada e mal remunerada.

É importante caracterizar o tipo de família da RMSP, as fontes de dados são amplas, entre elas destaca-se a Pesquisa de Condições de Vida, da Fundação SEADE - Sistema Estadual de Análises de Dados, realizada em 1990. Pode-se verificar que:

70,5% das famílias da RMSP são nucleares completas, formadas por casais e filhos somente;

15,5% são nucleares ampliadas onde se inclui parentes ou outros agregados;

A família quebrada, em que está faltando um dos cônjuges, são 12,9% do total das famílias com crianças ou adolescentes;

67,6% das famílias possuem crianças ou adolescentes.

A PED - Pesquisa de Emprego e Desemprego, publicado pelo SEADE trimestralmente, mostra a taxa de desemprego pela posição que o morador do domicílio ocupa na família, os dados abaixo referem-se aos dados coletados no 3º trimestre de 1993. Calculou-se a média dos 3 meses que compõem o trimestre, obtendo-se como porcentagem da PEA: população economicamente ativa, que estavam desempregados (taxa de desemprego) os seguintes valores: 8,5% dos chefes de família; 13,20% dos cônjuges; 23,83% dos filhos.

Outros dados de interesse fornecem embasamento para a pesquisa a que se propõe este projeto, observou-se que:

- mais da metade das famílias da RMSP: Região Metropolitana de São Paulo, é carente em educação, o que significa que nelas o chefe ou o segundo componente com 18 anos ou mais e não estudante conseguiu concluir, no máximo, o 1º grau;

- mais de um terço de indivíduos da RMSP tem menos de 18 anos de idade, dos quais 43,8% vivem em famílias com precárias condições sócio-econômicas e são carentes em três ou mais dimensões da pobreza: habitação, educação, saúde, renda e inserção no mercado de trabalho;

- menores de 18 anos constituem quase a metade das pessoas que compõem as famílias mais pobres, enquanto que nas mais ricas eles são 30% dos integrantes;

- entre os menores de 18 anos o contingente feminino é ligeiramente maior que o masculino: 51% e 49% respectivamente. Cabe ressaltar que este comportamento não é específico da RMSP e, de modo geral, independe da faixa etária considerada, trata-se de uma característica demográfica brasileira;

- 91,3% dos menores de 18 anos ocupam a posição de filho nas famílias da RMSP;

- aproximadamente um terço das crianças e adolescentes entre 10 e 17 anos da RMSP estavam inseridos no mercado de trabalho, contribuindo com cerca de 28,4% do total da renda familiar, caracterizando o trabalho desta faixa da população como indispensável para a sobrevivência do grupo familiar;

- 13,3% das famílias da RMSP são chefiadas por mulheres, havendo uma tendência de quanto maior o nível de precariedade das famílias, maior o percentual daquelas chefiadas por mulheres;

- trabalha fora cerca de um terço das mães com crianças em idade pré-escolar: 0 a 6 anos. Analisando esta condição segundo a divisão da PCV, 1990 (SEADE, 1992a e b), em grupos homogêneos de famílias, segundo suas condições de vida ou carências tem-se:

- no grupo A (condições de vida mais favoráveis, cuja principal característica é não incluir famílias carentes de instrução) 20,5% das famílias da RMSP, 51,7% das mães trabalham fora;

- no grupo B (condições intermediárias na escala sócio-econômica, tendo por característica básica a não carência de moradia) 37,2% das famílias da RMSP, 22,3% das mães trabalham fora;
 - no grupo C (todas as famílias deste grupo são carentes em moradia) 19,7% das famílias da RMSP, 41,4 % das mães trabalham fora;
 - no grupo D (são a parcela mais carente da população em nível de instrução, mercado de trabalho e renda) 22,6% das famílias da RMSP, 32,2% das mães trabalham fora.
- considerando-se mães de crianças de 0 a 6 anos que freqüentam creche ou pré-escola, trabalha fora aproximadamente a metade destas mães. Analisando também segundo os grupos homogêneos de famílias, descritos no item anterior, observa-se que o percentual de mães que trabalham fora fica assim distribuído:

No grupo A: 59,4%;

No grupo B: 26,5%;

No grupo C: 65,1%;

No grupo D: 54,5%.

- 27,89% da PEA ocupam a posição de filhos nas famílias da RMSP segundo a PNAD, IBGE, 1990.

As mulheres, as crianças e os adolescentes são mão-de-obra auxiliar fundamental para aumento do rendimento do domicílio, descaracterizando a criança como passageiro constante nas situações em que a família busca

atender às necessidades de educação e lazer. Elas têm portanto um padrão de atividades e deslocamentos bem diferente do que o relatado na literatura internacional, bem como afetam de maneira diferente o padrão de atividades dos outros moradores do domicílio.

1.3 Métodos de pesquisa

Os procedimentos a serem tomados para atingir-se os objetivos propostos são:

- Fazer uma revisão bibliográfica sobre os métodos e enfoques usuais na modelagem de geração de viagens de passageiros urbanos.
- Comparar os métodos encontrados de acordo com seus fundamentos teóricos.
- Buscar na sociologia e psicologia o embasamento teórico necessário para estabelecer-se criteriosamente uma relação de causa e efeito entre as características do domicílio e o indivíduo, ou seja, caracterizar a influência do domicílio sobre o indivíduo através do estudo de variáveis tais como, papel desempenhado dentro da estrutura familiar, tipo de estrutura familiar, estágio no ciclo da vida da família, número de membros do domicílio, etc.
- Realizar um estudo de caso com a finalidade de obter dados quantitativos criteriosos que permitam a detecção estatística da influência do domicílio sobre o indivíduo, no que tange a mudanças no seu padrão de deslocamentos. Tal verificação estatística deve estar baseada em amostras aleatórias de dados individuais que permitam submeter as hipóteses a testes

estatísticos de significância da contribuição do grupo de variáveis domiciliares ao modelo de geração de viagens.

- Para mostrar a importância da presença do grupo de variáveis domiciliares, além da verificação de sua significância estatística, far-se-á um estudo dos erros obtidos ao se estimar o total de viagens de uma região utilizando-se modelos individuais com e sem estas variáveis.

1.4 Apresentação dos capítulos

O trabalho está dividido em 7 capítulos, e 2 anexos, o Anexo A e o B.

O Capítulo 1 faz uma apresentação geral da tese, da justificativa para a escolha do tema, objetivos e métodos.

O Capítulo 2 introduz de maneira geral a etapa de geração de viagens no contexto do modelo de análise de demanda por transporte.

O Capítulo 3 apresenta uma revisão bibliográfica sobre os enfoques e métodos usuais para modelar a geração de viagens, incluindo comparações, abrangências e limitações de cada enfoque e método.

O Capítulo 4 apresenta a importância da influência das variáveis domiciliares sobre o indivíduo sob a ótica da sociologia e da psicologia social.

O Capítulo 5 apresenta um estudo de caso, com resultados estatísticos extraídos de dados da Região Metropolitana de São Paulo fornecidos pela Cia.

do Metropolitano. Trata-se dos dados na íntegra da pesquisa Origem/Destino realizada em 1987.

O Capítulo 6 trata da estimação dos parâmetros de alguns modelos individuais e mostra a significância estatística da contribuição da inclusão das variáveis domiciliares, consistindo na verificação e na constatação da viabilidade da tese proposta.

O Capítulo 7 comenta os aspectos que compõe a contribuição efetiva desta tese.

O Capítulo 8 descreve as principais conclusões, recomendações e sugestões.

O anexo A resume os principais conceitos teóricos necessários para uma compreensão adequada do uso e abrangência da técnica de análise de regressão linear usando o método de mínimos quadrados.

O anexo B contém os *Box Plot's* usados na análise descritiva dos dados do capítulo 5.

2. A ANÁLISE DA DEMANDA POR TRANSPORTES

A “demanda por transportes é o desejo de uma entidade, uma pessoa ou um grupo de pessoas físicas ou jurídicas, de locomover alguma coisa, a si próprio, outras pessoas ou cargas, de um lugar para o outro”.

“O conhecimento da demanda por transportes de uma região ou de uma cidade é indispensável ao planejamento de transportes, na medida em que ele mostra os deslocamentos potenciais de pessoas ou mercadorias num espaço físico, ajudando a estabelecer prioridades no atendimento e a dimensão da oferta de transportes, além de indicar a quantidade e a localização, atual ou futura, da população beneficiada por um determinado projeto de transporte” (Kawamoto, 1993, p.35 e 36).

A busca de um gerenciamento de transportes mais eficiente necessita da busca de orientação em análises do comportamento da demanda, facilitando a alocação de recursos e a tomada de decisões sobre políticas de transportes, gerando dados quantitativos para um planejamento de transportes mais eficaz.

O principal objetivo da análise de demanda é compreender os determinantes da demanda e a maneira como eles interagem e afetam a evolução do volume de tráfego, analisar os fatores que possivelmente influem no volume de tráfego, estudar a forma como esta influência ocorre, projetar fatores para uma data futura e finalmente, estimar o provável volume de tráfego futuro, mesmo para outras regiões com características diferentes daquelas que deram origem ao modelo.

Os resultados destas análises são geralmente modelos matemáticos ou estatísticos que relacionam medidas de demanda por transportes (geralmente volume de tráfego) com medidas do nível de atividades sócio-econômicas e/ou características do usuário de transportes.

Previsões podem ser feitas quando o pesquisador estiver convencido de que compreendeu suficientemente os determinantes da demanda e o mecanismo de interação entre eles, bem como o método escolhido para desenvolver o modelo de previsão.

A análise de demanda é distinta da previsão de tráfego. A previsão de tráfego por si só, consiste, principalmente, em analisar a variação ocorrida no volume de tráfego de uma dada região, ao longo de tempo e projetá-la para uma data futura levando-se em conta a tendência do passado.

Tradicionalmente dois tipos de questões são respondidas pela análise de demanda por transportes (Kanafani, 1983):

- Avaliação de estratégias relacionadas a oferta de transportes, por exemplo, impactos devido a mudanças sócio-econômicas.
- Avaliação de alternativas para o uso do solo, por exemplo, previsão de tráfego.

As limitações básicas referem-se ao fato de a demanda estar ligada à expressão do comportamento humano e portanto não pode ser tratada como um simples agregado de decisões de indivíduos racionais que maximizam suas utilidades, enfoque dado pela maior parte dos modelos de demanda por transportes urbanos (Kanafani, 1983; Bodmer, 1984).

2.1 Os modelos de demanda por transportes

Conhecer a relação funcional existente entre os fatores que afetam a demanda é tarefa difícil mas de grande importância, sendo preciso saber quais as características do fator a serem medidas e como medi-las.

No atual estágio de desenvolvimento da análise da demanda, apenas tem-se uma leve idéia sobre a maneira como as variáveis interferem na quantidade de viagens entre duas zonas, os métodos de quantificação de demanda necessariamente passam pela especificação *a priori* de uma função matemática que via de regra é denominada modelo de demanda.

O uso mais comum dos modelos de demanda está ligado à estimativa do nível de uso de componentes do sistema de transporte num determinado cenário. Dado que é praticamente impossível representar o cenário em toda a sua extensão através de variáveis, espera-se que o modelo contenha pelo menos as principais variáveis que influem no nível de uso de componentes de transporte.

A maioria dos modelos de transporte urbano é baseada em hipóteses sobre o comportamento do usuário urbano e é validada pela observação do comportamento relacionado às viagens ou deslocamentos, contudo não é possível realizar experimentos controlados para avaliar como os usuários reagem diante de situações hipotéticas. Logo os modelos criados são úteis somente em situações semelhantes àquela que originou os resultados.

A demanda por transportes é consequência de outras demandas tais como a necessidade de trabalhar, estudar, fazer compras, desenvolver atividades de

lazer, etc. Diz-se então que a demanda por transportes deriva da demanda por outras atividades.

O postulado básico dos modelos de demanda por transportes urbanos é que esta demanda é relacionada ou mesmo proveniente da demanda por atividades urbanas, que por sua vez depende de características sócio-econômicas.

A oferta de atividades está ligada à ocupação do solo e às características do sistema de transportes.

Todo esse contexto gera um processo de escolha por parte do usuário. Esse processo é o objeto de estudo da análise de demanda por transportes e é também onde se tem desenvolvido a maior parte das pesquisas.

As hipóteses importantes na análise de demanda estão relacionadas a esse processo de escolha (Kanafani, 1983) e abrange os seguintes aspectos:

- Quantas viagens deve-se fazer num dado período de tempo? Este aspecto é intitulado “geração de viagens”.
- Qual o destino de cada viagem? Este aspecto é estudado sob o título “distribuição de viagens”.
- Qual o modo a ser usado? Aspecto estudado sob o título “divisão modal”.
- Qual rota utilizar? Aspecto estudado sob o título “atribuição de viagens às rotas”, ou alocação.

- Qual o período do dia em que se escolhe realizar o deslocamento? Pela manhã, a tarde, no horário de pico de tráfego?

Entender esse processo de escolha e modelar satisfatoriamente é essencial para a condução da análise de demanda por transporte urbano.

As duas maiores questões que envolvem este processo de escolha são:

- As escolhas são independentes entre si ?
- Se são dependentes, são simultâneas ou seqüenciais?

A literatura tem aceito como uma hipótese muito irrealista que este processo de escolha tenha suas etapas independentes, é difícil mesmo imaginar que a escolha da rota a seguir independa do destino ou do modo de viagem (Kanafani, 1983).

Tem-se considerado, portanto, que as escolhas são dependentes. Se estas forem consideradas simultâneas tem-se os modelos de demanda chamados “modelos diretos”, se seqüenciais tem-se os modelos chamados de “modelos seqüenciais”.

- **Os modelos diretos de demanda**

Os modelos diretos são aqueles com os quais procura-se representar a demanda por um dado modo e motivo de viagem, entre um determinado par de origem e destino.

São modelos do tipo:

$$T_{ijmr} = \alpha_{ijmr} \prod_k X_{ijmr}^k \beta_{ijmr}^k, \forall i, j, m, r$$

onde:

T: total de viagens

i: origem

j: destino

m: modo

r: rota

k: número real

α e β são parâmetros a serem estimados

X: são variáveis que medem vários atributos das zonas de demanda, destino, modo e das rotas.

Porém à medida que aumenta o número de zonas de origem e destino, o número de motivos e o número de modos de viagem, o uso de modelos diretos torna-se cada vez mais trabalhoso, basta verificar que para 206 zonas de tráfego tem-se um número de pares de origem-destino igual a 42.436 (número de elementos que existe numa matriz quadrada de ordem 206). O que não é irrealista pois esta foi a subdivisão da região Metropolitana de São Paulo feita em 1966 pela Companhia do Metropolitano de São Paulo para o planejamento da primeira linha do trem metropolitano.

Na verdade admite-se na literatura que o uso de modelos diretos torna-se impossível, pois para cada par de origem-destino as variáveis das demais zonas devem estar presentes, uma vez que incrementos sistemáticos no tempo ou custo de viagem em outros pares levam os indivíduos a mudarem de emprego

ou de residência afetando a demanda em outras zonas (Kawamoto, 1993). Tantos são os fatores numa complexa rede urbana de transporte com vários modos que o número de combinações possíveis de alternativas e seus atributos podem ser tão numerosos que excedam o limite realístico da discriminação humana, o que torna o processo de decisão humano extremamente difícil, senão impossível (Kanafani, 1983; Oppenheim, 1995).

- **Modelos seqüenciais de demanda**

Neste contexto os modelos seqüenciais visam simplificar as dificuldades apresentadas nos modelos diretos. Quando se considera o processo de escolha de forma seqüencial divide-se o problema em partes menores, de acordo com uma seqüência de decisão que se acredita que um indivíduo toma antes de viajar de i para j , por motivo n usando o modo de viagem m . Supõe-se que em primeiro lugar um indivíduo decide exercer uma atividade, depois decide onde exercer esta atividade, a seguir escolhe o modo de viagem e finalmente a rota (Kawamoto, 1993).

Tem-se então um processo de escolha em quatro etapas:

- geração de viagens,
- distribuição de viagens aos destinos,
- divisão modal e
- atribuição de viagens às rotas.

Como toda simplificação, esta também tem suas limitações, algumas das quais são brevemente discutidas em Oppenheim (1995).

No início da década de 1970 políticos americanos e europeus preocuparam-se com os grandes congestionamentos nos sistemas viários das grandes cidades, bem como com o consumo dos derivados de petróleo. Visando adotar medidas para solucionar estes problemas impulsionaram pesquisas que pudessem avaliar a reação dos usuários de transportes em relação às mudanças pretendidas, fossem de caráter técnico ou político. Desenvolveram-se então modelos comportamentais para a análise da demanda, ou seja, modelos baseados no comportamento do homem. Os modelos até então existentes eram baseados nas características das zonas de tráfego (modelos agregados) e não podiam avaliar os impactos das novas políticas de transportes sobre o comportamento dos usuários.

Isto causou nos últimos 20 anos, uma concentração da maior parte das pesquisas realizadas na área de demanda, na etapa de divisão modal (Kawamoto, 1993). Assistiu-se, também nos últimos 20 anos, uma estagnação, até mesmo uma ausência, de estudos na área de geração de viagens. Pode-se dizer que, num certo sentido, este fato ocorreu em virtude de as atenções estarem voltadas para as novas descobertas acerca da divisão modal.

Mas também nos últimos 20 anos a sociedade assistiu uma redefinição na estrutura familiar através da reformulação dos valores e papéis dos indivíduos de ambos os sexos, que trouxeram para a sociedade contemporânea um novo padrão de comportamento, um novo padrão de atividades, portanto um novo padrão de deslocamentos e por conseguinte mudanças na demanda por transportes.

3. GERAÇÃO DE VIAGENS

Segundo Khisty (1990) a geração de viagens é o processo pelo qual uma quantidade de atividades urbanas é convertida em número de viagens, ou seja, quantifica-se a relação entre atividades urbanas e número de deslocamentos. Pode-se exemplificar observando que o número de viagens realizadas para um shopping center é bem diferente que o número de viagens realizadas para um complexo industrial que ocupe a mesma área em espaço físico.

Os resultados finais de um estudo típico de geração de viagens, no contexto do modelo tradicional de análise de demanda, seqüencial de quatro etapas, envolve um estudo do número de viagens produzidas e o número de viagens atraídas por uma região geográfica ou zona de tráfego.

Os termos produção e atração de viagens não são definidos segundo as direções ou destinos das viagens, mas em termos do uso do solo associado à região onde as viagens são geradas (Papacostas, 1987). A produção de viagens é definida baseando-se em viagens que envolvem zonas residenciais e a atração de viagens é definida baseando-se em viagens que envolvem zonas não residenciais. No esquema abaixo a zona 1 produziu 2 viagens e a zona 2 atraiu 2 viagens (independentemente da origem ou destino de cada viagem).

Desta forma vê-se que há uma equivalência entre viagens residenciais produzidas e viagens não residenciais atraídas. Esta divisão surge em virtude de os modelos de distribuição de viagens mais usuais (2ª etapa do modelo seqüencial) incorporarem esta abordagem, bem como, ser mais viável estimar as viagens produzidas como função das características sócio-econômicas das regiões ou dos indivíduos em questão, que estimar as viagens atraídas como função da intensidade e oportunidades encontradas nas zonas de atração de viagens (Papacostas, 1987).

As viagens podem ainda ser classificadas em viagens baseadas no domicílio, (cuja origem ou o destino são o domicílio); e viagens não baseadas no domicílio, que são aquelas nas quais, nem o destino nem a origem são o domicílio. Estima-se que 80% das viagens de passageiros urbanos são baseadas no domicílio (Yu, 1982).

As viagens não baseadas no domicílio geralmente são estimadas através dos deslocamentos de veículos (Bruton, 1993).

A geração de viagens engloba o número total de viagens, baseadas ou não nos domicílios.

Os estudos da etapa de geração de viagens são parte vital do processo de planejamento de transporte. É essencial que sejam entendidos os determinantes atuais da produção de viagens antes que se procure entender a natureza futura da demanda por viagens.

Os modelos de geração de viagens tratam de quatro assuntos básicos: produção de viagens, atração de viagens, de pessoas ou bens.

Este trabalho aborda a modelagem de geração de viagens de passageiros urbanos, mas especificamente a produção de viagens de passageiros urbanos.

A maior parte dos estudos acerca da produção de viagens de passageiros leva em consideração fatores tais como: renda, disponibilidade de automóveis, valor do solo, densidade demográfica da região, tamanho da família, renda familiar, a estrutura etária da população, o grau de urbanização da região, etc. (Bruton, 1993; Ortúzar & Willumsen, 1994).

Os resultados obtidos na literatura até o presente momento, com a inclusão destes fatores, são descritos e discutidos no decorrer deste capítulo.

3.1 Enfoques atuais para a geração de viagens

Desde os anos 50 várias técnicas têm sido propostas para modelar a geração de viagens e estão baseadas em 2 procedimentos gerais:

- identificação de variáveis que possuem efeitos significantes sobre a geração de viagens;

- desenvolvimento de um modelo matemático-estatístico onde os efeitos destas variáveis possam ser medidos conjuntamente.

Um modelo geral pode ser descrito por (Kanafani, 1983).

$Y_p = X(D_p)$ onde:

Y_p = número de viagens realizadas com o propósito p.

D_p = vetor de variáveis ligadas à demanda por viagens com o propósito p.

Richardson (1990) considera que existiram 4 grandes eras ao discutir os enfoques utilizados no planejamento e modelagem de transportes. A primeira, pré 1969, é designada como a era dos modelos agregados; a segunda foi a era dos modelos desagregados (até meados de 1970), que foi seguida pela era dos modelos comportamentais. A era dos modelos de atividades, surgiu na década de 80 e continua até hoje.

Os dois primeiros enfoques foram muito absorvidos pelos modelos de geração de viagens, coexistindo ainda hoje em centros de relevância na pesquisa de transportes.

Os modelos de geração de viagens podem ser classificados como modelos originados por 2 enfoques gerais: o enfoque agregado e o desagregado.

Os termos “agregado” e “desagregado” são relativos. Considera-se desagregado o modelo que mantém um maior grau de informação em nível

individual. Por exemplo modelos que incluem renda média familiar são mais desagregados que os que incluem renda total ou média de uma região.

3.1.1 Modelos agregados

Os modelos agregados partem do pressuposto que variáveis significantes para explicar a geração de viagens estão ligadas às características sócio-econômicas e demográficas da região geográfica ou zona de tráfego em questão. São também chamados de modelos baseados em zonas. São modelos largamente utilizados e muitos exemplos são ainda hoje encontrados na literatura, Saka, 1993; Wilson e Megeed, 1990; Herendeen *et al.* 1993; Metrô-SP, 1968, 1975, 1991; etc. Procuram relacionar o uso do solo ao sistema de transportes através de equações matemáticas e estatísticas com o objetivo principal de explicar o número de viagens geradas na região segundo variáveis que incluem:

- total de empregos da região;
- renda média da região por unidade de área ou per capita ou familiar, etc.;
- total de domicílios;
- número de habitantes total ou por km².

A literatura tem mostrado que a maior proporção de variação do comportamento do usuário ocorre dentro da zona. Esta variação não é explicada pelo modelo agregado (Douglas, 1973; Ortúzar & Willumsen, 1994), que pode explicar as variações existentes somente entre zonas.

As insatisfações advindas do uso de modelos agregados são claras e extensas na literatura (Douglas, 1973; Daly & Ortúzar, 1990; Vuren, Gunn & Daly,

1995; Dey & Fricker, 1994; Saka, 1993; Wilmot, 1995; Kumarage & Wirasinghe, 1995; Kawano *et al.*, 1990). As principais baseiam-se no fato de os modelos baseados em zonas serem condicionados pelo tamanho e natureza do sistema existente na zona da qual o modelo foi derivado, portanto há dificuldades na transferência do modelo para outras regiões. Os modelos existentes só oferecem boas estimativas quando aplicados na região onde se desenvolveu o estudo ou em regiões muito semelhantes. Caso contrário muitos ajustes são necessários incluindo nova coleta de dados e recalibração do modelo.

Tal é a magnitude desta questão que métodos de transferência têm sido objetos de estudos. Por exemplo, Soumya & Fricker (1994) apresentam um método bayesiano para a transferência de modelos a nível nacional para modelos a nível local onde a informação nacional é considerada como uma informação *a priori*.

Estas dificuldades colaboram para a inadequação destes modelos para fazer previsões.

Modelos agregados não levam em conta o caráter comportamental do processo de geração de viagens, no máximo captam o comportamento de um segmento da população: os habitantes de tal região, os possuidores de automóveis, etc. Para solucionar estes problemas alguns autores sugerem reduzir o tamanho da região, zonas pequenas, contudo, implicam num grande número delas e isto traz duas conseqüências:

- Custos mais altos devido à maior complexidade de coleta de dados e calibração.

- Erros amostrais maiores que comprometem a precisão dos estimadores.

Segue abaixo alguns exemplos, bastante atuais, da aplicação deste enfoque.

- **Saka, 1993**, obteve um modelo de demanda por transporte para a região de Fredericksburg, Va. onde a geração de viagens é relacionada ao uso do solo e com variáveis sócio-econômicas da região. A primeira calibração do modelo ocorreu no ano de 1985, ele foi recalibrado em 1990 e efetuou-se uma previsão para 2010 mediante alterações no modelo de 1990. A análise de regressão linear foi utilizada, escolhendo-se as variáveis pelo método *stepwise*. As variáveis são:

T: número de viagens geradas por domicílio, variável dependente,

H: número total de domicílios da região,

E: número total de empregos da região,

O modelo passou por algumas etapas, partindo de um modelo básico descrito por:

$$T = \exp(-1,814 + 0,17 \ln H + 0,24 \ln E)$$

- **Wilson e Megeed, 1990**, obtiveram outro modelo para a região sudoeste da província de New Brunswick, Fredericton, no Canadá, correspondendo a uma área total de 30.000 km². O ano base para a calibração do modelo foi 1979, utilizando dados interpolados de 2 censos. Também as variáveis explicativas são relacionadas a características sócio-econômicas e demográficas, selecionadas pelo método *stepwise* através da análise de regressão linear. O modelo é:

$$Y = -2086,4 + 71,26 \text{ DENS} + 2,81 \text{ FOR} + 0,251 \text{ ADI}$$

onde:

Y = número de viagens diárias produzidas pela região,

DENS = número de habitantes por km²,

FOR = área de extração de madeira.

ADI = renda média per capita na região.

- A Cia do Metropolitano de São Paulo, durante os anos de 1968, 1975, 1987 e 1991 formulou modelos agregados para descrever a geração de viagens classificadas por diversos motivos. O modelo desenvolvido em 1968 era expresso por:

$$\log V = -0,8153 + 0,1772 \log R + 1,0181 \log D - 0,036 Z_1 - 0,1824 Z_3$$

onde:

V: é o número de viagens total,

R: é renda média por domicílio,

D: é densidade demográfica bruta: população / área em ha.

Z₁: é uma variável *dummy* para indicar algumas regiões.

Z₃: é uma variável *dummy* para indicar outras regiões

Um outro modelo, calibrado em 1975, era dado por:

$$V = 88,2334 X_1 + 0,04676 X_1 X_2$$

onde:

V = número de viagens por motivo de trabalho,

X_1 = população em milhares,

X_2 = renda média familiar em unidades monetárias vigentes.

Um terceiro modelo, obtido em 1991, é:

$$V = 0,186267 \text{ Pop} + 0,0025 \text{ Ren}$$

onde:

V = número de viagens por motivo de trabalho,

Pop = população da zona,

Ren = renda total da zona.

Os impactos que as mudanças atuais nas estruturas familiares têm sobre o comportamento humano influenciam as decisões do indivíduo também enquanto usuário de transportes, o que faz com que a projeção das viagens geradas usando enfoques agregados seja inadequada.

3.1.2 Modelos desagregados

A inadequação dos modelos agregados levou à busca de outros enfoques que tivessem unidade de análise mais homogênea que a zona de tráfego. Diz-se que os modelos são mais desagregados quanto maior for o nível de informação individual preservada. Há duas abordagens básicas: desagregar-se em nível de domicílios e em nível individual.

3.1.2.1 Modelos baseados em domicílios

Desde a década de 70 a unidade amostral mais difundida na literatura é o domicílio e não o indivíduo. O argumento principal está baseado na relevância das interações interpessoais dentro do domicílio. Mais especificamente, crê-se ser necessário usar o domicílio para incorporar seu efeito sobre o indivíduo (Vuren, Gunn & Daly, 1995; Wooton & Pick, 1967), através de variáveis que caracterizem o domicílio tais como:

- número de pessoas empregadas,
- número de moradores ou membros do domicílio,
- renda familiar,
- número ou disponibilidade de automóveis, etc.

O único caso encontrado na revisão bibliográfica, que incorporou com sucesso variáveis não tradicionais como as citadas, é apresentado por Said & Young (1990) e Said, Young & Ibrahim (1991).

Os autores perceberam os erros que poderiam ser cometidos ao tratar-se grupos de domicílios como sendo homogêneos numa sociedade multi-étnica como a do Kuwait, onde 40% são kuwaitianos, 38% árabes e 21% asiáticos. Os kuwaitianos são caracterizados como uma população muito jovem, os árabes e asiáticos são de idade mediana onde o contingente masculino é o dobro do feminino. As mulheres kuwaitianas não têm participação significativa na população economicamente ativa, ao contrário das demais, devido às leis do Kuwait no que diz respeito à necessidade de participação na força de trabalho por parte dos estrangeiros.

Os tamanhos dos domicílios são muito diferentes. O número de membros da família kuwaitiana é muito grande, em média é 10, dos árabes é 5,4, e dos asiáticos é 3,9 (o mesmo que o tamanho médio das famílias da Região Metropolitana de São Paulo segundo o SEADE (1992a)).

Os autores utilizaram modelos lineares generalizados para explicar o número médio de viagens domiciliares segundo variáveis tais como: número de adultos no domicílio, número de crianças, número de automóveis e algumas interações. Os modelos foram ajustados separadamente, por grupo étnico. Os resultados obtidos foram bastante satisfatórios.

Outro trabalho que merece destaque é o apresentado por McDonald & Stopher (1983). Além das variáveis tradicionais:

- número de membros do domicílio

- número de automóveis
- número de membros empregados
- número de membros licenciados para dirigir
- renda familiar

os autores introduziram outras variáveis, tais como:

- número de membros maiores que 12 anos
- tipo do domicílio: multifamiliar ou unifamiliar
- raça
- número de pessoas em idade pré-escolar
- estrutura do domicílio

A estrutura do domicílio foi descrita inicialmente por algumas categorias, que são reagrupadas de várias formas ao longo da pesquisa:

- domicílio unipessoal classificado por sexo;
- famílias sem um dos cônjuges;
- casais sem crianças;
- casais com crianças;
- famílias adultas com crianças;
- famílias adultas sem crianças;
- grupo convivente sem relação de parentesco.

Crerérios subjetivos, sem muito rigor, levaram a classificação das famílias como sendo adultas.

Alguns modelos lineares foram ajustados para diversas combinações das variáveis e de seus níveis. Análise de variância e análise de classificação múltipla foram utilizadas (Stopher & McDonald, 1983). Os autores concluíram que as variáveis ligadas à estrutura do domicílio, embora significantes, não merecem destaque, devendo, até mesmo, ser eliminadas. Argumentam que variáveis ligadas às características do domicílio não são alvos estratégicos de políticas de planejamento de transportes, bem como apresentam dificuldades para categorizar os domicílios.

Exemplos de modelos baseados em domicílios:

- Rengaraju e Satyakumar, 1994, propuseram um modelo utilizando regressão linear múltipla. A finalidade era estimar o número de viagens de categorias pouco representadas na amostra quando se utiliza análise de categorias para estudar o número de viagens geradas. Um exemplo dos modelos obtidos é apresentado a seguir:

$$Y = 1,205 + 0,23 X_1 + 0,12 X_2 + 0,09 X_1 X_2,$$

onde:

Y = número médio de viagens por domicílio,

X₁ = tamanho do domicílio,

X₂ = renda total mensal do domicílio.

- White, 1976, utilizou dados do Western Midlands Rural Travel Survey - 1972 e desenvolveu o modelo abaixo:

$$T/U = - 0,926 (1/U) + 1,938 - 0,152 P/U + 0,955 E/U + 1,568 C/U$$

onde:

T = número médio de viagens por domicílio,

P = número de pessoas do domicílio com idade superior ou igual a 5 anos,

E = número de pessoas do domicílio empregadas em período integral,

C = 1 se o domicílio possui mais de um automóvel,

0 caso contrário.

U = combinação linear dos resíduos de um pré-modelo, que tem a finalidade de transformar as variáveis originais para atingir-se a hipótese de homocedasticidade necessária ao uso da técnica de regressão (variâncias iguais).

3.1.2.2 Modelos baseados em indivíduos

O objetivo da desagregação em nível individual é a obtenção de um modelo que represente o número de viagens que um indivíduo faz, por unidade de tempo, em função de algumas características dos indivíduos ou dos domicílios.

Segundo Ortúzar & Willumsen (1994), modelos individuais foram propostos pela primeira vez por Januz Supernak em 1979, que mesmo vislumbrando algumas vantagens e dificuldades que serão comentadas no item seguinte, incentivou-se com os bons resultados obtidos em suas pesquisas, acreditando poder incorporar as inter-relações domiciliares num modelo individual.

Em 1982, Tye *et al.* apresentam uma breve discussão sobre um modelo de geração desagregado em nível individual. Trata-se do modelo Logit aplicado à

geração de viagens. O objetivo era estimar a probabilidade de um indivíduo realizar uma ou mais viagens. Os resultados não foram satisfatórios e a pesquisa nesta direção não teve continuidade.

Supernak, Talvitie & De John (1983) apresentam a versão final do que chamam de modelo de geração de viagens categorizado individualmente. Trata-se da Análise de Categorias aplicada a dados individuais coletados em Baltimore em 1977.

Um destaque especial será dado a este trabalho por ser o único encontrado na revisão bibliográfica com enfoque individual na modelagem de geração de viagens.

O modelo foi desenvolvido em 4 estágios, utilizando 1903 indivíduos. No primeiro estágio escolheram-se arbitrariamente muitas variáveis que segundo o autor poderiam ser consideradas como importantes para explicar as variações na mobilidade individual:

- Sexo;
- Idade: considerada em 4 faixas etárias;
- Disponibilidade de automóvel. Considerada de duas maneiras:
 - a) quantitativamente: nenhum, um, dois ou mais automóveis
 - b) qualitativamente: nunca disponível, às vezes e sempre disponível
- Renda individual: considerada como o quociente renda familiar / número de membros do domicílio, classificada em 3 faixas;

- Situação ocupacional: empregado ou não;
- Raça: branca ou não;
- Tipo de emprego: classificado em 3 tipos;
- Tipo de família: classificado segundo a composição de seus membros: individual, casal sem filhos, famílias com filhos menores que 5 anos, famílias com filhos entre 5 e 12 anos e famílias com filhos maiores que 12 anos.

Entre os cruzamentos possíveis obtidos com os níveis das variáveis acima, 100 foram escolhidos subjetivamente para formar as categorias.

Para determinar quais variáveis são realmente importantes na construção destas categorias, examinou-se os indivíduos segundo cada variável isoladamente da seguinte maneira. Através do teste de comparação de 2 médias, ao nível de significância de 1% e utilizando a distribuição normal (teste Z com $\alpha = 0,01$) testou-se:

- A diferença entre os sexos para cada nível das demais variáveis.
- A diferença duas a duas entre os 3 níveis de idade para cada nível das demais variáveis...

e assim por diante, para cada variável.

Os testes acima são para 2 parâmetros: número médio de viagens diárias e tempo médio de viagem.

As variáveis que sistematicamente não apresentavam diferenças significantes entre seus níveis, foram eliminadas. São elas: raça, tipo de emprego e tipo de família. Também foram eliminados alguns níveis das variáveis que foram mantidas no modelo.

O estágio 2 começa com a repetição do estágio 1 com algumas alterações nos níveis das variáveis que permaneceram no modelo. Também utilizou-se a análise de regressão linear para ajustar o modelo $N_i = b_{ij} + m_{ij}N_j$ onde:

N_k = número de viagens de categoria $k = i, j$.

b_{ij} = intercepto da relação entre as categorias 'i' e 'j'.

m_{ij} = coeficiente angular da relação entre as categorias 'i' e 'j'.

A categoria 'i' é considerada similar à categoria 'j' e portanto eliminada da análise, se $N_i \cong N_j$. Sendo a relação de proximidade representada pelo símbolo $\cong$, definida segundo o critério abaixo:

$N_i \cong N_j$ se:

$$\gamma_{ij} > 0.9$$

$$0.75 < m_{ij} < 1.25$$

$$|b_{ij}| < 0.10$$

onde γ_{ij} = coeficiente de correlação linear entre as taxas de viagens das categorias 'i' e 'j'.

O final do estágio 2 obteve como resultado que somente a idade, considerada em 3 faixas e a situação ocupacional, considerada em 2 níveis, são altamente significantes. As demais variáveis, sexo, disponibilidade de automóvel e renda poderiam ser usadas ou não. Segundo o autor, completou-se o conjunto final de variáveis, acrescentando-se somente a disponibilidade de automóvel, para tornar este modelo compatível com os demais (sic). Para confirmar a saída das variáveis sexo e renda realizou-se uma análise de variância bidimensional.

Com os cruzamentos dos níveis das 3 variáveis resultantes deste processo de escolha: idade, situação ocupacional e disponibilidade de automóvel, formou-se 8 categorias consideradas importantes, esquematizadas abaixo. Cada quadrado representa uma categoria, contendo o número de indivíduos.

O estágio 3 oferece uma análise, com bons resultados, da capacidade de transferência do modelo.

O estágio 4 diz respeito à comparação deste modelo com um modelo baseado em domicílios.

3.1.3 Considerações sobre modelos desagregados

A literatura tem apontado a superioridade dos modelos desagregados, em nível de domicílios, quando comparados com modelos agregados. Várias são as vantagens: maior capacidade de transferência (Supernak, 1981); previsões mais confiáveis (Douglas, 1973; Daly & Ortúzar, 1990; etc); custos reduzidos devido à possibilidade de calibração com menores amostras (Horowitz, 1983; Tye *et al.*, 1982); são mais avançados por possuírem um caráter mais comportamental (Kollo & Purvis, 1989), etc.

Os custos envolvidos na calibração de modelos desagregados de forma geral, são muito inferiores aos custos envolvidos para a calibração de modelos agregados. É o caso, por exemplo, dos modelos de utilidade aleatória que descrevem o processo de escolha do usuário dentro do contexto de análise de demanda por transportes.

Os modelos desagregados em nível de indivíduos podem ser calibrados utilizando-se algumas centenas de unidades amostrais, enquanto que um modelo representativo do comportamento de uma família ou de uma zona requer algumas dezenas de milhares de unidades amostrais. Intervalos de confiança precisos têm sido obtidos com amostras com 200 observações, contudo para fins de estimação da mobilidade de uma zona é necessário amostra bem maior para representar todos os níveis das variáveis.

Não está claro na literatura como as características domiciliares, de forma geral, influenciam o comportamento do indivíduo, muito menos como medir esta influência. Contudo existe um consenso no sentido de aceitar esta influência nas decisões do usuário de transporte. Este é o motivo principal para a popularidade dos modelos desagregados em nível domiciliar, como se só isto bastasse para captar esta influência.

Apesar disto, não foi possível até o presente momento incorporar-se as inter-relações domiciliares de forma geral num modelo com enfoque individual, nem no modelo com enfoque domiciliar.

Concluindo, informações importantes continuam escondidas, havendo perda de informação ao tratar-se de forma similar domicílios que são homogêneos segundo algumas características gerais e são diferenciados quanto a renda, disponibilidade de automóveis, tamanho (número de membros), etc.

Mesmo assim o modelo individual não teve sustentação suficiente para ser difundido, nem razões lógicas fortes foram apresentadas em seu favor, excetuando-se a sua maior capacidade de transferência.

3.2 Discussão dos métodos analíticos usuais para a modelagem de geração de viagens

Existem dois tipos principais de métodos para o estudo da geração de viagens:

- a análise de regressão linear e

- a análise de categoria.

Ambos podem ser aplicados em diferentes níveis de desagregação.

Convém organizar a discussão destes métodos em três sub-itens, dois deles dedicados somente a esclarecer o que vem a ser a análise de categorias e análise de regressão linear, e um terceiro que os compara.

Num terceiro item a discussão amplia fundamentos teóricos mais elaborados para ambas as técnicas e inclui a noção de modelos lineares gerais que engloba ambas e ainda a análise de variância e covariância.

3.2.1 A análise de categoria

Segundo Stopher & McDonald (1983) a análise de categoria aparece como método alternativo na modelagem de geração de viagens de passageiros urbanos, desde a década de 1950. No final da década de 60 torna-se mais popular. Este método é conhecido como classificação cruzada na América do Norte e análise de categoria no Reino Unido. Segundo Douglas (1973) este método foi o mais utilizado para dados desagregados até a década de 70 e o crescente interesse no uso de modelos desagregados fez ressurgir a técnica de análise de categorias formulada de uma maneira mais desagregada, mais recentemente.

A análise de categoria nada mais é do que a obtenção da taxa média de viagens para categorias (grupos) de unidades amostrais, partindo do pressuposto que estes grupos são homogêneos internamente.

Por exemplo, tratando-se de domicílios, produz-se uma classificação dos domicílios ou grupos de domicílios considerados homogêneos, segundo os cruzamentos dos valores de algumas variáveis descritivas do próprio domicílio. Geralmente algumas destas variáveis descritivas são de caráter sócio-econômico e cada cruzamento obtido gera uma categoria de domicílios. Obtém-se a taxa média de mobilidade para cada cruzamento e esta taxa é associada a cada categoria (Kanafani, 1983 e Douglas, 1973).

Até a década de 70, a análise de categoria foi o método mais utilizado para dados desagregados (Douglas, 1973) e ainda hoje é largamente utilizado a despeito da grande disponibilidade de recursos computacionais eficientes que proporcionam o uso de técnicas mais acuradas.

A fragilidade desta técnica de análise como usualmente aplicada consiste no fato de não se tratar de um processo de modelagem propriamente dito e sim um processo classificatório simplesmente, ou seja, um processo de análise descritiva de dados estatísticos, ao invés de um processo de inferência estatística como o é a análise de regressão. Isto significa que não se pode avaliar a qualidade do ajustamento (não se tem sequer um modelo na acepção da palavra) nem se pode fazer previsões (Said & Young, 1990; Stopher & McDonald, 1983). Os resultados obtidos limitam-se ao cálculo das médias das categorias, sem preocupação de entender e quantificar a importância dos fatores em estudo na descrição do padrão de deslocamentos.

Ressalta-se também que a análise de categoria só atinge seu objetivo para amostras muito grandes, onde todas as categorias possam ser bem representadas (Douglas, 1973).

É comum ter-se categorias sem representatividade na amostra e até mesmo categorias vazias. Rengaraju e Satyakumar (1994) formularam um modelo de regressão linear para explicar o número de viagens geradas pelos domicílios de categorias vazias ou quase vazias. Usam mínimos quadrados ponderados pelo número de domicílios em cada categoria. As variáveis explicativas são os valores das variáveis, que cruzadas, formam as categorias. Entre as conclusões ressaltam que o modelo criado é mais eficaz, que a própria análise de categoria, para explicar as variações nas taxas de viagens e detectam as variáveis de maior influência através da interpretação dos parâmetros estimados.

Supernak, Talvitie & De John (1983) descrevem como resolver o problema da necessidade da amostra ser muito grande, através do uso de várias técnicas estatísticas com a finalidade de diminuir o número de categorias necessárias e portanto o tamanho da amostra. Além de uma exaustiva descrição estatística dos dados através de gráficos e tabelas que apontem evidências de categorias desnecessárias, um teste de hipóteses estatísticas é conduzido com a finalidade de verificar níveis desnecessários dentro de cada variável. Posteriormente, com as categorias remanescentes procede-se uma regressão linear entre as taxas médias de viagens (T_k) num modelo do tipo ($T_i = a + bT_j$). O intuito é agrupar numa única categoria aquelas em que resultem em, aproximadamente, $a = 0$ e $b = 1$, portanto em taxas médias quase iguais.

A partir de então apresentam como resultado principal e final o recálculo das taxas médias das categorias que sobraram, como usualmente se faz.

Stopher & McDonald (1983) sugerem o uso da Análise de Classificação Múltipla (MCA); método muito utilizado em Ciências Sociais (segundo os autores) para analisar dados quantitativos. Este método está baseado em

análise de variância e permite alguns critérios de redução do número de categorias e a estimação das médias dos grupos utilizando a média geral adicionada a um desvio causado pelo grupo.

Apesar destes trabalhos, não está bem estabelecida na literatura, uma forma de escolha de quantas e quais categorias utilizar (Said & Young, 1990). Entretanto, o problema maior não é este, nem o fato de os dados estarem agrupados em categorias, mas a fragilidade desta abordagem, reside na pouca informação contida nos resultados gerados, já que desprezam-se informações importantes que poderiam ser extraídas dos dados, além das médias de cada categoria.

A literatura aponta evidências de que métodos simples, porém mais acurados que o método usual de análise de categorias, podem ser utilizados, sugere ainda que o uso da técnica de regressão para um grau mais adequado de categorização dos domicílios, seja melhor investigado (Douglas, 1973).

3.2.2 A análise de regressão linear

A análise de regressão linear é um método bem conhecido por pesquisadores de muitas áreas do conhecimento e portanto dispensa uma apresentação mais detalhada.

Constrói-se modelos onde se explica o número de viagens geradas pela unidade amostral de interesse através de relações lineares (nos parâmetros) de variáveis, em geral, sócio-econômicas.

Em muitos casos, a linearidade pode ser alcançada através do uso de variáveis dummy (Lewis, 1970; Montgomery & Peck, 1983; etc.).

“Construir um modelo de regressão é um processo iterativo. Usualmente várias análises são necessárias como implementações na estrutura do modelo e falhas nos dados são descobertas. Um bom programa de computador para fazer a regressão é necessário neste procedimento. O computador não é um substituto para o pensamento criativo sobre o problema. A análise de regressão requer a inteligência e o hábil uso do computador” (Montgomery & Peck, 1983 p.7).

“O conjunto de dados a ser utilizado deve resultar de um experimento controlado especificamente delineado para a coleta de dados ou de registros históricos existentes” (Montgomery & Peck, 1983 p.9). A exigência básica de não correlação nos erros do modelo de regressão geralmente é garantida no processo de aleatorização de coleta da amostra.

“Este é um ponto essencial da análise de regressão, porque já as conclusões das análises dependem dos dados, o esforço para uma boa coleta de dados pode produzir mais benefícios, incluindo uma análise simplificada e uma aplicação mais geral do modelo.

Os dados... devem ser representativos do sistema estudado... Sem esta representatividade... a maioria das técnicas usadas em análise de regressão pode ser seriamente distorcida levando a conclusões erradas” (Montgomery & Peck, 1983 p.5).

Duas são as críticas mais freqüentes encontradas na literatura:

- o mau uso que se faz desta técnica de análise em virtude da não verificação das hipóteses necessárias (White, 1976),
- a má interpretação dos resultados por aqueles que desconhecem os fundamentos estatísticos da análise de regressão e contudo sabem operar os *software* disponíveis para os cálculos (Douglas, 1973).

Embora ambos causem conclusões inverídicas a não verificação das hipóteses merece atenção especial. É necessário que os erros sejam não só aleatórios, mas também não relacionados, de variâncias iguais e normalmente distribuídos. Estas hipóteses são necessárias para que se possa aplicar os testes de significância conhecidos (t-student e F-Snedecor) bem como para que se possa olhar para as estimativas obtidas pela técnica de mínimos quadrados como sendo provenientes de estimadores não viesados e de variância mínima. Isto quer dizer que no caso da violação destas hipóteses os estimadores não são confiáveis, podem ser viesados e portanto não têm mais variância mínima. O mesmo ocorre no caso da presença de forte grau de multicolinearidade.

É interessante observar que, mesmo para modelos desagregados em nível de domicílios, para algumas regiões, não se consegue atingir a hipótese de homocedasticidade dos erros, ou seja, os resíduos não apresentam a mesma variância (White, 1976; Said, Young & Ibrahim, 1991). Outros artifícios, como transformações das variáveis originais (Siqueira, 1983) são então utilizados para que se possa obter resultados confiáveis através dos estimadores de mínimos quadrados no modelo de regressão linear.

Tal é a importância, uso e aplicabilidade desta técnica que o anexo A dedica-se a tratar estas questões com maiores detalhes, aconselhando-se sua leitura, onde as dúvidas mais comuns com relação a interpretação dos resultados e verificação da adequação das hipóteses aos dados podem ser esclarecidas.

3.2.3 O modelo linear geral

A utilização do modelo linear para descrever a taxa de viagens em função de outras variáveis traz consigo as vantagens de se obter uma medida da qualidade do ajuste, testes de significância sobre os efeitos de cada variável, uma medida da importância de cada uma para o modelo, e em alguns modelos a possibilidade de fazer-se previsões com uma margem de erro conhecida, etc.

Traz também um conjunto de hipóteses mais fortes sobre os dados, o que é dispensável para o uso de análise de categorias como usualmente se faz, como a normalidade, a independência e a homocedasticidade (igualdade das variâncias). O que não fica claro para o usuário comum de técnicas estatísticas é que o modelo linear engloba técnicas como análise de variância, de

covariância e de regressão (Neter & Wasserman, 1974 e Neter, Wasserman & Kutner, 1983).

- Se todas as variáveis independentes são qualitativas, tem-se o modelo de análise de variância para efeitos fixos (j): $Y_{ij} = u + u_j + e_{ij}$.
- Se todas são quantitativas tem-se o modelo de análise de regressão: $Y_{ij} = u + \beta X_{ij} + e_{ij}$.
- Se algumas são quantitativas e outras qualitativas, tem-se o modelo de análise de covariância: $Y_{ij} = u + u_j + \beta(X_{ij} - \bar{X}_{..}) + e_{ij}$.

onde:

u : é uma média geral, componente constante

u_j : são os efeitos específicos do grupo e sujeitos à restrição $\sum n_i u_i = 0$

β : é o coeficiente de regressão da relação de Y e X

X_{ij} : chamadas de variáveis explicativas ou preditoras, são não aleatórias

e_{ij} : são os erros ou desvios aleatórios do modelo, têm distribuição $N(0, \sigma^2)$ e são independentes

É interessante ressaltar que a análise de variância e a de covariância podem ser formuladas como um modelo de análise de regressão e que testar a igualdade dos efeitos dos tratamentos é equivalente a testar a nulidade dos parâmetros do modelo de regressão (Neter, Wasserman & Kutner 1983).

Said & Young (1990) obtêm resultados muito bons e interessantes com a aplicação do modelo linear geral em dados domiciliares no Kuwait, onde a especificidade das características dos domicílios inviabilizam o uso de análise de categorias na forma tradicional.

Rengaraju & Satyakumar (1994) pretendendo estruturar a análise de categoria como técnica de inferência estatística, propõem o uso do modelo linear somente com a finalidade de estimar as taxas de viagens de categorias vazias ou mal representadas e mesmo reconhecendo a superioridade deste estimador em relação à média simples das demais categorias, não se arriscaram a abandonar definitivamente a forma usual de fazer a análise de categorias.

Supernak, Talvittie & De John (1983) utilizam as técnicas de análise de variância e de regressão somente com o intuito de diminuir o número de categorias com o objetivo de estimar as taxas médias de viagens da forma usual.

Conclui-se que de forma geral (excetuando-se Said & Young (1990)) está havendo um desperdício do uso de modelos lineares no sentido do uso de uma ferramenta poderosa simplesmente para diminuir categorias, pois o estimador final continua sendo a média simples dos n dados disponíveis na categoria, desprezando-se inclusive a possibilidade de efetuar-se previsões, verificar-se a qualidade do ajustamento, a magnitude e a importância dos efeitos dos fatores que influenciam o padrão de deslocamentos.

3.2.4 Comparações entre os métodos analíticos

Uma discussão inicial sobre este assunto pode ser encontrada em Douglas & Lewis (1970, a - b; 1971 a - b) bem como em Starkie (1970).

A questão básica levantada quanto ao método de análise de categorias está no fato de que, sob este título, o que se encontra na literatura são somente cálculos de médias das categorias e a denominação desta tabela de médias como sendo um modelo. Neste contexto, o método de análise de categorias é muito rudimentar do ponto de vista estatístico, em virtude não só da fragilidade estatística da média, mas sobretudo da existência de alternativas mais apropriadas para tratar dados estatísticos com o fim de fazer inferências.

A situação mais próxima a esta prática com características de modelagem é a consideração do modelo de análise de variância, $Y_{ij} = u + u_{ij} + e_{ij}$, onde os estimadores de mínimos quadrados também são as médias dos grupos ou categorias. Apesar de neste contexto serem disponíveis testes estatísticos dos efeitos dos tratamentos, não se pode fazer previsões nem estimar uma possível relação de causa e efeito, quando devidamente fundamentada, entre os fatores estudados e a variável dependente.

O fato de os dados estarem agrupados em categorias não é problema para o uso de modelos lineares. Contudo, para a aplicação da técnica de regressão alguns cuidados devem ser observados e sugere-se a leitura de Haitovsky (1973).

O modelo linear, na maioria das vezes representado pelo método da regressão, preenche os quesitos para a adequada interpretação e produção de resultados estatísticos e mostra-se superior aos demais quanto ao tamanho de amostra necessário, bem como quanto à obtenção de dados quantitativos.

Ao utilizar-se o modelo de análise de regressão para o estudo da geração de viagens, o objetivo é estimar $E(Y/X_i) = \alpha + \beta X_i$ ou genericamente $Y = \beta X$. Algumas vantagens deste modelo são:

- testes podem ser feitos sobre os parâmetros, bem como podem ser determinados intervalos de confiança.
- previsões podem ser feitas com a estimação das margens de erros ($\hat{Y}$) através do chamado intervalo de previsão (o mesmo não pode ser feito através da técnica de análise de variância).
- pode-se estimar as médias de viagens para valores dos fatores de influência (X_h) não amostrados, porém dentro do campo de variação de X estudado, com a determinação da precisão destes estimadores através dos intervalos de confiança para $\hat{Y}$ (também não pode ser feito através da técnica de análise de variância).
- pode-se interpretar as diferenças e detectar os fatores mais importantes, através do teste estatístico de hipóteses sobre os parâmetros.

É interessante analisar a razão entre as variâncias dos estimadores gerados por ambos os métodos, análise de categorias e análise de regressão, bem como analisar a razão entre as amplitudes dos intervalos de confiança.

Em ambos os casos, os estimadores obtidos pela técnica de mínimos quadrados na regressão linear, são mais ou tão precisos quanto os obtidos pela análise de categorias, utilizando-se amostras bem menores, conforme apontam os cálculos abaixo.

Seja $\hat{Y}$ o estimador de mínimos quadrados de Y no modelo de regressão $Y = X\beta + e$, quando $X = x$, sendo x o valor assumido pela variável X .

A expressão analítica para $\text{var}(\hat{Y})$ é dada por $\sqrt{x'(X'X)^{-1}x} \cdot \sigma^2$.

Aplicando a expressão acima para as n observações pode-se mostrar (Draper & Smith, 1981) que para qualquer modelo linear vale:

$$\sum_{i=1}^n \frac{V(\hat{Y}_i)}{n} = \text{traço} \{X(X'X)^{-1}X'\} \frac{\sigma^2}{n} = p \frac{\sigma^2}{n}$$

onde p = número de parâmetros a serem estimados (incluindo o intercepto quando for o caso).

O que torna razoável a aproximação de $\hat{V}(\hat{Y}_i)$ em média por $p \frac{\hat{\sigma}^2}{n}$.

Quando se utiliza a análise de categorias ou análise de variância a fatores fixos, o estimador em questão é a média de cada categoria i obtida pelas n_i observações correspondentes: $\bar{X}_i$.

$$\text{Ora } V(\bar{X}_i) = \frac{\sigma^2}{n_i}.$$

Ao se comparar ambas as variâncias:

$$\hat{V}(\bar{X}_i) = \frac{\hat{\sigma}^2}{n_i}$$

e

$$\hat{V}(\hat{Y}_i) = p \frac{\hat{\sigma}^2}{n}$$

Observa-se que $n = \sum_{i=1}^k n_i$ onde k é o número de categorias.

$$\text{Se } n_i \text{ são todos iguais então } n = kn_i \text{ e } \frac{\hat{V}(\hat{Y}_i)}{\hat{V}(\bar{X}_i)} = \frac{\frac{p\hat{\sigma}^2}{kn_i}}{\frac{\hat{\sigma}^2}{n_i}} = \frac{p}{k}.$$

Então se o número de variáveis for menor que o número de categorias $p < k$, tem-se $\hat{V}(\hat{Y}_i) < \hat{V}(\bar{X}_i)$.

Esta é a situação mais comum na prática. Para duas variáveis, por exemplo idade e renda, cada uma subdividida em 2 níveis, tem-se $p = 2$ e $k = 2 \cdot 2 = 4$ categorias, obtidas pelos cruzamentos dos níveis das duas variáveis.

Os intervalos de confiança obtidos através dos dois estimadores são:

$$\bar{X}_i \pm t_{n_i-1} \frac{\hat{\sigma}}{\sqrt{n_i}} \text{ e } \hat{Y}_i \pm t_{n-p-2} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{p}$$

A razão entre suas amplitudes é dada por:

$$\frac{IC(\bar{X}_i)}{IC(\hat{Y}_i)} = \frac{\frac{t_{n_i-1}}{\frac{\hat{\sigma}}{\sqrt{n_i}}}}{t_{n-p-2} \frac{\hat{\sigma}}{\sqrt{n}} \sqrt{p}} = \frac{\frac{t_{n_i-1}}{\sqrt{n_i}}}{t_{n-p-2} \sqrt{\frac{p}{n}}}$$

Para que ambos os intervalos tenham mesma amplitude (razão igual a 1) tem-se:

$$n_i = \frac{t_{n-p-2}^2}{t_{n_i-1}^2} \frac{n}{p}$$

Se $t_{n-p-2}^2 \simeq t_{n_i-1}^2$, o que ocorre para graus de liberdade acima de 20 se o nível de confiança é de 95% (t fica praticamente constante para graus de liberdade acima de 120, independentemente do nível de confiança), então:

$$n_i = \frac{n}{p} \text{ ou } n = n_i p \text{ com } n_i > 20 \text{ ou seja:}$$

Se tivermos duas variáveis (idade e renda) basta $n = 2n_i = 40$ observações no total para que o intervalo de confiança de $\hat{Y}_i$ tenha a mesma amplitude que o intervalo de confiança da categoria i no método usual, se esta contiver 20 observações ou mais.

4. INFLUÊNCIA DAS CARACTERÍSTICAS DO DOMICÍLIO SOBRE O INDIVÍDUO

No início dos anos 80, surgiu grande interesse em enriquecer os modelos de geração de viagens com teorias e métodos advindos das ciências do comportamento. A principal hipótese que norteou estes estudos é que as circunstâncias sociais na qual o indivíduo vive, contribuem consideravelmente para ampliar ou restringir seu padrão de atividades. Por exemplo, um casal sem filhos ou com filhos adultos deve ser mais móvel que um casal com filhos bem pequenos, em idade pré-escolar.

Contudo, os resultados não foram satisfatórios. No presente trabalho, buscou-se na sociologia e psicologia social a sustentação teórica para a inclusão de características do domicílio, no modelo individual de geração de viagens.

4.1 Inserção deste tema no contexto geral de análise de demanda por transportes

Deslocamento de pessoas no espaço urbano faz parte das atividades desenvolvidas pela população numa sociedade industrial. As viagens são o meio ou o instrumento que possibilita a participação dos indivíduos em atividades, mesmo as mais primárias (Kutter, 1973; Carpenter & Jones, 1983; The 1985 International Conference on Travel Behaviour, etc). Portanto uma análise dos deslocamentos deve começar com uma análise do comportamento humano no que se refere ao desempenho de atividades.

Durante os últimos 30 anos, vários autores dedicaram-se ao estudo de modelos de análise de demanda baseados em atividades. Fox (1995) oferece uma revisão da literatura sobre este enfoque buscando entender o comportamento individual e do domicílio em resposta às necessidades de viagens e a satisfação de demanda por transportes geradas pelas diversas atividades.

A maior parte dos chamados modelos de atividades permite simular o modo como se usa o tempo, mediante uma série de restrições, tais como aspectos físicos, espaciais e sociais. Há um consenso geral em aceitar que a característica essencial é interpretar que os deslocamentos são uma consequência da escolha de realizar-se atividades, portanto uma consequência da escolha de como se utilizar o tempo (Clarke, 1986; Clarke *et al*, 1981).

De forma geral pode-se subdividir as variáveis envolvidas nestes modelos em 3 grupos:

- Variáveis que identificam a demanda por atividades: são aquelas que avaliam as tentativas do indivíduo em realizar as atividades necessárias. São derivadas não somente de necessidades biológicas, como ir ao supermercado para abastecer-se de alimentos, mas também de forças sociais, por exemplo a necessidade de assumir os cuidados de uma casa ou cuidados de filhos.
- Variáveis que identificam a oferta de atividade: são aquelas que analisam as restrições no tempo e no espaço para a realização de tais atividades. Podem ser derivadas de papéis desempenhados dentro do ciclo da vida, por exemplo, a maternidade pode modificar a demanda de atividades ligada ao lazer, trabalho fora de casa, atividades culturais, etc. Estas variáveis

constituem fator de restrição para algumas atividades, entretanto fator gerador de outras.

- Variáveis que representam o comportamento do indivíduo: segundo Damm (1983) o estudo de hipóteses empíricas ou teóricas neste grupo de variáveis é o mais importante campo de desenvolvimento destes modelos.

Excluindo itens específicos relacionados as atividades, pode-se ampliar este contexto para os modelos de demanda de forma geral.

As hipóteses envolvidas podem ser divididas em alguns grupos:

- Hipóteses sobre os fatores que afetam o comportamento humano na hora de decidir-se por participar de certa atividade: prioridades, valores filosóficos e morais, preferências, etc.
- Hipóteses sobre a influência de características individuais: idade, estilo de vida, escolaridade, etc.
- Hipóteses sobre as características ligadas à família: número de filhos, número de automóveis, renda familiar, etc.
- Hipóteses sobre as restrições do ambiente externo: tempo, distância, modo de transporte disponível, horário comercial, etc.

Estes grupos de hipóteses podem ser incluídos em 2 mais gerais:

- Hipóteses Lógicas: são uma expressão explícita causal e coerente de porque as pessoas agem como agem com relação à participação ou

desenvolvimento de atividades. Não são hipóteses passíveis de teste empíricos. Por exemplo, várias hipóteses podem ser formuladas para tentar explicar o porquê as pessoas escolhem (se é que escolhem) ter ou não ter filhos. Em muitos casos as explicações transcendem o próprio entendimento do indivíduo questionado e qualquer teste aplicado é pura aproximação da realidade.

- Hipóteses que surgem dedutivamente: são resultados da análise de fatores que se espera que expliquem o comportamento de alguma variável. Estas hipóteses são passíveis de análise empírica. Normalmente as variáveis que se deseja explicar são relacionadas a dados de deslocamentos e não de atividades propriamente ditas.

Neste último grupo é que se insere o contexto desta tese.

Este trabalho desenvolve-se com o intuito de mostrar a necessidade de acrescentar variáveis domiciliares ao modelo individual de geração de viagens de passageiros urbanos. Tais variáveis devem incorporar fatores tais como:

- 1) a caracterização do grupo com o qual convive o indivíduo, do ponto de vista das relações de parentesco;
- 2) o papel que o indivíduo desempenha no grupo;
- 3) o estágio no ciclo da vida do indivíduo ou do grupo familiar.

4.2 A necessidade de maior compreensão do homem em seu meio ambiente

Entre as limitações e problemas encontrados nos modelos tradicionais de transportes em geral, destaca-se a ausência de relação entre a demanda por transportes e fatores gerados por necessidades psicológicas, biológicas e sociais do homem.

Ash (1977) oferece uma breve discussão sobre as doutrinas a respeito do homem, destacando que a psicologia ocupa, no estudo do homem, uma posição ímpar e de comando uma vez que todas as atividades sociais, econômicas, políticas, artísticas, etc, têm seu centro em indivíduos.

Os psicólogos não são os únicos que tentam compreender os seres humanos e seu comportamento. Para os psicólogos é fácil acreditar que uma teoria psicológica científica constitui a única base sólida para uma compreensão do homem e de seu comportamento. Contudo é possível fazer-se outras suposições e explicar, de outra forma, o comportamento humano (Baldwin, 1967).

É preciso convir que, antes da investigação psicológica, muito se compreende das coisas humanas. Poder-se-ia, até, reconhecer a superioridade, em muitos aspectos, da compreensão diária sobre o conhecimento hoje apresentado pela psicologia. Mas a compreensão diária não proporciona um conjunto científico de teoria e fatos a respeito do homem. Tudo isto fornece apenas um conhecimento ingênuo que deve ser diferenciado da compreensão conceitual.

Em cada ciência pode-se encontrar uma teoria ingênua ou de senso comum, que pode ser um auxílio para a construção de teoria científica bem como uma fonte de hipóteses enganadoras que podem dificultar seu desenvolvimento por longo tempo, à semelhança do homem das neves que caminhou o dia inteiro, no gelo, em direção ao norte. Ao final do dia, exausto, quando conferiu sua posição descobriu que estava mais ao sul que pela manhã. Percebeu então ter caminhado o dia inteiro sobre um *iceberg* que começava a deslocar-se.

No entanto Baldwin (1980) afirma que existe um fundo comum, a partir do qual se desenvolve uma compreensão dos fenômenos naturais. A ciência desenvolve-se imperceptivelmente, a partir de respostas que os leigos encontram para as perguntas que lhes ocorrem, como exemplo cita o fato de uma criança, sem os conhecimentos da Física, ao desejar lançar uma bola o mais longe possível, atira-a para cima, ao invés de atirar em linha reta.

Contudo é necessário ter em mente uma perspectiva dos fatores mais importantes e de maior influência na formação do pensamento a respeito da natureza social do homem, que não sejam provenientes só do senso comum.

4.3 A importância das características domiciliares e do grupo familiar inserido no contexto da sociologia e psicologia social

São necessárias algumas considerações gerais sobre a influência do grupo familiar sobre o indivíduo, tendo por base alguns aspectos relevantes tanto sob a ótica da sociologia como da psicologia social.

A forma sociológica de estudar o homem ensina que para entender os seres humanos é necessário estudar os homens em seu ambiente, observar as forças exercidas sobre eles e suas condições sociais. As inter-relações entre as ações dos membros da sociedade revelam regularidades e tendências que podem ser estudadas por si mesmas.

A psicologia social afirma ser necessária a compreensão adequada do ambiente em que as pessoas agem, através de uma melhor compreensão dos fatos psicológicos, das maneiras pelas quais os seres humanos compreendem as condições e forças em que atuam.

A interação da sociologia com a psicologia social pode ser exemplificada na necessidade de descobrir o que significa ser um pai ou uma nora, após ter conhecido o esquema formal de relações de parentesco num grupo (Ash, 1977).

A própria abordagem da ciência social tem-se modificado no sentido de buscar explicações comportamentais aos fenômenos. Isto não significa que as ciências sociais estejam realizando investigações psicológicas e sim que o ponto de partida de todas as investigações a respeito dos processos sociais está em suposições definidas a respeito dos objetivos e capacidades dos indivíduos, devendo haver portanto uma concepção compreensiva da natureza humana.

Contudo, nos dados normalmente utilizados na modelagem tradicional de transportes, o comportamento do indivíduo é separado do comportamento do grupo e como consequência, informações importantes permanecem escondidas numa verdadeira caixa preta.

Segundo Chinoy (1975), a sociologia parte de dois princípios básicos: o comportamento dos seres humanos revela padrões regulares e repetitivos e os seres humanos são animais sociais e não criaturas isoladas. A estreita dependência do indivíduo em relação ao meio social possibilita a explicação de alguns aspectos do comportamento humano, sem que isto signifique que ele seja tão somente um instrumento passivo da sociedade.

O processo de sociabilização executa duas funções importantes: prepara o indivíduo para os papéis que há de desempenhar e transmite o conteúdo da cultura de uma geração para a outra. Segundo Chinoy (1975) o principal órgão nesses processos é geralmente, a família ou o grupo de parentesco em que convive.

A explicação mais aceita para a presença quase universal da família encontra-se na natureza da própria sociedade. Uma teoria amplamente divulgada centraliza esta explicação nas funções executadas pela família para a manutenção e continuidade da existência social organizada (Chinoy, 1975).

4.4 A importância do tema na etapa de geração de viagens

A família ou o grupo com o qual se convive é uma dimensão importante na organização do cotidiano, como polarizadora de grande parte da sociabilidade, bem como uma forma de enfrentamento das condições adversas de existência.

A importância do papel social do indivíduo, tanto no domicílio como na sociedade em geral, não é definida somente por sua participação na população

economicamente ativa mas também pela idade, pelo sexo, pelo estágio no ciclo da vida, pelo papel que desempenha no grupo com o qual convive, etc.

Cada indivíduo, num domicílio, a cada início de um dia, seleciona um conjunto de atividades derivado de suas necessidades biológicas e papéis sociais, limitações físicas e do meio ambiente em que vive.

Algumas atividades são obrigatórias, outras não e estão sujeitas a restrições de tempo, renda, custos, etc.

Esta abordagem permite definitivamente adicionar uma perspectiva social à forma tradicional de planejamento de transportes oferecendo uma análise mais completa do processo de decisão do indivíduo dentro do domicílio, levando-se em conta as interações entre os membros.

O uso de teorias de ciências sociais no planejamento de transportes é quase tão recente quanto as próprias teorias sociológicas. Somente nos últimos 15 anos tem-se reconhecido um amplo contexto social relacionado com vários aspectos da modelagem de transportes (Klausner, 1983).

A estrutura da família é considerada como o maior determinante do comportamento diário de seus membros definindo atividades específicas (Clarke *et al*, 1981; Hanson & Hanson, 1991). Por exemplo, numa família composta por um casal jovem sem filhos, provavelmente ambos trabalham em tempo integral e as atividades ligadas a compras são realizadas na hora do almoço ou no caminho de volta para casa.

A chegada do primeiro filho modifica o padrão de atividades do domicílio. Muitas vezes as viagens da mãe por motivo de trabalho são reduzidas,

aumentando as viagens por motivo de compras (trinta por cento das viagens de mães de filhos pequenos são do tipo casa - compras - casa, segundo Clarke *et al*, 1981).

À medida que os filhos crescem, participam de outras atividades levando as mães ou outros membros da família a desempenharem o papel de motoristas dos filhos.

Quando a família envelhece, não mais são necessários os complexos arranjos requeridos no atendimento de filhos pequenos e filhos em idade escolar, simplificando o padrão de atividades.

Supernak, Talvitie e De John (1983) cujo trabalho, com enfoque na modelagem individual, foi destacado anteriormente, concluíram que a estrutura familiar, do ponto de vista da composição da família, em nada contribui para explicar as variações no padrão de viagens do indivíduo. Esta conclusão foi baseada na não significância das diferenças de taxas médias de viagens entre os níveis de estrutura familiar utilizados. Concluíram também que os modelos domiciliares devem ser abandonados, uma vez que seu uso justificava-se pela relevância atribuída aos efeitos das inter-relações de seus membros, partindo do pressuposto que bastava usar o domicílio como unidade amostral para captar estes efeitos. Em outras palavras, os modelos individuais podem ocupar o lugar dos modelos domiciliares, pois não há mais necessidade de se estudar a estrutura do grupo à qual pertence o indivíduo.

Apesar destas conclusões percebe-se o incômodo provocado pela negação dos efeitos do grupo familiar, ao admitirem que não está claro como introduzir tais efeitos no modelo individual, nem mesmo no modelo domiciliar.

Concomitantemente, McDonald & Stopher (1983) tentavam fazer o mesmo, ou seja, incluir os efeitos das inter-relações do grupo familiar no modelo domiciliar.

Estes últimos autores verificaram que a estrutura familiar é tão significativa quanto as demais variáveis, ou seja, não tinham coeficiente de explicação maior. Concluem então que não há vantagens em seu uso nos modelos de geração de viagens, devendo-se abandonar a inclusão deste tipo de variável. Acrescentam ainda como fatos favoráveis a esta decisão, uma possível dificuldade de obtenção destes dados, aliado ao fato de a estrutura familiar do ponto de vista das relações de parentesco, não poder ser modificada por estratégias políticas de planejamento de transportes.

Entretanto acredita-se que uma parte significativa das variações no padrão de deslocamentos ocorre devido a variações dentro dos domicílios, a despeito de serem homogêneos quanto a algumas características gerais. O mesmo motivo que levou à rejeição dos modelos agregados em zonas de tráfego ou geográficas está levando a rejeição dos modelos desagregados domiciliares (Ortúzar e Willumsen, 1994).

Acredita-se que variações significantes no padrão de deslocamentos são provenientes das diferenças existentes entre os indivíduos que compõem estes domicílios, que com suas diversas estruturas e em diferentes estágios do ciclo da vida afetam seu papel dentro do grupo familiar.

Por conseguinte, a explicação para estas variações deve ser procurada no indivíduo inserido nos diversos contextos de agrupamento familiar.

Um certo conhecimento ‘ingênuo’ fortalece a busca de explicações nesta direção. Ou seja, pode-se imaginar facilmente, sem investigações científicas até então, que na situação hipotética abaixo, encontre-se algumas explicações para a diferença no padrão de deslocamentos dos indivíduos em questão.

Imagine dois domicílios, A e B, com as mesmas características gerais, medidas através do uso de variáveis usualmente empregadas na construção de modelos desagregados encontrados atualmente. Por exemplo, ambos os domicílios possuem: 3 componentes, dos quais 2 estão empregados, mesma renda mensal familiar e 2 automóveis. Suponha que o indivíduo A, seja o mesmo indivíduo colocado nestes dois contextos, domicílios A e B. Imagine que:

O domicílio A é composto por:

- uma mulher de 36 anos, de nível universitário, que assume o papel sócio-econômico de chefe de família e é um dos membros empregados.
- um homem de 18 anos, estudante universitário em período integral que na estrutura familiar é filho da chefe de família e
- indivíduo A: uma mulher de 33 anos, de nível secundário de instrução, o outro membro empregado em período integral, que na família assume o papel de irmã da chefe de família.

E que o domicílio B é composto por:

- um homem de 30 anos, de nível universitário, que assume o papel sócio-econômico de chefe de família e é um dos membros empregados.

- uma criança de 6 anos, que na estrutura familiar é filho do chefe de família e
- Indivíduo A: a mulher de 33 anos, de nível secundário de instrução, o outro membro empregado em período integral, que na família assume o papel de cônjuge e mãe.

Pode-se imaginar atividades muito diferentes desempenhadas pelo indivíduo A, para satisfazer os anseios e necessidades de ambos os domicílios, basta observar que no domicílio B deve haver atividades derivadas exclusivamente da presença da criança e portanto um padrão de deslocamentos característicos para cada um de seus membros.

Razões que vão além do senso comum, fundamentos lógicos, advindos de teorias que auxiliam na compreensão do homem foram encontradas na sociologia e em alguns fundamentos da psicologia social.

Modelos de escolha originados em teorias econômicas ou só psicológicas realmente não explicam o comportamento do indivíduo. São na verdade modelos usados para descrever o comportamento de um grupo homogêneo de pessoas (Fox, 1995).

Estas diferenças no padrão de atividades implicam em diferenças no padrão de deslocamentos. Um chefe de uma família nuclear simples (um casal sem filhos) tem um padrão diferente de atividades, portanto de deslocamentos, se comparados a chefes de famílias nucleares completas (casais com filhos), que por sua vez é diferente do padrão de um chefe de uma família quebrada (sem um dos cônjuges) ou de um chefe de um grupo convivente (sem relação de parentesco).

O ciclo de desenvolvimento das famílias, ou o ciclo vital, tem sido utilizado em muitos estudos em diversas áreas das ciências, sob distintos referenciais teóricos, através de índices específicos, adequados às questões que se propõe explicar (Kitamura, 1988; Kunert, 1994; Pas & Koppelman, 1987).

Além de incorporar a dimensão temporal, estabelece elos entre a dinâmica interna da família e o contexto social. Para alguns autores o conceito de estágio no ciclo da vida constitui um instrumento extremamente valioso do ponto de vista das conexões entre o tempo biológico ou individual e o tempo histórico ou social (Montali, 1990).

Em famílias identificadas num mesmo momento do ciclo vital existem variações nos patamares de renda, gerando especificidades nos arranjos familiares (ou do grupo convivente) nos quais estão inscritos os papéis socialmente atribuídos aos diversos membros. A partir do núcleo doméstico é delineado o horizonte de escolhas percebidas como possíveis e as práticas que informam sua estratégia de sobrevivência.

A cada estágio do ciclo vital predominam também projetos de vida específicos, resultando em práticas diversas no enfrentamento de condições objetivas. Segundo Montali (1990) não existe permutabilidade dos diferentes membros da família no que concerne a alguns papéis, por exemplo, o acesso ao emprego, na medida em que a posição no interior da família impõe pressões e limitações de diferentes maneiras.

Especificamente falando sobre a região metropolitana de São Paulo, sabe-se que o grau de participação da mulher no mercado de trabalho depende de vários fatores, dentre os quais a maternidade e a idade estão entre os mais

relevantes (Pesquisa de Emprego e Desemprego na Grande São Paulo - SEADE), mesmo levando-se em consideração o sensível aumento da presença da mulher no mercado de trabalho (SEADE, 1992a e b).

As possibilidades que a mulher tem de responder às demandas do mercado estão estreitamente condicionadas por fatores individuais e familiares. Em qualquer análise sobre o padrão de atividades da mulher, relacionado a trabalho, um aspecto crucial é o da sua posição na divisão social e sexual do trabalho, prioritariamente definida a partir de suas funções biológicas, o que a condiciona a um papel específico dentro da estrutura familiar.

O estado conjugal é uma das variáveis que mais afetam a participação da mulher em atividades econômicas de mercado. As taxas mais elevadas de atividade, encontram-se entre as mulheres separadas que não contam com o apoio financeiro de um cônjuge, desempenhando o papel de chefes de família (Bruschini, 1986).

Trinta por cento dos domicílios do mundo são chefiados por mulheres e em algumas regiões do mundo esta proporção é bem maior. Na região Metropolitana de São Paulo o SEADE (1992^a) estimou em 17,3%.

Existe também um grande número de famílias não nucleares (ou seja, não composta por casais, com ou sem filhos).

As mudanças na estrutura da família, por exemplo, o impacto do divórcio, alterou e tem alterado o tempo gasto com diversos papéis desempenhados dentro da estrutura familiar. Constatou-se que num período de 30 anos, os homens americanos experimentaram uma redução de 43% no tempo gasto no convívio com filhos pequenos (George, 1993).

Mudanças na estrutura familiar são vistas também como a primeira resposta para mudanças na condição econômica seguida por transições na vidas dos outros membros da família.

As mudanças na duração do tempo de desempenho de papéis dentro da estrutura familiar têm conseqüências sociais importantes (George, 1993) e têm conseqüências importantes também no padrão de deslocamentos que por sua vez precisa ser modelado adequadamente na etapa de geração de viagens urbanas dentro do contexto de análise de demanda por transportes de passageiros. Contudo tais variações estão sendo subestimadas pelos métodos tradicionais de análise.

4.5 Cuidados a serem tomados na obtenção de resultados quantitativos

Dentro do contexto específico de geração de viagens de passageiros, algumas condições devem ser satisfeitas para obtenção de dados adequados que possam fornecer resultados úteis e confiáveis e de maior abrangência que os fornecidos pelos métodos convencionais.

1) A escolha do indivíduo como unidade amostral básica

A desagregação em nível de indivíduos aumenta a possibilidade de transferência do modelo para outras regiões, uma vez que a maior parte da variação de comportamento está sendo explicada por fatores intrínsecos ao comportamento do indivíduo e não por fatores inerentes à zona de tráfego,

como nos modelos agregados ou às características do domicílio, como nos modelos desagregados existentes atualmente (Daly, 1982; Supernak, Talvitie & De John, 1983).

Este enfoque também propicia uma redução de custos em virtude da possibilidade de menores amostras fornecerem resultados satisfatórios.

2) Escolha criteriosa das variáveis a fim de que elas sejam realmente adequadas como representativas das necessidades e restrições do desempenho de atividades e portanto de deslocamentos.

Estas variáveis, importantes na descrição do comportamento, podem ser facilmente obtidas e têm papel fundamental na descrição de mudanças do padrão de deslocamentos como será visto posteriormente.

3) Estabelecimento de uma relação de causalidade entre as variáveis em estudo. Para tal, considerações teóricas devem ser feitas, fora do âmbito do conjunto de dados. Neste específico caso a base para estas considerações são provenientes da teoria sociológica do homem e da psicologia social.

4) Os métodos de análises devem ser quantitativos e essencialmente estatísticos. Tais métodos incorporam, entre outras ferramentas, os testes de significância, que auxiliam na confirmação, corroboração ou rejeição das relações de causa e efeito pré-concebidas.

No estudo apresentado nesta tese, optou-se pelo ajuste de um modelo linear através do uso de análise de regressão. Dentro deste contexto geralmente não faz sentido falar-se sobre o melhor conjunto de variáveis a ser incluído no modelo. Uma equação de regressão pode ser usada para diversos propósitos.

Um conjunto de variáveis pode ser o melhor para um e não para outro fim. Portanto não há um único conjunto. Este propósito deve estar em mente na hora da escolha do método de seleção de variáveis.

Podem haver vários subconjuntos adequados entre uma totalidade de variáveis envolvidas com o fenômeno em estudo. Estes subconjuntos trazem luz no sentido de entender a estrutura dos dados e do processo em questão (Chatterjee & Price, 1977; Shao, 1993; Weiss, 1995; Zheng & Loh, 1995).

A partir de então pode-se formular um modelo que se acredite estar realmente próximo dos mecanismos comportamentais reais, que expressem a relação causal estabelecida. Isto é essencial para afirmar que o comportamento é determinado na forma indicada pelo modelo e que assim será no futuro, havendo então reais possibilidades de efetuar-se previsões e interpretar-se adequadamente os coeficientes estimados.

Finalizando, em Baldwin (1980) encontram-se algumas considerações próprias para reflexão durante o processo de obtenção de dados quantitativos, em qualquer área da ciência, que estão reproduzidas parcialmente a seguir:

- Em todas as tentativas eruditas da compreensão do homem há uma dedicação à busca da verdade e uma disposição para aceitar os resultados, qualquer que seja o ponto a que leve.
- Quem busca a verdade, constantemente descobre como estava errado, cego e inconsciente. O fato de passar a vida toda descobrindo os erros anteriores e apesar disso, continuar a acreditar na acessibilidade da verdade, exige certa humildade e uma autêntica confiança.

- O esforço da pessoa para libertar-se de pressuposições e dominar seu desejo de um resultado específico deve acompanhá-la constantemente, desenvolvendo-se um mais alto senso de honestidade.
- Finalmente a sujeição de nossas crenças à revisão de autoridades exige um tipo de disciplina que não está muito longe da exigida para deixar que hipóteses prediletas venham a ser desmentidas pelos fatos.
- É impossível entender ou prever a totalidade do comportamento humano mas é necessário entender os mecanismos que o influenciam sob a ótica do problema em questão. O que é possível é a detecção de fatores e tendências gerais e significantes através de uma série de observações.

5. O ESTUDO DO CASO DA REGIÃO METROPOLITANA DE SÃO PAULO: ANÁLISES ESTATÍSTICAS DESCRITIVAS

Constatou-se que a pesquisa Origem/Destino da Cia do Metropolitano de São Paulo poderia fornecer grandes subsídios dentro do contexto desta análise em virtude da disponibilidade de algumas variáveis adequadas, bem como devido à sua abrangência e qualidade. A pesquisa abrange toda a RMSP e dela foram extraídos os dados para a estimação dos parâmetros de um modelo desagregado individual.

A abrangência e a criteriosa condução da pesquisa do Metrô merecem uma descrição para que se possa avaliar os resultados obtidos, inclusive à luz da origem dos dados.

5.1 Descrição sumária da pesquisa origem/destino - Metrô - SP - 1987

Sob o nome geral de Pesquisa Origem e Destino, a Cia do Metropolitano de São Paulo realiza quatro pesquisas que permitem levantar o padrão sócio-econômico, os locais de emprego e matrículas escolares, a matriz origem e destino das viagens internas e também externas à Região Metropolitana de São Paulo, que será citada neste trabalho por RMSP. As três primeiras são aplicadas aos domicílios e a última é aplicada aos passageiros dos veículos que cruzam os limites da RMSP.

Em 1987 a área de estudo abrangeu os 38 municípios da RMSP e coincidiu com a área de pesquisa. Subdividiu-se a área de estudo em 254 zonas identificadas pela homogeneidade de uso do solo. A divisão orientou-se ainda pela manutenção do número de zonas dentro dos limites operacionais adequados e pela necessidade de representar a estrutura urbana atual.

Os municípios que compõem a RMSP são: Arujá, Barueri, Biritiba Mirim, Caieiras, Carapicuíba, Cajamar, Cotia, Diadema, Embu, Embú-Guaçu, Ferraz de Vasconcellos, Francisco Morato, Franco da Rocha, Guararema, Guarulhos, Itapeverica da Serra, Itapevi, Itaquaquecetuba, Jandira, Juquitiba, Mairiporã, Mauá, Mogi das Cruzes, Osasco, Pirapora do Bom Jesus, Poá, Ribeirão Pires, Rio Grande da Serra, Santo André, Santa Isabel, Santana do Parnaíba, São Bernardo do Campo, São Caetano do Sul, São Paulo, Suzano, Salesópolis, Taboão da Serra e Vargem Grande Paulista.

Por razões operacionais da Cia Metropolitana de São Paulo, assegurando-se os limites aceitáveis de critérios de confiabilidade por eles estabelecidos, a pesquisa domiciliar foi aplicada em 204 zonas que possuíam número significativo de pessoas residentes. As zonas restantes estão divididas em dois blocos: 8 delas são grandes agrupamentos urbanos que não possuem residentes, porém com grande poder de atração de viagens; as outras 42 têm ocupação rarefeita e não foram pesquisadas.

5.1.1 O plano de amostragem da O/D 1987- Metrô

Após a análise de várias alternativas o grupo de pesquisa da Cia do Metropolitano de São Paulo verificou que os cadastros dos consumidores de

energia elétrica das empresas distribuidoras Eletropaulo: Eletricidade de São Paulo S.A. e CESP: Companhia Energética do Estado de São Paulo são suficientemente abrangentes e possuem informações que permitem avaliar o nível sócio-econômico de seus moradores segundo manual explicativo da O/D - 1987 - Metrô. Entre outras informações, estes cadastros possuem a localização do domicílio e o consumo de energia elétrica.

Estudos anteriores realizados na RMSP, por exemplo a pesquisa O/D - 1977, indicam que as principais variáveis consideradas: total de viagens por domicílio, habitantes por domicílio, renda familiar, entre outras, apresentam grande variabilidade, mesmo quando se consideram regiões geográficas não muito grandes, por exemplo as zonas O/D. Sendo assim, esquemas amostrais que levassem em conta os domicílios como unidades primárias de amostragem (amostragem casual simples) exigiriam tamanhos de amostra por zona O/D-87 incompatíveis com os recursos do projeto.

O esquema de estratificação de unidades compostas (um conjunto de domicílios) e a utilização da amostragem de conglomerados em dois estágios por estrato possibilitam atingir maior homogeneização das variáveis citadas dentro dos estratos e dos conglomerados destes estratos.

Para agrupar os domicílios em conglomerados foi utilizado um estudo de correlação entre renda e consumo de energia elétrica, desenvolvido pela FIPE - Fundação Instituto de Pesquisas Econômicas. Os cadastros das concessionárias estão separados em livros, que agrupam conjuntos de domicílios próximos que são utilizados como roteiros de leitura de consumo. Constatou-se que os moradores de domicílios constantes em um mesmo livro apresentam o mesmo nível sócio-econômico. Por conseguinte estes livros foram divididos em três

faixas de consumo médio de energia, que constituem os três estratos de unidades domiciliares. Um quarto estrato foi formado pelas favelas.

Os livros dentro de cada estrato foram definidos como conglomerados (conjunto de domicílios) sendo que no caso das favelas o conjunto de barracos constitui um conglomerado. O sorteio da amostra foi feito em duas etapas: inicialmente sortearam-se, dentro de cada estrato (determinada faixa de renda) de cada zona, alguns livros; a seguir, nos livros selecionados, sortearam-se os domicílios que compuseram amostra. No caso de favelas foi feito o sorteio dos barracos.

Foram contatados 26 mil domicílios cobrindo toda a RMSP correspondendo a aproximadamente 111 mil indivíduos.

Em pesquisas do porte da O/D-87 - Metrô - SP, alguns erros são inevitáveis durante o processo de coleta e transcrição de dados. Assim correções foram feitas para eliminar dados inconsistentes. Por exemplo, famílias com dois chefes foram eliminadas, uma vez que segundo os critérios da pesquisa, cada família admite somente um chefe.

O conjunto inicial de dados já corrigidos é constituído de 109.838 indivíduos, agrupados em 27.970 famílias que ocupam 25.845 domicílios da RMSP. Isto corresponde à 0,7125% da RMSP, considerando-se o CENSO IBGE - 1991 que atribui a RMSP um total de 15.416.416 habitantes.

5.1.2 Dados coletados pela O/D - 1987 - Metrô

Para a coleta de dados da pesquisa domiciliar, os domicílios foram visitados obedecendo a seqüência preestabelecida, a partir das zonas periféricas da RMSP para as zonas situadas no centro do Município de São Paulo, sendo realizadas até três visitas a cada domicílio.

Os trabalhos de coordenação de campo foram realizados por empresas privadas e o gerenciamento da pesquisa, por equipes da Companhia do Metropolitano de São Paulo e das empresas participantes.

O questionário aplicado foi dividido em três partes:

- **Primeira parte:**

Informações relativas ao domicílio: tipo do domicílio, condição de moradia, valor do aluguel/prestação.

Informações relativas à família: tempo de residência no bairro, tamanho da família, itens de conforto familiar (automóvel, televisão, geladeira, etc.).

Informações relativas às características sócio-econômicas de cada pessoa da família: posição ocupada dentro da estrutura familiar, idade, sexo, nível de instrução, ocupação profissional, classe de atividade da empresa em que trabalha, faixa de renda mensal.

- Segunda parte:

Informações sobre o local de trabalho e/ou estudo: endereço do trabalho e/ou estudo.

- Terceira parte:

Informações relativas às viagens realizadas no dia anterior à entrevista pelas pessoas do domicílio: origem, destino, motivo, modo e duração.

O período de aplicação compreendeu os meses de outubro, novembro e os dez primeiros dias de dezembro de 1987, quando foram aplicados cerca de 80% dos questionários e março, abril, maio e junho de 1988, quando foram aplicados os 20% questionários restantes.

Convém ressaltar o auto padrão de análise e critérios utilizados na pesquisa O/D da Cia do Metropolitano de São Paulo. Detalhes como caracterização do respondente qualificado (o entrevistado), definição do que é um domicílio, conceito de família, moradores a serem computados como fazendo parte do domicílio, caracterização das relações de ligação familiar entre os membros, caracterização do que é uma viagem, etc. podem ser encontrados em detalhes no manual da O/D - 1987 da Cia do Metropolitano de São Paulo.

5.1.3 Características gerais dos dados coletados

- **Ano Base**

Com a publicação parcial dos dados do censo IBGE - 1991 o Departamento de Planejamento de Transporte da Cia. do Metropolitano recalculou as estimativas para número de viagens e número de habitantes, mas manteve o ano base de 1987, quando fornece dados já tabulados.

Para o caso dos dados utilizados nesta tese, considerou-se os dados coletados, portando o ajuste pelo censo 91 não foi considerado.

- **Contagem do número de viagens**

O conceito do que é uma viagem está diretamente ligado ou ao destino ou à origem. Por exemplo, o trajeto residência-trabalho-residência será computado como sendo uma viagem com destino a trabalho por dia e uma viagem com destino a residência, portanto duas viagens. As viagens consideradas podem ter tanto a origem como o destino não sendo o domicílio, ou seja, não se trata somente de viagens baseadas no domicílio.

5.2 Escolha das variáveis a serem analisadas

Optou-se pela coleta das seguintes variáveis:

- sexo
- idade

- renda individual (salind) e renda familiar (salfam): ajustados para setembro de 1987, considerados em números de salários mínimos.
- nível de instrução: inst
- número de pessoas da família: npesfam
- situação ocupada dentro da estrutura familiar: sitfam
- número de viagens diárias realizadas por pessoa do domicílio: viagpes. Considerou-se o número de viagens realizadas no dia anterior ao da entrevista.

- Para a situação ocupada dentro da estrutura familiar, seguiram-se as definições estabelecidas pela equipe de planejamento da O/D - 1987:

chefe de família: é o principal arrimo de família. Se este só passar os fins de semana em casa, ele não é considerado morador e a esposa será considerada a chefe da família. Se os moradores do domicílio não tiverem nenhum grau de parentesco, o chefe é aquele que assim o foi considerado pelos demais. Só há um chefe por família.

cônjuge: o termo é usado para a pessoa (de ambos os sexos) que seja legalmente casada ou viva em união consensual com o chefe.

filho(a): além dos filhos do casal são registrados nesta situação os filhos de criação ou os tutelados.

outras posições ocupadas: outras categorias definidas pelo Metrô: parentes, agregados, empregados residentes no domicílio e visitantes.

5.2.1 Estrutura familiar e índice do ciclo de vida das famílias: ICV

Duas outras variáveis foram geradas a partir das originais: o tipo de estrutura familiar (clasfam) e o índice de ciclo de vida da família (ICV, referido na análise de dados como icvclas).

Para o tipo de estrutura da família optou-se pela classificação da Fundação SEADE - Sistema Estadual de Análises de Dados, empregada na PCV- Pesquisa de Condições de Vida, realizada entre Junho a Agosto de 1990 em aproximadamente 5500 domicílios (Seade, 1992a e 1992b).

As famílias são classificadas como:

- 1- Nuclear simples: casal sem filhos.
- 2- Nuclear completa: casal com filhos.
- 3- Nuclear ampliada: casal com filhos mais outros membros, sejam parentes, visitantes, empregados residentes, etc.
- 4- Família quebrada: família com filhos caracterizada pela ausência de um dos cônjuges.
- 5- Grupo convivente: agrupamento de pessoas sem relação de parentesco com o chefe
- 6- Grupo familiar: agrupamento de pessoas com relação de parentesco com o chefe.
- 7- Unipessoal.

Para a análise desenvolvida neste trabalho juntaram-se as classificações 5 e 6 sob o título de grupo convivente.

Como o O/D-87 não dispõe diretamente do tipo de estrutura das famílias e sim dos dados dos moradores do domicílio divididos por famílias dentro do domicílio, foi necessário identificar a família de cada indivíduo e criar um código para sua posição ocupada dentro da família e um código para a própria família. O código da família foi determinado como sendo a soma dos códigos de seus membros, de tal forma que permita identificar o tipo de família da seguinte forma:

o chefe tem código.....	100.000,
cônjuge tem código.....	10.000,
filho tem código.....	100,
os demais agregados tem código.....	1.

Se uma família tem código 110209 (soma dos códigos de seus membros) significa que ela é nuclear completa ampliada, pois tem 1 chefe, um cônjuge, dois filhos e nove membros agregados quer sejam parentes, visitantes, etc.

O índice de ciclo de vida, ICV, é calculado a partir das idades dos membros da família nuclear completa. Ele permite diferenciar famílias em fases diferentes do ciclo da vida, que não seriam diferenciadas pela classificação acima. É o caso, por exemplo, de casais jovens com filhos adolescentes e casais maduros com filhos pequenos, onde ambas as famílias são nucleares completas.

Na medida em que cada membro da família solidariza-se com os demais modificam-se as suas possibilidades de inserção na sociedade. O indicador de ciclo de vida familiar auxilia na identificação do potencial de participação das famílias na sociedade, fator intimamente ligado a idade e ao papel do indivíduo dentro da estrutura familiar. Este índice foi desenvolvido por Bussab

& Wagner (1994). Os autores apresentam ainda outros índices baseados na classificação das famílias a partir da idade do chefe ou do cônjuge ou dos filhos, mas que não permitem uma classificação mais adequada de famílias, que sem representarem casos anômalos, não seguem os padrões temporais expressos nas faixas etárias.

Com a finalidade de garantir uma classificação mais adequada destas famílias os autores apresentam o seguinte índice:

$$ICV = \text{Log}[(IDMC * IDMF * DCNFV) / (\text{media}(IDMC) * \text{media}(IDMF) * \text{media}(DCNFV))]$$

onde:

IDMC = idade média dos cônjuges

IDMF = idade média dos filhos

DCNFV = diferença de idade entre o cônjuge mais novo e o filho mais velho

O denominador tem o efeito operacional de diminuir a magnitude dos números, que quando em torno da média oscilam próximos de 1 e quando muito inferiores (superiores) a média são inferiores (superiores) a 1.

O logarítmo foi adotado para obter-se redução de escala e para a introdução de valores positivos e negativos, referencial que facilita a identificação da posição relativa da família frente as demais.

Para evitar distorções decorrentes da inclusão de casos anômalos no estudo, Bussab e Wagner (1994) sugerem a introdução de filtros para a seleção das famílias, considerando-se apenas aquelas onde:

- a diferença entre a idade do chefe e a do cônjuge esteja entre - 20 (20 negativo) e 25 anos.
- a idade média dos filhos esteja entre 1 e 50 anos.
- a diferença entre a idade do cônjuge mais novo e a do filho mais velho esteja entre 14 e 50 anos.
- a diferença entre a idade do filho mais velho e a do filho mais novo esteja entre 0 e 25 anos.

Das 27.970 famílias disponíveis nos dados, 428 não passaram no filtro.

O conjunto de dados final, após a atribuição do ICV para todas as famílias é constituído de: 107.789 indivíduos de 27.542 famílias ocupando 25.440 domicílios (0,7% da RMSP).

Os pontos de corte do indicador de ciclo de vida foram baseados na técnica de análise de agrupamento (*cluster analysis*) com a finalidade de gerar-se grupos homogêneos de famílias, cuja adequabilidade dos resultados foi verificada através de análise discriminante, tendo como referencial discriminatório as idades dos membros da família e na fase final o próprio ICV.

Os pontos de corte propostos pelo SEADE são:

Família jovem: é aquela com $ICV < -0,395$.

Família adulta: é aquela com $-0,395 \leq ICV < 0,840$.

Família velha: é aquela com $ICV \geq 0,840$.

Exemplo:

Uma família composta por um chefe com 35 anos, o cônjuge com 35 anos e um filho com 10 anos tem $icv = 0,047$, portanto é considerada uma família adulta. Uma família com um chefe com 22 anos, o cônjuge com 18 anos e um filho com um ano tem $icv = -1,4971$ portanto é uma família jovem.

O ICV como definido anteriormente é aplicável somente para famílias nucleares completas, ou seja, casais com filhos, quer as famílias sejam ampliadas ou não.

Contudo o SEADE, através de Bussab & Wagner (1994) ampliou este conceito para outras estruturas de família. O procedimento está baseado na atribuição do valor do índice de ciclo de vida calculado nas famílias nucleares completas para as demais famílias que apresentem semelhança na composição das idades do chefe ou dos cônjuges.

Considere como exemplo uma família nuclear simples, ou seja um casal sem filhos, cuja idade média dos cônjuges seja 25 anos.

Considere todas as famílias nucleares completas, cuja idade média entre os cônjuges seja também 25 anos. Calcula-se o icv médio destas famílias e atribui-se este valor à família nuclear simples de mesma idade média entre os cônjuges.

Para valores de idade média do casal da família nuclear simples que não haja correspondente nas famílias nucleares completas, toma-se o valor do icv médio anterior mais próximo, ou seja, toma-se o valor atribuído a menor e mais

próxima idade média entre cônjuges que tenha representação nas famílias nucleares completas.

Após esta etapa, realiza-se uma regressão monotônica com a finalidade de deixar o icv monotonicamente crescente com a idade média do casal.

Para os demais tipos de família, onde não há cônjuges e sim um chefe, o procedimento é análogo, tendo por base a idade do chefe.

5.3 Análises estatísticas preliminares dos dados da Região Metropolitana de São Paulo

Muitas tabulações foram feitas. Há cerca de 2.000 páginas com análises estatísticas detalhadas para cada nível das variáveis e cruzamentos entre elas. A finalidade principal foi detectar influências nas variações do número de viagens individual e em segundo lugar descrever minuciosamente a composição dos indivíduos da amostra disponível.

Tal magnitude de trabalho teve dois incentivos essenciais. Em primeiro lugar, a qualidade e critérios rigorosos adotados na elaboração da pesquisa O/D-87 da Cia. do Metropolitano de São Paulo - SP. Estes fatos convidam e encorajam qualquer empenho em obter resultados pois, se conduzidos de forma adequada, certamente terão alta credibilidade. O segundo incentivo foi a disponibilidade de um bom, senão o melhor, sistema de análise estatística de dados o SAS: Statistical Analysis System, devidamente documentado (SAS, 1989, 1990, 1991) e instalado no Laboratório Avançado de Informática do Departamento de Estatística Matemática Aplicada e Computacional do Instituto de

Geociências e Ciências Exatas da UNESP - Rio Claro - SP, em meados de 1996.

Os dados foram analisados em três etapas distintas.

Na primeira delas dispunha-se de dados agrupados. Para cada nível das variáveis, por exemplo, faixas de renda, faixas de idade, sexo, etc, dispunha-se do número de habitantes e do número de viagens totais destes habitantes. A partir de então calcula-se a mobilidade média para aquele específico cruzamento fazendo-se: número de viagens total dividido pelo número de habitantes daquele grupo.

Diversos modelos de geração de viagens foram ajustados e minuciosamente analisados através da técnica de regressão linear múltipla, utilizando-se mínimos quadrados ponderados pelo número de habitantes de cada grupo e variáveis dummy, que descrevem os diversos níveis das variáveis qualitativas e das variáveis quantitativas agrupadas em faixas, como por exemplo faixas de renda.

Muitas conclusões interessantes e relevantes para atingir-se os objetivos deste trabalho puderam ser obtidas.

Contudo verificou-se que dados individualizados, ao invés de dados agrupados, trariam subsídios e recursos muito mais abrangentes que os anteriores. Em muito superariam as conclusões e a obtenção de alguns resultados já fortemente evidenciados nas primeiras análises.

Abandonou-se portanto, quase dois anos de trabalho com o primeiro conjunto de dados, reiniciando-se as análises, desde as mais simples, com os dados

individualizados, na íntegra, fornecidos pela equipe de planejamento da pesquisa Origem/Destino - 1987 - Metrô - SP, no mês de janeiro de 1997.

A segunda etapa de análises concentrou-se na correção do arquivo completo de dados e na atribuição do índice de ciclo de vida da família de cada indivíduo, dado não disponível na pesquisa O/D-87, mas passível de ser calculado. Em seguida, classificou-se as famílias de cada indivíduo segundo sua estrutura familiar, conforme descrito anteriormente.

Análises gerais foram feitas e retirou-se 10 amostras aleatórias de cerca de 50 domicílios cada uma. Estas amostras eram compostas, em média por uns 250 indivíduos.

Análises estatísticas foram realizadas com estas amostras, diversos modelos de geração de viagens foram ajustados, utilizando-se a técnica de regressão linear, não mais usando mínimos quadrados ponderados, uma vez que os dados não eram agrupados.

A terceira etapa concentrou-se na análise de 30 amostras aleatórias de mesmo tamanho médio, cerca de 265 indivíduos cada, só que com uma diferença básica das anteriores, foram selecionados aleatoriamente indivíduos no lugar de domicílios. No primeiro caso, todos os membros do domicílio, necessariamente, faziam parte da amostra.

Em nenhuma das três etapas, obteve-se resultados contraditórios. Pelo contrário, a cada etapa, cresciam as evidências favoráveis à inclusão da influência das características do domicílio sobre o padrão de viagens do indivíduo.

Não é viável e nem de interesse ou de bom senso expor a totalidade das conclusões e tabulações obtidas.

Optou-se por descrever detalhadamente os principais resultados que foram obtidos nas duas últimas etapas de análises.

A primeira parte de resultados apresentados levou em consideração a totalidade dos dados obtidos e nas análises finais, com a finalidade de obter-se modelos de geração de viagens, considerou-se as 30 amostras aleatórias de indivíduos (e não de domicílios).

5.3.1 Resultados da análise estatística descritiva dos dados individualizados

É aconselhável que uma análise estatística comece com técnicas exploratórias de dados, adquirindo-se sensibilidade e informações adicionais sobre a variável estudada.

Com base na análise descritiva dos dados tiraram-se várias conclusões. Entre outras, verificou-se haver um padrão de deslocamentos diferente para cada nível das variáveis explicativas, o que direcionou a tomada de decisões com relação à técnica de análise a ser usada, bem como dos procedimentos relacionados à formulação do modelo individual de geração de viagens a ser ajustado. Os principais resultados estão apresentados a seguir.

5.3.1.1 Correlação de Pearson

A correlação de Pearson (ρ) (Draper & Smith, 1981) é um número entre -1 e 1, que quando próximo de 1, em módulo, indica forte relação linear entre as variáveis. A Tabela 1, abaixo, mostra os valores obtidos com os 109.838 indivíduos pesquisados.

Tabela 1 - Correlação de Pearson calculada com 109.838 indivíduos da O/D-87, Metrô - SP*

viagpes	viagpes	sexo	papel na família	inst	estrutura da família	índice de ciclo de vida	idade	salário individual	número de membros da família
sexo	-0,1442	-							
papel na família	-0,1379	0,1985	-						
inst	0,3331	-0,0196	-0,2014	-					
estrutura da famíl	0,0073	0,0725	0,1898	0,0343	-				
índ ciclo de vida	0,0632	0,0140	0,0117	0,0622	0,0250	-			
idade	0,0097	0,0343	-0,4873	0,2273	0,0833	0,3444	-		
salário individual	0,2654	-0,2227	-0,3166	0,4536	0,0172	0,0207	0,2380	-	
n membros fam	-0,0332	-0,0212	0,2918	-0,1711	0,0292	0,0996	-0,2720	-0,1134	-
salário familiar	0,1779	-0,0029	0,0785	0,3249	0,0182	0,0865	0,0456	0,4845	0,0818

* Todos os valores são altamente significantes, excetuando-se a correlação entre sexo e salário familiar.

A significância estatística dos valores das correlações de Pearson dizem respeito ao teste para $H_0: \rho = 0$ para $n = 109838$ indivíduos.

Os valores mais próximos de +1 ou de -1 estão destacados a seguir:

- -0,4873: é a correlação entre a idade individual e o papel que o indivíduo ocupa dentro da estrutura familiar (idade e sitfam).

- +0,4845: é a correlação entre o salário individual e o salário familiar (salind e salfam).
- +0,4536: é a correlação entre o salário individual e o nível de instrução do indivíduo (salind e inst).
- +0,3444: é a correlação entre a idade do indivíduo e o índice de ciclo da família (idade e icvclas).
- +0,3331: é a correlação entre o número de viagens do indivíduo e a instrução (viagpes e inst).
- +0,3249: é a correlação entre o salário familiar e o nível de instrução do indivíduo (salfam e inst).
- -0,3166: é a correlação entre o salário individual e o papel que o indivíduo ocupa dentro da estrutura familiar (salind e sitfam).

Convém observar também que as variáveis mais correlacionadas com o número individual de viagens são instrução (0,3331); salário individual (0,2654) e salário familiar (0,1779).

5.3.1.2 Classificação dos indivíduos segundo as variáveis de interesse: distribuição de freqüências

É conveniente ter em mente a composição dos indivíduos distribuídos segundo os níveis das variáveis explicativas pesquisadas: sexo, idade, nível de

instrução, salário individual, salário familiar, índice de ciclo da vida familiar (ICV) e tipo de estrutura da família.

Na atribuição do índice de ciclo de vida, foram eliminadas 428 famílias entre as 27.970 disponíveis. Estas famílias representam casos anômalos, segundo critérios descritos anteriormente. Com isto as distribuições de frequências a seguir estão baseadas em 27.542 famílias, totalizando 107.789 indivíduos que ocupam 25.440 domicílios (0,7% do número de habitantes da Região Metropolitana de São Paulo).

A variável número de viagens do indivíduo merece destaque especial e será a última a ser apresentada num item separado. Isto se deve ao fato de ser conveniente primeiro ter em mente um retrato da composição dos habitantes da RMSP, para depois avaliar-se o número de viagens individual já à luz da composição da população.

- SEXO

A distribuição da população pesquisada quanto ao sexo segue o padrão conhecido para o Brasil: 52,1% dos habitantes são do sexo feminino e 47,9% são do sexo masculino, segundo as estimativas obtidas entre os indivíduos pesquisados.

- NÍVEL DE INSTRUÇÃO

Tabela 2 - Nível de instrução- distribuição de frequências

INST	Freq.	Porcentagem	Freq. Acum.	Porc. Acum.
< 4ª série	49537	46,0	49537	46,0
4ª ≤ inst < 8ª	28893	26,8	78430	72,8
8ª ≤ inst < 2ºg	12376	11,5	90806	84,2
2ºg ≤ inst < sup	10638	9,9	101444	94,1
≥ superior	6345	5,9	107789	100,0

72,8% dos indivíduos pesquisados não possuem nem o 1º grau completo.
Somente 11,5% tem o 1º grau completo.

9,9% tem o 2º grau completo.

5,9% possuem o 3º grau completo.

- PAPEL OCUPADO DENTRO DA ESTRUTURA FAMILIAR

Convém lembrar como a Cia. do Metropolitano de São Paulo caracterizou o papel do indivíduo dentro da estrutura familiar. Seguiu-se as definições estabelecidas pela equipe de planejamento da O/D - 1987 descritas no item 5.2.

Tabela 3 - Papel ocupado dentro da estrutura familiar-distribuição de frequências

SITFAM	Freq.	Porcentagem	Freq. Acum.	Porc. Acum.
chefe	27542	25,6	27542	25,6
cônjuge	21180	19,6	48722	45,2
filho	47471	44,0	96193	89,2
parente	9366	8,7	105559	97,9
agregado	1019	0,9	106578	98,9
empreg.	1008	0,9	107586	99,8
visit.	203	0,2	107789	100,0

Observa-se que entre os indivíduos pesquisados:

25,6% são chefes de família

19,6% ocupam o papel de cônjuges

44,0% ocupam o papel de filhos

Ao considerar-se só os chefes de família, verificou-se que 16,87% são do sexo feminino, ou seja, 16,87% das famílias pesquisadas são chefiadas por mulheres. A estimativa do SEADE (1992) para a RMSP era de 17,3%. Entre os chefes do sexo feminino:

10,4% tem menos que 30 anos. Esta porcentagem dobra para os chefes do sexo masculino, 20,1%.

53,6% tem 50 anos ou mais, enquanto que para o sexo masculino este número passa a 71,6%.

- **ESTRUTURA DA FAMÍLIA**

Seguiu-se o critério do SEADE para classificar as famílias segundo descrição feita no item 5.2.1.

Tabela 4 - Estrutura familiar- distribuição de frequências

CLASFAM	Freq.	Porcentagem	Freq. Acum.	Porc. Acum.
nuclear simp.	3097	11,2	3097	11,2
nuclear comp.	13770	50,0	16867	61,2
nuc. comp. amp.	4313	15,7	21180	76,9
quebrada	3209	11,7	24389	88,6
grupo	1559	5,7	25948	94,2
unipes.	1594	5,8	27542	100,0

Entre as 27.542 famílias usadas neste trabalho, observou-se que:

84,2% tem até 5 membros sendo que 0,4% tem 10 ou mais membros.

11,2% são casais sem filhos, nucleares simples.

65,7% são famílias nucleares completas, ampliadas ou não.

11,7% são famílias quebradas.

5,8% são unipessoais.

5,7% são grupos conviventes.

Este padrão está dentro do esperado para a RMSP segundo estimativas do SEADE (1992a).

- SALÁRIO INDIVIDUAL

Entre os indivíduos pesquisados, verificou-se que:

55,4% não tem renda individual.

82,6% ganham até 5 salários mínimos, incluindo os que não têm renda.

7,1% ganham mais que 10 salários mínimos.

- SALÁRIO FAMILIAR

7,3% não tem renda familiar.

23,3% tem renda familiar inferior a 3 salários mínimos, incluindo os que não têm renda.

65,6% tem renda familiar no máximo igual a 10 salários mínimos.

14,2% tem renda familiar maior que 20 salários mínimos.

- IDADE

A idade foi considerada pela O/D-87, Metrô - SP, em anos completos. Então crianças menores que 1 ano, são computadas como zero anos. Verificou-se junto à equipe de planejamento do Metrô - SP que um deslocamento de uma pessoa com uma criança de colo leva à computação de viagens para ambos separadamente, decorre daí deslocamentos para idades muito baixas.

A Tabela 5 abaixo mostra a composição etária dos indivíduos pesquisados.

Tabela 5 - Distribuição de idade em anos completos

IDADE	Porcentagem	Porcentagem Acumulada
até 6	15,1	15,1
7 até 14	16,0	31,1
15 até 35	37,0	68,1
36 até 65	27,3	95,4
mais que 65	4,6	100,0

- **ÍNDICE DE CICLO DE VIDA DAS FAMÍLIAS**

31,6% das famílias pesquisadas são consideradas jovens e somente 3,3% são consideradas velhas.

Comparando-se as Tabelas 6 e 7 abaixo verifica-se que este padrão não segue as estimativas do SEADE (1992a) para a RMSP.

Tabela 6 - Índice de ciclo de vida (ICV) para os dados coletados na O/D-87 - Metrô - SP

ICV	Freq.	Porcentagem	Freq. Acum.	Porc. Acum.
fam. jov.	8700	31,6	8700	31,6
fam. adul.	17933	65,1	26633	96,7
fam. velh.	909	3,3	27542	100,0

Tabela 7 - Índice de ciclo de vida (ICV) segundo dados coletados pelo SEADE na Pesquisa de Condições de Vida de 1990

FAMÍLIAS	Porcentagem	Porcentagem Acumulada
jovens	36,1	36,1
adultas	36,1	72,2
velhas	27,8	100,0

Conforme já comentado anteriormente, o ICV foi criado para famílias nucleares completas, ampliadas ou não. O ICV é ampliado para as demais famílias, baseando-se nos valores do ICV médio das famílias nucleares completas com mesma idade média de chefe ou cônjuges que a família para a qual deseja-se atribuir o ICV.

Após este procedimento, uma regressão monotônica é processada com a finalidade de tornar o icv monotonicamente crescente com a idade média do casal. Este procedimento não foi aplicado para a classificação das famílias dos dados pesquisados na O/D-87.

Para verificar possíveis classificações inadequadas derivadas da ausência do crescimento monotônico do ICV com a idade, alguns cuidados foram tomados.

Construiu-se um gráfico, Gráfico 1 abaixo, para verificar as oscilações no crescimento do ICV médio das famílias nucleares completas de acordo com a idade média do casal.

Gráfico 1 - ICV médio por idade média do casal (IDMC) das famílias nucleares completas

Levando-se em consideração que as famílias são classificadas de acordo com a divisão do icv em 3 grandes faixas, somente as famílias no limiar das faixas poderiam ter sua classificação alterada, o que não altera o retrato do quadro geral das famílias pesquisadas.

Para tornar isto mais evidente comparou-se as características gerais de cada tipo de família, jovem, adulta e velha obtidas com os dados da O/D-87 com as características gerais apresentadas para a Região Metropolitana de São Paulo por Bussab & Wagner (1994) e referenciadas na Tabela 8 como resultados do SEADE.

A Tabela 8, abaixo descreve as idades médias do chefe, cônjuge, filhos e do casal para cada classe de família: jovem, adulta e velha.

Tabela 8 - Idade média dos chefes, cônjuges e filhos das famílias nucleares completas da RMSP, segundo as etapas do ciclo da vida familiar

Etapas do ciclo da vida familiar	Idade Média							
	Casal		Chefe		Cônjuge		Filhos	
	O/D-87	SEADE	O/D-87	SEADE	O/D-87	SEADE	O/D-87	SEADE
jovem	28,8	30,6	30,56	32,2	27,05	29,1	3,82	4,5
adulta	44,72	42,7	46,63	44,4	42,81	41,0	14,0	13,3
velha	72,69	58,6	74,11	60,5	71,27	56,8	37,85	25,5

Observa-se uma grande semelhança entre as duas estimativas, indicando a não existência de classificações inadequadas neste trabalho.

A diferença principal apontada entre o padrão das famílias segundo o ICV nas duas estimativas apresentadas nas Tabelas 7 e 8 anteriores mostram que, segundo o SEADE 27,8% das famílias da RMSP são velhas enquanto nos dados, da mesma RMSP, extraídos da O/D-87 este número é de 3,3%. É fato aceito na literatura e reafirmado pelo SEADE que as famílias estão envelhecendo.

Ao se considerar somente as famílias nucleares completas, onde todos os procedimentos indicados por Bussab & Wagner (1994) foram realizados, esta diferença se mantém, nas mesmas proporções.

A conclusão que se tira a partir destes fatos é que esta diferença não é derivada da ausência da regressão monotônica e sim das características inerentes às famílias pesquisadas.

Esta técnica de regressão não é usual, para processá-la um programa deveria ser especificamente desenvolvido. Julgou-se ser maior o prejuízo de não incluir o ICV na classificação das famílias, em virtude das considerações já levantadas sobre seu significado na interação entre os membros do domicílio.

5.4 Número de viagens diárias individuais - análise estatística descritiva

A distribuição de frequências da variável número de viagens individual é apresentada na Tabela 9, a seguir.

Tabela 9 - Distribuição de freqüências do número de viagens diárias individuais

Nº de viagens (individuais)	Freq.	Porcent.	Freq. Acum.	Porc. Acum.
0	34388	34,7	34388	34,7
1	647	0,7	35035	35,4
2	46395	46,8	81430	82,2
3	4400	4,4	85830	86,6
4	9117	9,2	94947	95,8
5	1433	1,4	96380	97,3
6	1567	1,6	97947	98,9
7	411	0,4	98358	99,3
8	380	0,4	98738	99,7
9	131	0,1	98869	99,8
10	83	0,1	98952	99,9
11	39	0,0	98991	99,9
12	25	0,0	99016	100,0
13	16	0,0	99032	100,0
14	15	0,0	99047	100,0
15	4	0,0	99051	100,0
16	2	0,0	99053	100,0
17	3	0,0	99056	100,0
18	1	0,0	99057	100,0
19	1	0,0	99058	100,0
20	1	0,0	99059	100,0
21	1	0,0	99060	100,0
24	1	0,0	99061	100,0

Cabe ressaltar alguns valores:

82,2% realizam até 2 viagens. Convém lembrar que as viagens são contadas de acordo com o seguinte exemplo: um trajeto do tipo residência-trabalho-residência é contado como sendo 2 viagens, uma para o trabalho e outra com destino à residência.

4,1% realizam 4 ou mais viagens diárias.

34,7% não realizaram viagens. Devido à representatividade deste número, algumas tabulações gerais foram feitas para melhor caracterizá-los.

5.4.1 Caracterização dos indivíduos com número de viagens diárias igual a zero

Do total de indivíduos pesquisados, 4268 não responderam ao número de viagens realizadas. 34,7% responderam não ter feito nenhuma viagem. Entre estes, observou-se que:

65,5% são do sexo feminino.

64,4% não completam nem até a 4ª série do primeiro grau.

1,9% possuem nível superior de instrução.

11,4% são chefes de família.

29,2% são cônjuges de famílias.

44,3% são filhos, dentro da estrutura familiar.

79,7% não tem rendimentos.

10,1% ganham mais que 2 (dois) salários mínimos ou seja 89,9% ganham menos que 2 salários mínimos. Entretanto 30% dos indivíduos que não viajaram tem renda familiar inferior a 4 salários mínimos e 30% superior a 10 salários mínimos.

60,5% são de famílias adultas, segundo a classificação do índice de ciclo de vida familiar.

A distribuição etária também merece destaque:

33,3% tem até 6 anos de idade.

7,3% tem de 7 a 14 anos.

26,0% tem de 15 a 35 anos.

26,7% tem de 36 a 65 anos.

6,7% tem mais que 65 anos.

5.4.2 Número médio de viagens diárias individuais segundo os níveis das variáveis de interesse

Algumas tabulações foram realizadas com o intuito de analisar o número médio de viagens diárias do indivíduo de acordo com algumas classificações das variáveis de interesse: idade, sexo, renda, índice de ciclo de vida familiar, estrutura familiar e instrução.

- Indivíduos de famílias velhas são os que menos viajam, realizando em média 0,97 viagens, seguidos de indivíduos de famílias jovens, com 1,44. Os membros de famílias adultas são os que mais viajam, realizam em média 1,77 viagens.
- A média de viagens de indivíduos cresce, segundo a estrutura familiar na seguinte ordem:

indivíduo de família nuclear simples (1,54); nuclear completa ampliada (1,61); de grupo convivente (1,62); de família nuclear completa (1,68); de família quebrada (1,69) e unipessoal (1,86).

- De acordo com o papel desempenhado dentro da estrutura familiar, o número médio de viagens individuais cresce de empregado residente (0,59); para visitante (0,62); parente (1,26); cônjuge (1,34); membro agregado (1,63); filho (1,65) até o chefe (2,11).
- O indivíduo do sexo masculino tem média de viagens 1,90 enquanto que do sexo feminino tem 1,43.

- O número médio de viagens cresce de acordo com o salário individual e a renda familiar conforme mostra o gráfico 2.

É interessante notar que o salário individual estima mais precisamente o número médio de viagens que o salário familiar, basta observar a grande oscilação do número de viagens para cada valor do salário familiar.

Gráfico 2 - Número médio de viagens diárias individuais por renda pontual

Nº médio de viagens diárias
individuais

Nº médio de viagens diárias
individuais

Estas oscilações ficam suavizadas quando se utiliza faixas de renda, conforme mostra o Gráfico 3 a seguir.

Gráfico 3 - Número médio de viagens diárias individuais por faixas de renda

A variação do número médio de viagens, segundo a composição etária, está representada nos Gráficos 4 e 5 abaixo.

Grafico 4 - Número médio de viagens diárias individuais por faixas de idade

Grafico 5 - Número médio de viagens diárias individuais por idade pontual

- É nitidamente crescente o número médio de viagens individual de acordo com o aumento do nível de instrução.

Considerou-se 5 níveis para instrução descritos na Tabela 10 abaixo.

Tabela 10 - Número médio de viagens diárias individuais por nível de instrução

Nível de instrução	Nº médio de viagens diárias individuais
até 4ª série do 1º grau	1,14
4ª série completa até 1º grau incompleto	1,77
1º grau completo até 2º grau incompleto	2,19
2º grau completo até superior completo	2,45
superior completo	2,82

5.4.3 Análise da distribuição do número de viagens diárias individuais segundo alguns cruzamentos dos níveis das variáveis domiciliares

A análise ora apresentada está baseada num recurso de apresentação de dados estatísticos, muito útil e de fácil aplicação, quando se deseja comparar a evolução da distribuição estatística de uma variável segundo outras variáveis qualitativas. Neste caso, deseja-se estudar a variação no padrão do número de viagens do indivíduo, segundo as variáveis de família: estrutura da família, índice de ciclo de vida da família e papel ocupado pelo indivíduo dentro da

estrutura familiar. O recurso ora citado é referenciado na literatura como *box plot*.

5.4.3.1 Descrição do *box plot*

O *box plot* apresentado abaixo, é um diagrama na forma de uma caixa onde estão descritos os principais elementos de uma análise descritiva de dados estatísticos (SAS/STAT, 1989; SEADE, 1992b).

Figura 1 - *Box Plot*

Para utilizar o *box plot*, imagina-se o eixo real Y onde se coloca as laterais superior e inferior da caixa localizadas respectivamente no 3º e 1º quartis da distribuição: Q_3 e Q_1 . A linha horizontal central corresponde à mediana ou 2º quartil (Q_2), enquanto que o sinal de $+$ representa a média.

As linhas verticais pontilhadas estendem-se até o maior e o menor valor da variável, desde que estes não ultrapassem o primeiro quartil (Q_1) menos uma vez e meia o intervalo interquartil (IIQ) ou terceiro quartil (Q_3) mais uma vez e meia o intervalo interquartil (IIQ).

O intervalo interquartil (IIQ) é dado pela diferença entre o terceiro quartil e o primeiro quartil ($Q_3 - Q_1$).

Para distribuições simétricas a linha horizontal, que representa a mediana, coincide com a média, sinal de $+$ e divide a caixa em duas partes iguais, cada uma contendo 25% das observações analisadas. Logo a caixa contém 50% das observações analisadas numa faixa de valores da variável de tamanho expresso pelo intervalo interquartil. Para distribuições assimétricas, pode ou não aparecer a divisão da caixa ao meio se a mediana coincidir com o 1º ou 3º quartil.

Os valores mais extensos que $Q_3 + 1,5\text{IIQ}$ (terceiro quartil mais um e meio intervalo interquartil) ou $Q_1 - 1,5\text{IIQ}$ (primeiro quartil menos um e meio intervalo interquartil), porém dentro do intervalo $[Q_1 - 3\text{IIQ}; Q_3 + 3\text{IIQ}]$ são identificados com o número zero e são considerados discrepantes. Os pontos fora desta faixa são considerados muito discrepantes e assinalados por um asterisco (*).

Convém apresentar alguns *box plot*, para o número de viagens diárias individuais. Eles estão agrupados segundo os critérios abaixo:

Bloco A - Comparação entre os chefes de famílias de diversas estruturas familiares.

Bloco B - Comparação entre os chefes de famílias em diversos estágios do ciclo da vida familiar, para cada tipo de estrutura familiar.

Bloco C - Comparação entre os chefes em relação aos diversos estágios do ciclo da vida familiar.

Bloco D - Comparação entre os cônjuges em relação aos diversos estágios do ciclo da vida familiar.

Bloco E - Comparação entre os filhos em relação aos diversos estágios do ciclo da vida familiar.

Bloco F - Comparação entre os diversos estágios do ciclo da vida de acordo com faixas de salário individual.

Bloco G - Comparação entre os diversos estágios do ciclo da vida de acordo com o nível de instrução do indivíduo.

Seria enfadonho comentar em detalhe cada um destes grupos, mas convém fazer algumas observações gerais e deixar para o leitor análises mais minuciosas, observando os *box plot's* no Anexo B. Reproduz-se aqui o Bloco A e parte do Bloco B (B1).

Bloco A

chefes

Observa-se a variação da distribuição do número de viagens, como um todo, para chefes de acordo com as diversas estruturas familiares, destacando-se como de maior mobilidade os chefes de famílias compostas por casais com filhos e agregados (famílias nucleares completas e ampliadas).

Bloco B-1

chefes

Nº de viagens individual

Nesta série de *box plot* observa-se a variação proveniente da diferença entre os estágios no ciclo da vida das famílias de chefes de famílias nucleares simples (casais sem filhos): chefes de famílias jovens tem maior mobilidade que chefes de famílias adultas ou velhas.

No Bloco A (p.127 e Anexo B) observa-se nítida diferença do padrão de viagens individual, visto através do deslocamento de toda a distribuição da variável número de viagens, para os chefes de diversas estruturas familiares. Observe que o primeiro quartil coincide com o terceiro quartil para famílias nucleares completas (não ampliadas) e que há um aumento significativo na mobilidade de chefes de famílias nucleares completas ampliadas.

No Bloco B (p. 128 e Anexo B) fica bem evidenciada a diferenciação provocada pelo estágio do ciclo da vida familiar para cada tipo de estrutura familiar.

No Bloco C (Anexo B) mais uma vez, observa-se que em famílias jovens os chefes têm maior mobilidade.

No Bloco D (Anexo B) reafirma-se a baixa mobilidade para membros de famílias velhas.

O Bloco E (Anexo B) mostra maior mobilidade dos filhos de famílias adultas. Os blocos C, D e E, juntos, revelam a diferença de mobilidade quando se leva em conta o papel que o indivíduo ocupa dentro da estrutura familiar.

O Bloco F (Anexo B) mostra diferenças de viagens para indivíduos dentro de cada estágio do ciclo da vida, ao se levar em conta faixas de renda específicas. Observa-se um aumento geral da mobilidade de cada grupo conforme aumenta a renda, mesmo nas famílias velhas.

No Bloco G (Anexo B) nota-se algo semelhante ao Bloco F, só que para diversas faixas de nível de instrução, ao invés de renda. Note o aumento significativo da mobilidade de indivíduos de famílias velhas.

6. VERIFICAÇÃO ESTATÍSTICA DA TESE E ANÁLISE DE SUA VIABILIDADE

Espera-se que bons resultados possam ser obtidos através de técnicas usuais de análise partindo-se de variáveis que realmente estejam detectando as diferenças de comportamento e de deslocamentos.

Optou-se pelo ajuste de um modelo linear através do uso de análise de regressão. Convém apresentar uma série de considerações sobre o modelo teórico, o processo de estimação e a correta interpretação dos resultados obtidos.

Este destaque justifica-se em virtude do amplo uso desta técnica e o desconhecimento de grande parte dos usuários, de alguns aspectos mais específicos, que quando negligenciados podem invalidar os resultados obtidos.

Também serve de alerta para que os usuários aprofundem seus conhecimentos e não se deixem enganar pela simplicidade de aplicação e interpretação dos resultados, uma vez que sua consolidação como ferramental estatístico de modelagem só é válida quando se faz um uso correto e adequado.

6.1 Regressão linear múltipla: notas metodológicas

Como regra geral, não se deve aplicar uma técnica estatística sem antes verificar se as suposições do modelo estão razoavelmente satisfeitas. Siqueira (1983) oferece um estudo detalhado com alguns exemplos das causas,

detecção e consequências dos desvios das suposições do modelo linear geral, que engloba o modelo de regressão linear.

A técnica de análise de regressão é largamente utilizada e freqüentemente é mal usada. Há vários abusos que devem ser mencionados, alguns estão descritos a seguir e podem ser examinados com mais detalhes em Montgomery & Peck (1983).

Os modelos de regressão devem ser usados como equações de interpolação e não extrapolação, ou seja, dentro do campo de variação de valores das variáveis explicativas que foram usados para a estimação dos parâmetros do modelo.

A presença de valores discrepantes, conhecidos na literatura como *outliers* ou pontos influentes, pode comprometer seriamente as estimativas obtidas através da técnica de mínimos quadrados. Podem causar heterocedasticidade e não aditividade do modelo de regressão (Siqueira, 1983).

A criteriosa análise destes pontos levam à decisão de eliminá-los da amostra quando são considerados pontos derivados de erros ou outros problemas indesejados na coleta e transcrição de dados. Mas por ser um ponto influente não necessariamente deve ser eliminado, em alguns casos estes pontos trazem evidências do comportamento do fenômeno estudado, até então desconhecidas.

O fato de obter-se um coeficiente de determinação muito alto numa dada equação de regressão, não significa que as variáveis são desta forma relacionadas, a não ser que haja um conhecimento prévio do fenômeno que torne isto viável. Montgomery & Peck (1983) mostram um coeficiente de determinação (r^2) igual a 0,9842 para um modelo que explica o número de

peessoas com deficiências mentais através da variável explicativa número de licenças para recebimento de ondas de rádio. Um outro exemplo é a obtenção de um coeficiente de determinação de 0,87 quando se muda a variável explicativa do exemplo anterior para o número de cartas recebidas no ano corrente por pessoas com o nome igual ao primeiro nome do presidente americano.

A escolha do modelo sem intercepto (o termo constante) não pode ser baseada na comparação dos valores dos coeficientes de determinação (r^2) dos modelos com e sem intercepto, pois estes valores são não comparáveis (Searle, 1971; Uyar & Erdem, 1990; Kvalseth, 1985; Hawkins, 1980). Normalmente a confusão instala-se quando o conjunto de dados disponível está muito longe da origem e o comportamento das relações entre as variáveis é muito diferente daquele próximo à origem.

Outro engano comum é o uso indiscriminado que se faz da regressão linear sem verificar a adequabilidade das suposições básicas.

Pode-se resumir em poucas palavras efeitos devastadores da violação das hipóteses do modelo de regressão:

- Em geral um desvio de uma suposição altera o nível de significância dos testes de hipóteses sobre os parâmetros de interesse, geralmente aumentando-o (Siqueira, 1983).
- Os testes F-Snedecor e t-Student só podem ser utilizados corretamente sob a hipótese de normalidade dos erros, caso contrário a distribuição da estatística do teste não é mais F-Snedecor ou t-Student. Devido a forte

desvio da normalidade dos dados também ficam prejudicados os intervalos de confiança e previsão por dependerem das estatísticas F e t.

- As propriedades estatísticas dos estimadores de mínimos quadrados não são válidas, ou seja, os estimadores tornam-se viesados e não têm variâncias mínimas.

Siqueira (1983) oferece um estudo mais detalhado das conseqüências da violação das suposições para o modelo linear geral.

Alguns procedimentos que devem ser tomados ao utilizar-se a análise de regressão linear não são de conhecimento geral, portanto decidiu-se incluir no Anexo A algumas definições e conceitos básicos para evitar-se problemas como os descritos acima.

A leitura do Anexo A pode esclarecer dúvidas quanto as causas, detecção e efeitos dos desvios das suposições do modelo, da presença de multicolinearidade e de *outliers*, bem como de critérios para a escolha de modelos com ou sem intercepto.

- **As variáveis explicativas**

Algumas das variáveis disponíveis para o ajuste do modelo são qualitativas: sexo, estrutura familiar, nível de instrução, estágio no ciclo da vida familiar e papel ocupado dentro da família. Elas aparecem no modelo linear como variáveis *dummy*, uma vez que o comportamento da mobilidade depende do

nível da variável em questão podendo ser necessário um modelo não linear caso este procedimento não fosse adotado.

Além destas utilizou-se as seguintes variáveis: nº membros da família, idade em anos completos e renda individual e familiar em número de salários mínimos.

- **A escolha do modelo de regressão linear sem intercepto**

Com base em algumas considerações detalhadamente descritas no Anexo A, seção A-5, optou-se pelo modelo linear sem intercepto.

- **A verificação das hipóteses e diagnóstico em regressão**

A verificação das hipóteses usuais do modelo de análise de regressão, bem como a detecção de outliers e multicolinearidade foram baseadas em algumas estatísticas descritas no Anexo A: os pontos de corte destas estatísticas não são definitivos para avaliar se uma observação deve ser eliminada. São indicativos de que uma análise daquela observação é necessária para, só então, tomar-se a decisão de eliminá-la ou não.

Detecção de *outliers*

Através da análise de 6 estatísticas: diagonais da matriz H , *hat matrix*, D de Cook, exame dos resíduos *studentizados* e cálculo de: *covratio*, *dffits*, *dfbetas*.

Deteção de multicolinearidade

Através de medidas usuais: exame dos autovalores e autovetores bem como a decomposição espectral de $X'X$, número da condição espectral, fator de inflação de variância.

A verificação das hipóteses

Verificação das hipóteses usuais do modelo de análise de regressão é feita mediante criteriosa análise de resíduos.

O SAS, Statistical Analysis System permite o cálculo e a impressão destas estatísticas. Realizou-se estes procedimentos para os dados agrupados da 1ª etapa de análises e para uma amostra aleatória de 232 indivíduos da etapa final.

Todas estas análises mostraram-se satisfatórias sem comprometimento de nenhuma das hipóteses do modelo, seja normalidade, independência ou homocedasticidade dos erros e sem a presença de *outliers* ou multicolinearidade que possa comprometer as propriedades estatísticas dos estimadores de mínimos quadrados.

6.1.1 Métodos de seleção de variáveis e testes da contribuição de uma ou mais variáveis

Métodos de seleção de variáveis, num modelo de regressão linear, são amplamente discutidos na literatura estatística. O método mais utilizado é o

stepwise ou passo a passo que muitos autores recomendam. Uma breve discussão e apresentação de outros métodos pode ser encontrada em Draper & Smith (1981) entre outros.

Antes de procurar algum método de seleção de variáveis, os procedimentos usuais de diagnóstico em regressão devem ser considerados buscando-se eliminar observações discrepantes e detectar a presença de forte grau de multicolinearidade, pois poderiam invalidar os resultados do método escolhido.

Convém fazer uma descrição sumária dos passos do método *stepwise*, uma vez que será utilizado na análise dos dados da O/D-87:

1 - Escolhe-se a variável mais correlacionada com a variável dependente Y; neste caso, nº de viagens individual.

2 - Ajusta-se a regressão de Y sobre esta variável.

3 - Calcula-se r: o coeficiente de determinação parcial entre Y e todas as demais variáveis (menos a escolhida no passo 1).

$$r_{Y,X_1} = \frac{SQReg(Y|X_1 \text{ e } X_2) - SQReg(Y|X_2)}{SQRes(Y|X_2)} = \frac{\text{Contribuicao de } X_1}{SQRes(Y|X_2)}$$

onde:

$SQReg(Y|X_i)$ = soma de quadrados da regressão de Y levando em consideração X_i como variável explicativa.

$SQRes(Y|X_i)$ = soma de quadrados do resíduo da regressão de Y sobre X_i .

Escolhe-se a de maior r para entrar no modelo.

4 - Testa-se a contribuição de todas as variáveis que estão no modelo no passo 2 para verificar se a inclusão de uma variável não implica na exclusão de outra que já pertença ao modelo. Repete-se o passo 2 para incluir as variáveis que ainda não foram testadas.

5 - O processo termina quando nenhuma outra variável fora do modelo tem contribuição significativa e todas as que já estão são significantes; ou quando a variável que deve entrar no modelo, por ter maior coeficiente de correlação parcial é a mesma que deve sair. Note que uma variável que entra no modelo não permanece necessariamente.

O teste da contribuição da cada variável ao modelo, está baseado na estatística F-Snedecor.

Prova-se que, sob a hipótese $H_0: \beta_1 = 0$; β_1 do modelo de regressão de Y sobre X_1 e X_2 tem-se:

$$\frac{\text{Contribuicao de } X_1 \text{ ao modelo}}{\frac{SQRes(Y|X_1 \text{ e } X_2)}{n - (k + 1)}} = \frac{SQReg(Y|X_1 \text{ e } X_2) - SQReg(Y|X_2)}{\frac{SQRes(Y|X_1 \text{ e } X_2)}{n - (k + 1)}} \sim F_1; n - (k + 1)$$

onde k = número de variáveis explicativas do modelo maior. Neste caso k = 2.

Cabe observar que sempre $SQReg(Y|X_1 \text{ e } X_2)$ será maior que a $SQReg(Y|X_1)$ ou que a $SQReg(Y|X_2)$. Ao levar-se em conta mais variáveis a $SQRes$ diminui

necessariamente e a SQT (soma de quadrados total) mantém-se igual para as três regressões, Y sobre X_1 , Y sobre X_2 e Y sobre X_1 e X_2 . Ela não depende de quem sejam as variáveis explicativas (veja Anexo A). Convém lembrar que $SQT = SQReg + SQRes$. Sendo assim o teste da contribuição de cada variável, testa se o aumento na soma de quadrados da regressão, obtido com a inclusão de mais uma variável, é realmente significativo para o modelo e não somente uma consequência matemática dos cálculos.

O teste é análogo a considerar-se no lugar da contribuição de X_1 ao modelo, a diferença obtida entre as somas de quadrado dos resíduos de ambos os modelos, com ou sem a variável de interesse, fazendo-se:

$$\frac{\frac{SQRes(Y|X_1) - SQRes(Y|X_1 \text{ e } X_2)}{SQRes(Y|X_1 \text{ e } X_2)}}{n - (k + 1)} \sim F_{k-1; n - (k + 1)}$$

Este teste pode ser estendido para testar-se a contribuição de um grupo de variáveis explicativas. Basta especificar adequadamente a estatística F. Por exemplo, para testar-se a contribuição de X_3 , X_4 e X_5 testa-se:

$H_0: \beta_3 = \beta_4 = \beta_5 = 0$ fazendo-se:

$$\frac{\frac{[SQRes(Y|X_1, X_2) - SQRes(Y|X_1, X_2, X_3, X_4, X_5)] / (k - g)}{SQRes(Y|X_1, X_2, X_3, X_4, X_5)}}{n - k + 1} \sim F_{k-g; n - (k + 1)}$$

ou

$$\frac{[SQRes(\text{do modelo reduzido}) - SQRes(\text{do modelo completo})] / (k - g)}{QMRes(\text{do modelo completo})} \sim F_{k-g; n - (k + 1)}$$

onde:

QMRes = quadrado médio residual

g = n° de variáveis do grupo a ser testado

Este teste é encontrado nos textos de inferência estatística sob um título mais geral: testes de hipóteses lineares gerais sobre os parâmetros. Tais testes estão disponíveis nos softwares, ou sistemas de análises estatísticas, mais completos.

Convém ressaltar que no método *stepwise* o teste é feito individualmente para cada variável, a cada passo do programa.

Normalmente os programas imprimem a estatística t-Student apresentada para cada variável explicativa num modelo de regressão. Esta estatística é equivalente à estatística F do teste da contribuição de cada variável separadamente, $F_{1; n-(k+1)} = t_{n-(k+1)}^2$.

É interessante deixar claro algumas diferenças entre o teste conjunto da contribuição de um grupo de variáveis e o teste isolado para cada variável.

Suponha que se queira testar a contribuição do grupo de variáveis X_1 , X_2 e X_3 . O teste apropriado é para $H_0: \beta_1 = \beta_2 = \beta_3 = 0$, ou seja, uma hipótese só que engloba os três parâmetros.

Suponha agora que se realizou 3 testes separados, um para cada parâmetro: $H_i: \beta_i = 0$; $i = 1, 2, 3$ cada qual com um nível de significância α . Quais as consequências?

Quando se testa conjuntamente ou globalmente $H_0: \beta_1 = \beta_2 = \beta_3 = 0$, o teste é conjunto para os três parâmetros e é realizado através de testes apropriados para hipóteses estatísticas multivariadas. Tem-se o controle sobre a fixação do valor do nível de significância α . O que não ocorre numa afirmação global resultante de 3 testes separados H_1, H_2 e H_3 cada qual com nível de significância α .

Os resultados dos 3 testes não permitem concluir globalmente que a probabilidade de rejeitá-los, caso sejam verdadeiros, seja α . O nível de significância conjunto ou global é bem maior que cada α .

Dizer que as 3 hipóteses, H_1, H_2 e H_3 , são verdadeiras ao mesmo tempo, trata da intersecção dos 3 eventos sendo o nível de significância novo, no mínimo igual a $\alpha_1 + \alpha_2 + \alpha_3$ (Wichern & Johnson, 1992). Isto está baseado no seguinte:

$$P(\text{todas as } H_i \text{ serem verdadeiras}) = 1 - P(\text{pelo menos uma } H_i \text{ ser falsa}) \geq$$

$$1 - P(\text{união dos eventos: } H_i \text{ é falsa}) = 1 - \sum_{i=1}^3 P(H_i \text{ ser falsa}) =$$

$$1 - \sum_{i=1}^3 (1 - P(H_i \text{ ser verdadeira})) = 1 - (\alpha_1 + \alpha_2 + \alpha_3)$$

onde $P(.)$ = probabilidade

Uma alternativa é escolher $\alpha_i = \frac{\alpha^*}{3}$ onde α^* é o nível de significância global desejado e o denominador é o número de testes isolados que se fará para obter um resultado global. Um exemplo deste procedimento são os critérios adotados para a construção de intervalos de confiança de Bon Ferroni (Wichern & Johnson, 1992).

Contudo, estando disponível o teste multivariado, melhores conclusões podem ser obtidas. São testes do tipo:

$$H_0: \beta_1 = \beta_2 = \beta_3 = 0$$

H_a : pelo menos uma combinação linear dos β_i ; $i = 1, 2, 3$ é diferente de zero.

A não aceitação de H_0 leva a inferir que há no mínimo uma combinação linear dos parâmetros que seja diferente de zero e que é responsável pela rejeição de H_0 .

6.2 Significância estatística da contribuição ao modelo das variáveis ligadas à influência das características domiciliares sobre o indivíduo

Conforme diretrizes já apresentadas, a técnica de mínimos quadrados para estimar os parâmetros de um modelo de regressão linear oferece resultados confiáveis após a verificação e tratamento adequado dos dados com a finalidade de eliminar dados discrepantes ou a existência de forte multicolinearidade.

Com esta finalidade e baseando-se nas análises descritivas já apresentadas, optou-se por eliminar o salário familiar e a idade do indivíduo. Algumas considerações que nortearam esta decisão foram:

- O salário familiar é altamente correlacionado com o salário individual (correlação de Pearson = 0,4845).
- A oscilação do número de viagens individual para um salário familiar fixo é muito maior que para o salário individual fixado, conforme mostra o Gráfico 2, indicando que o salário individual é mais preciso para estimar o número de viagens.
- Ao manter-se o salário familiar no modelo, aparecem autovalores de $X'X$ muito próximos de zero, o que é indesejável.
- A idade é altamente correlacionada com o papel que o indivíduo ocupa na família (correlação de Pearson = -0,4873) e com o índice de ciclo de vida familiar (correlação de Pearson = 0,3444) e ainda com o nível de instrução (correlação de Pearson = 0,2273). O conteúdo de informação das demais variáveis é mais interessante do que a da idade por si só.

Embora ainda permaneceram algumas variáveis com correlação alta, como por exemplo o salário individual com o nível de instrução (0,4536), os métodos de detecção de multicolinearidade não apontaram ser necessárias novas exclusões, obtendo-se um grau de multicolinearidade considerado não prejudicial para a condução da análise.

Mesmo que nas análises estatísticas descritivas já esteja evidente a variação no padrão de viagens do indivíduo, é necessário estabelecer-se um critério estatístico mais elaborado que possa mostrar a significância da inclusão de variáveis domiciliares no modelo de geração de viagens individual. Testes foram conduzidos segundo os critérios discutidos na seção anterior. Estão baseados na estatística F do teste da contribuição do grupo de variáveis: papel ocupado pelo indivíduo dentro da estrutura familiar, tipo de estrutura do domicílio, nº de membros do domicílio e índice de ciclo de vida familiar.

A estatística F, da contribuição do grupo de variáveis domiciliares, é altamente significativa, ou seja, encontrou-se uma oscilação para $F_{\text{observado}}$ de 1,96 a 4,33. Todos significantes para $\alpha = 3\%$ e em alguns casos $\alpha = 0,01\%$ (0,0001) para amostras de 222 indivíduos em média. O valor de F passa para 1001,39 ao utilizar-se todos os dados disponíveis, ou seja, 99.061 indivíduos, que responderam a renda e o número de viagens diárias além das demais variáveis inseridas no modelo. Isto significa que a contribuição deste grupo de variáveis é altamente significativa para o modelo de análise de regressão desenvolvido.

Partindo deste pressuposto e para dirimir dúvidas quanto um possível acaso, retiraram-se outras 29 amostras aleatórias de tamanho médio igual a 265 indivíduo, entre os 107.789 indivíduos disponíveis inicialmente.

Para cada uma destas 30 amostras, eliminaram-se as variáveis salário familiar e idade conforme já discutido anteriormente. A partir de então selecionaram-se as variáveis de um modelo de regressão pelo método *stepwise* 30 vezes, ou seja, 30 modelos foram ajustados. Os parâmetros estimados bem como as principais estatísticas estão descritos na Tabela 11 e 13.

O número de vezes em que cada variável permaneceu no modelo está descrito abaixo:

Sexo: 18 vezes, ou seja, em 18 das 30 amostras, o sexo permanece no modelo ao final do procedimento *stepwise*.

Nº de pessoas da família: 17.

Salário individual: 26.

Índice de ciclo de vida da família: 21.

Papel ocupado pelo indivíduo dentro da estrutura familiar: 20.

Nível de instrução: 30.

Tipo de estrutura da família: 25.

Quando uma variável do grupo familiar não estava presente no modelo, outra deste mesmo grupo estava. De maneira que todos os modelos resultantes incluíram alguns representantes das variáveis de família.

Outro procedimento conduzido foi estimar o número total de viagens dos 107.789 indivíduos da amostra completa, tendo por base os modelos resultantes com as 30 amostras de tamanho médio 265.

60 estimativas foram feitas, uma para cada modelo estimado:

- 30 estimativas foram feitas tendo por base 30 modelos resultantes do método *stepwise* partindo do conjunto completo de variáveis, incluindo portanto as variáveis de família no modelo final (Tabela 11 e 13).
- 30 estimativas foram feitas tendo por base 30 modelos resultantes do método *stepwise* tendo como ponto de partida o conjunto de variáveis completo, menos as variáveis de família: número de membros do domicílio, papel ocupado dentro da estrutura familiar, tipo de estrutura familiar e índice de ciclo de vida familiar (Tabela 12 e 13).

Tabela 11 - Estimação dos parâmetros dos modelos incluindo as variáveis domiciliares⁴

obs	amostra	dumysex	salind	dinst2	dinst3	dinst4	dinst5	npesfam	dsitf2	dsitf3	dsitf4	dfam2	dfam3	dfam4	dfam5	dfam6	dicv2	dicv3	rsquare	F-Snedecor	n
1	ind1	-	0,0569	0,9169	1,2557	0,6618	1,4470	-	-	0,4131	-0,7067	-	0,6578	-	1,2330	-	0,6320	-	0,6451	40,34	232
2	ind2	-	0,0480	0,8262	0,8763	-	-	-	-	-	-	0,7374	0,8995	1,0368	-	-	0,5508	-	0,6427	55,50	223
3	ind3	-	0,0455	0,6445	1,2043	1,7757	1,2481	-	-	0,6270	-	-	1,0573	-	1,0220	-	0,7362	-	0,6037	41,47	254
4	ind4	-	0,0187	0,3341	0,8696	-	0,9469	-	-	-	-	0,9188	1,0935	0,6405	1,8915	-	0,4421	1,3464	0,5950	34,67	246
5	ind5	0,3061	0,0576	0,6342	1,0971	-	1,6369	0,1575	-	-	-0,9250	-	-	-	0,8143	-	0,2937	-	0,6512	53,52	267
6	ind6	-	0,0598	0,5409	0,9207	0,9413	-	-	-	0,6024	-	0,4737	0,7322	-	1,4262	1,8175	0,3425	-	0,5960	38,06	268
7	ind7	-	0,0548	0,9724	1,2849	2,6286	1,6090	-	-	0,5341	-	0,5845	1,0224	-	-	-	-	-	0,5926	48,18	273
8	ind8	-	0,0670	0,3999	1,0360	-	-	-	-	-	-0,9564	1,1608	1,4598	0,9948	2,5935	1,0553	0,3548	1,0586	0,6249	39,84	274
9	ind9	-	0,0361	0,6734	1,1973	1,3622	1,7545	0,0588	-	0,7256	-	-	0,3438	-	-	1,5861	0,5202	-	0,6336	44,44	267
10	ind10	0,6436	0,0420	0,9935	1,0374	1,0856	1,7487	0,0600	-	-	-	0,4432	-	-	-	-	-	-	0,6081	47,71	254
11	ind11	0,2598	0,0727	1,0138	1,6579	1,4525	1,8144	0,0701	-	0,7584	-	-0,4125	-	-	-	0,9761	-	-	0,7078	62,02	266
12	ind12	-	0,0758	0,5003	1,0879	0,5094	-	0,0818	-	0,3518	-	0,4331	-	-	0,7742	-	0,3505	1,2359	0,6209	40,29	256
13	ind13	0,3544	0,1215	0,3887	-	0,4035	-	0,1015	-	0,3840	-	-	-	-0,4100	-	-	0,4392	-	0,6736	65,78	263
14	ind14	0,5269	0,0338	0,6440	1,0990	3,0250	1,6813	0,0877	-	0,4525	-	-	-	-	-	-	-	-	0,6173	56,46	288
15	ind15	0,6567	0,0628	0,5318	1,2164	0,6424	0,9473	0,1379	-	-	-0,6173	-	-	0,9324	-	-	-	-	0,5235	31,62	268
16	ind16	0,5857	-	0,7985	1,0556	1,9308	1,0802	0,0716	-	-	-	-	-	-	-	1,6519	0,4273	-	0,5864	45,01	262
17	ind17	0,7714	0,0367	0,8412	1,6526	1,4806	1,9644	0,0636	-	-	-0,8506	-	-	-	0,9260	-	0,4306	-	0,6009	42,46	292
18	ind18	0,3589	0,0483	0,8469	1,2568	1,0952	0,9722	-	-	0,4200	0,7885	0,3667	-	-	-	1,3895	0,3796	-	0,6219	39,63	276
19	ind19	0,6532	0,0253	1,2438	0,7846	1,0690	1,4293	0,1180	-	-	-	-	-	-	-	2,0086	-	-	0,5402	36,27	255
20	ind20	0,7211	-	0,9425	1,3289	1,0066	2,2279	-	-	-	-	0,5557	-	-	-	-	0,4256	-	0,5721	47,74	257
21	ind21	0,5107	0,0520	0,8561	1,2739	1,9303	1,6922	-	0,4459	0,4924	-	-	-	-	-	-	0,4469	-	0,6104	45,25	269
22	ind22	-	0,0776	0,6272	1,4285	1,2907	2,1652	0,0598	-	0,4063	-	-	-	-	-	-	0,5238	-	0,6540	59,53	260
23	ind23	-	0,0789	1,0774	1,5310	1,0205	0,8331	-	-0,8707	-	-	0,7348	0,4996	-	1,2466	-	0,4774	-	0,6852	55,95	267
24	ind24	-	0,0350	0,6394	1,7129	1,0961	0,8689	-	-	-	-	1,1960	1,2157	1,0069	-	-	-	-	0,6134	48,20	251
25	ind25	0,2928	0,0608	0,7129	1,1620	1,3088	1,8854	0,0635	-	0,4447	-	-	-	-	-	-	0,4946	-	0,6146	50,67	295
26	ind26	0,5945	0,0679	0,4112	-	0,6448	1,0326	0,0609	0,4849	0,5394	-	-	-	-	-	-	0,3887	-	0,6336	54,57	293
27	ind27	0,4900	-	0,9100	1,4665	2,1340	2,1675	-	-	0,3865	-	-	-	-	0,8366	-	0,6863	-	0,6876	74,82	280
28	ind28	0,3904	-	0,9176	1,5505	2,4514	2,5268	0,0774	-	-	-	0,4805	-	-	-	-	-	-	0,5619	48,55	272
29	ind29	0,6214	0,0437	0,8598	1,3086	1,2090	1,4821	0,0987	-	-	-	-	0,4056	-	-	-	-	-	0,5480	38,49	262
30	ind30	0,5086	0,0746	0,4512	1,0976	0,8430	-	-0,1208	-	-	-	1,2240	1,6264	1,3285	-	-	0,3353	-	0,5586	32,65	268

⁴ Veja descrição das variáveis na Tabela 13

Tabela 12 - Estimação dos parâmetros dos modelos excluindo as variáveis domiciliares⁵

Obs	Amostra	DUMYSEX	SALIND	DINST2	DINST3	DINST4	DINST5	RSQUARE	F-Snedecor	N
1	ind1	0,6302	0,0429	1,4425	1,8172	1,1598	1,9737	0,5733	50,60	232
2	ind2	0,7120	-	1,6247	1,9793	1,1962	1,7261	0,5470	52,65	223
3	ind3	0,9844	-	1,3310	1,9252	2,3958	2,3612	0,5067	51,15	254
4	ind4	0,5640	-	1,3679	1,8847	1,4606	2,2180	0,4804	44,56	246
5	ind5	0,9286	0,0419	1,2315	1,9285	0,8077	2,3944	0,5772	59,39	267
6	ind6	0,7861	0,0423	1,1372	1,5232	1,5650	0,9347	0,5019	44,00	268
7	ind7	0,8159	0,0352	1,3340	1,8119	3,0241	1,9269	0,5349	51,17	273
8	ind8	0,8360	0,0556	1,4451	2,0205	1,4283	1,0679	0,5262	49,61	274
9	ind9	0,9104	-	1,3491	1,9198	2,0798	2,7811	0,5108	54,72	267
10	ind10	1,0334	0,0384	1,3381	1,4203	1,3825	2,0243	0,5850	58,27	254
11	ind11	0,5931	0,0462	1,2886	1,9736	1,8835	2,2930	0,6475	79,60	266
12	ind12	0,9009	0,0483	1,1902	1,8046	1,3381	1,0698	0,5453	49,96	256
13	ind13	0,9284	0,0884	1,0473	0,9934	1,0497	1,0472	0,5753	58,01	263
14	ind14	1,0842	-	1,0762	1,6099	3,5233	2,2144	0,5764	77,03	288
15	ind15	1,1025	0,0510	1,0640	1,7266	0,9624	1,4338	0,4773	39,88	268
16	ind16	1,0362	-	1,2635	1,4507	2,3309	1,4845	0,5345	59,02	262
17	ind17	1,1131	0,0313	1,2298	2,0742	1,8154	2,3105	0,5688	62,89	292
18	ind18	0,8810	0,0407	1,3064	1,8217	1,7052	1,5776	0,5572	56,62	276
19	ind19	1,0670	-	1,6935	1,2732	1,4665	2,1813	0,5033	50,67	255
20	ind20	1,1476	-	1,3164	1,7371	1,3668	2,6214	0,5430	59,88	257
21	ind21	0,8367	0,0353	1,3490	1,9852	2,3300	2,2820	0,5722	58,63	269
22	ind22	0,7306	0,0565	1,1715	1,9637	1,7971	2,7844	0,5991	63,26	260
23	ind23	0,7893	0,0739	1,3905	1,9534	1,4225	1,3069	0,6294	73,87	267
24	ind24	0,9539	-	1,3067	2,6322	1,8775	1,9013	0,5517	60,55	251
25	ind25	0,9965	-	1,3263	2,0018	2,1171	2,7843	0,5526	71,64	295
26	ind26	1,0083	0,0435	1,1440	0,9841	1,4173	2,0474	0,5643	61,96	293
27	ind27	0,9246	-	1,3648	1,8567	2,5234	2,5066	0,6257	91,95	280
28	ind28	0,8803	-	1,3487	1,9690	2,6581	2,6448	0,5291	59,99	272
29	ind29	0,9549	0,0410	1,2229	1,6668	1,5770	1,7907	0,5140	45,12	262
30	ind30	0,9949	0,0572	1,0352	1,8411	1,4134	1,2375	0,5114	45,70	268

⁵ Veja descrição das variáveis na Tabela 13.

Tabela 13 - Descrição das variáveis

viagpes = nº de viagens individual

Dumysex = 1 se for sexo masculino
= 0 caso contrário

dsitf2 = 1 se o papel ocupado na família for o de cônjuge
= 0 caso contrário

dsitf3 = 1 se o papel ocupado na família for o de filho
= 0 caso contrário

dsitf4 = 1 se o papel ocupado na família for outro que não chefe, cônjuge
ou filhos
= 0 caso contrário

Dinst2 = 1 se o nível de instrução for até 1º grau incompleto
= 0 caso contrário

Dinst3 = 1 se o nível de instrução for até 2º grau incompleto
= 0 caso contrário

Dinst4 = 1 se o nível de instrução for até o superior incompleto
= 0 caso contrário

Dinst5 = 1 se o nível de instrução for superior completo
= 0 caso contrário

Dfam2 = 1 se a família for nuclear completa
= 0 caso contrário

Dfam3 = 1 se a família for nuclear completa e ampliada
= 0 caso contrário

Dfam4 = 1 se a família for quebrada
= 0 caso contrário

Dfam5 = 1 se a família for um grupo convivente
= 0 caso contrário

Dfam6 = 1 se a família for unipessoal
= 0 caso contrário

Dicv2 = 1 se a família for adulta segundo a classificação do índice de ciclo de vida (ICV) familiar
= 0 caso contrário

Dicv3 = 1 se a família for velha segundo a classificação do índice de ciclo de vida (ICV) familiar
= 0 caso contrário

idade = idade individual em anos completos

salind = salário individual em salários mínimos de setembro de 1987

salfam = salário familiar em salários mínimos de setembro de 1987

npesfam = número de membros do domicílio

As estimativas totais foram calculadas aplicando-se as equações de regressão estimadas para os 107.789 indivíduos, somando-se as mobilidades individuais estimadas.

Comparou-se estas estimativas com o número real de viagens totais dos 107.789 indivíduos pesquisados, o que resulta em 181.592 viagens.

Esta comparação foi baseada na seguinte porcentagem (chamada de erro):

$$\text{Erro} = \frac{\text{n}^\circ \text{ total de viagens estimadas} - \text{n}^\circ \text{ real de viagens (181.592)}}{\text{n}^\circ \text{ real de viagens (181.592)}}$$

Um erro = 0,10 significa que a estimativa é 10% maior que o número real de viagens.

Um erro = -0,05 significa que a estimativa é 5% menor que o número real de viagens.

Chamou-se de ‘errocom’ cada erro calculado com as 30 amostras usando os modelos que incluíam as variáveis de família.

Chamou-se de ‘errosem’ cada erro calculado com as 30 amostras usando os modelos que não incluíam as variáveis de família.

A média dos 30 ‘errocom’ é: -0,0454 com desvio padrão de 0,046.

A média dos 30 ‘errosem’ é: -0,1952 com desvio padrão de 0,052.

Fez-se as 30 diferenças (em módulo):

$|\text{errocom}| - |\text{errose}|$ obtendo-se um valor médio de -0,1448 com desvio padrão de 0,007.

Um teste t-Student para 2 médias provenientes de amostras dependentes ou emparelhadas (no caso as mesmas amostras, com e sem as variáveis de família) foi conduzido com a finalidade de testar se o erro médio de ambos os métodos eram iguais. A estatística $t_{\text{observada}} = -19,37$, que é altamente significativa, levando à rejeição da hipótese de que estes erros possam ser considerados iguais.

O módulo foi considerado para evitar aumentar a diferença entre o errocom e o errose em casos onde ambos tenham sinais diferentes. Por exemplo, se errocom = 0,10 e errose = -0,05,

$\text{dif} = |\text{errocom}| - |\text{errose}| = 0,10 - 0,05 = 0,05$ ao invés de $0,10 - (-0,05) = +0,15$.

A interpretação da diferença (dif) é a seguinte:

dif mede a diferença entre as distâncias em relação ao zero (o desejado é erro = zero) de ambos os métodos, com e sem as variáveis de família. Em primeira instância não importa se essa distância é para a esquerda ou para a direita do zero, importa é que é distante do zero! Sendo assim erros como -0,05 e 0,05 são considerados iguais, ao tomar-se o módulo.

Nos dados analisados, ambos os modelos com e sem as variáveis de família sempre subestimam, ou seja, os erros quase sempre são negativos. Contudo o interesse maior, no momento, é quantificar a margem de erro. Em trabalhos futuros pretende-se buscar com mais profundidade as razões das subestimativas.

A título de registro, se o módulo não fosse considerado, a estatística t -observada passa para 21,21; mais significativa que o valor anterior apresentado de -19,37.

Isto significa que ao se excluir as variáveis de família do modelo, comete-se em média um erro de -20% na estimação do total de viagens dos 107.789 indivíduos. Rejeita-se a hipótese de que esta média seja zero ($t_{\text{observado}} = -20,54$).

Quando se inclui as variáveis de família, comete-se em média um erro muito menor, -5% na estimação do total de viagens dos 107.789 indivíduos. Também se rejeita a hipótese de que esta média seja zero (no entanto o $t_{\text{observado}}$ é -5,39).

6.2.1 Alterações mediante a inclusão das variáveis: número de automóveis e ocupação principal

- **Número de automóveis**

Em relação ao número de automóveis da família observou-se que:

55,7% dos indivíduos estudados não possuem automóvel na família

32,6% possuem 1 automóvel

9,2% possuem 2 automóveis

2,5% possuem mais que 2 automóveis

Os *box plot* abaixo mostram a variação do número de viagens diárias individuais de acordo com o número de automóveis na família do indivíduo.

Número de Automóveis

Nº de viagens diárias individuais

Número de Automóveis

Nº de viagens diárias individuais

- **Ocupação principal**

Observou-se a seguinte divisão dos indivíduos segundo sua ocupação principal:

20,8% têm como ocupação principal estudar.

43,0% são aposentados, desempregados, afastados, nunca trabalharam ou desempenham a função de prendas domésticas.

36,2% têm como ocupação principal sua inserção no mercado de trabalho nas mais diversas formas, desde empregados em empresas, trabalhador rural, profissional autônomo, funcionário público a sócio ou proprietário de pequenos comércios a grande fazendeiro.

Os *box plot* abaixo mostram a variação do número de viagens de acordo com a ocupação principal.

Ocupação Principal

Nº de viagens diárias individuais

(*) inativos do ponto de vista de trabalho: são os 43,0 % que são aposentados, donas de casa, desempregados, etc.

A tabela 14, abaixo, mostra a alteração nos coeficientes de determinação (r^2) do modelo de regressão linear para explicar o número de viagens diárias individuais, de acordo com a inclusão das variáveis: número de automóveis da família e ocupação principal do indivíduo.

Tabela 14 - Alterações no ajuste geral dos modelos com e sem a inclusão das variáveis: número de automóveis da família e ocupação principal do indivíduo. Teste F-Snedecor e coeficiente de determinação (r^2)

		Tamanho de Amostra (n)					
Alterações nas variáveis do modelo	Estatística calculada	98877	209	200	232	224	247
SEM nº de automóveis	F-Snedecor	7890,16	17,901	17,712	21,686	18,476	27,977
Ocupação principal	r^2	0,6021	0,6416	0,6503	0,6592	0,6313	0,6998
COM nº de automóveis	F-Snedecor	7543,124	17,143	17,638	20,597	17,523	26,579
	r^2	0,6041	0,6446	0,6621	0,6607	0,6321	0,7008
COM Ocupação principal	F-Snedecor	9538,970	22,083	23,253	25,778	25,078	34,464
	r^2	0,6696	0,7115	0,7318	0,7196	0,7218	0,7620
COM nº de automóveis	F-Snedecor	9162,10	21,455	23,631	24,568	23,869	32,845
Ocupação principal	r^2	0,6709	0,7162	0,7449	0,7202	0,7222	0,7393

Observe que a variável número de automóveis da família não tem contribuição relevante ao modelo, ao contrário da variável ocupação principal do indivíduo que traz um aumento significativo ao r^2 .

Contudo convém ressaltar que o objetivo principal deste trabalho é diagnosticar a importância do grupo de variáveis domiciliares, destacando sua magnitude mesmo ao incluir-se variáveis tão importantes quanto as mencionadas: número de automóveis da família e ocupação principal do indivíduo.

A inclusão destas variáveis, no modelo de regressão, juntas ou separadamente, alteram a magnitude da significância estatística da contribuição do grupo de variáveis domiciliares. Entretanto a significância mantém-se muito alta como mostra a Tabela 15 abaixo onde a significância estatística vale para todo α (nível de significância) maior que α^* (nível de significância descritivo). Desta forma se $\alpha^* = 0,0001$ (ou 0,01%) significa que o valor de $F_{\text{observado}}$ é significativo para todo $\alpha > 0,0001$ (0,01%).

Tabela 15 - Teste F-Snedecor da contribuição estatística do grupo de variáveis domiciliares para modelos com e sem a inclusão das variáveis: número de automóveis da família e ocupação principal do indivíduo

		Tamanho de Amostra (n)					
Alterações das variáveis no modelo	Estatística calculada	98877	209	200	232	224	247
SEM nº de automóveis	F-Snedecor	1001,3900	1,9600	2,3600	4,0719	4,3310	3,3202
Ocupação principal	α^*	0,0001	0,0299	0,0076	0,0001	0,0001	0,0002
COM nº de automóveis	F-Snedecor	853,4536	1,7715	2,0388	3,8181	3,9407	3,3169
	α^*	0,0001	0,0554	0,0233	0,0001	0,0001	0,0002
COM Ocupação principal	F-Snedecor	258,2727	1,6094	2,3917	3,2234	3,9061	1,1588
	α^*	0,0001	0,0918	0,0069	0,0003	0,0001	0,3143
COM nº de automóveis	F-Snedecor	195,8374	1,6269	2,4714	3,0358	3,8877	1,1956
Ocupação principal	α^*	0,0001	0,0871	0,0052	0,0006	0,0001	0,2872

(*) α é o nível de significância descritivo, ou seja o valor de F observado é significativo para qualquer nível de significância maior ou igual ao apresentado na tabela.

6.3 Detalhamento da estimação de um modelo individual

O método *stepwise* de seleção de variáveis partiu do seguinte conjunto de variáveis: sexo, papel ocupado dentro do grupo familiar, nível de instrução, tipo de estrutura familiar, índice de ciclo de vida da família, salário individual e número de membros do grupo familiar.

O resultado obtido foi a eliminação do sexo e do número de membros do domicílio. O modelo resultante é apresentado abaixo onde a estatística t-Student, que testa a contribuição de cada variável está entre parênteses, logo abaixo do parâmetro estimado. A descrição das variáveis está na Tabela 13.

- A partir de uma amostra com 232 indivíduos.

$$\begin{aligned}
 \text{Viagpes} = & 0,41 \times \text{dsitf3} - 0,71 \times \text{dsitf4} + 0,92 \times \text{Dinst2} + 1,26 \times \text{Dinst3} + \\
 & \quad (2,43) \quad \quad (-2,22) \quad \quad (4,59) \quad \quad (4,25) \\
 & + 0,66 \times \text{Dinst4} + 1,45 \times \text{Dinst5} + 0,66 \times \text{Dfam3} + 1,23 \times \text{Dfam5} + \\
 & \quad (2,00) \quad \quad (3,38) \quad \quad (3,22) \quad \quad (2,60) \\
 & + 0,63 \times \text{Dicv2} + 0,06 \times \text{Salind} \\
 & \quad \quad (3,61) \quad \quad (3,56)
 \end{aligned}$$

$$r^2 = 0,6450 \quad r^2_{\text{ajustado}} = 0,6291 \quad F = 40,34 \quad \alpha = 5\%$$

Os resíduos deste modelo estão tabulados resumidamente abaixo, descritos no Gráfico 6 e não mostram haver indicações da ausência da suposição de normalidade dos dados.

⁶ $r^2_{\text{ajustado}} = 1 - \frac{(n-i)(1-r^2)}{n-p}$, $i=1$ se o intercepto $\neq 0$ e 0 caso contrário. Este valor é importante se n for muito próximo de p , neste caso r^2 é viesado e deve ser usado o coeficiente de determinação ajustado.

Gráfico 6 - Frequências dos resíduos do modelo estimado com 232 indivíduos da RMSP

O Gráfico nº 7 não indica a existência de autocorrelação nos erros.

Gráfico 7 - Dispersão dos resíduos do modelo estimado com 232 indivíduos da RMSP

O número de condição espectral para este modelo é 3,43 (não pode ser maior que 100, o que indicaria forte grau de multicolinearidade).

Outras medidas de diagnóstico foram calculadas, como os Dffits, Dfbetas, hii, etc. descritos no anexo A. Não há nenhuma indicação relevante quanto à inadequabilidade do modelo ajustado.

- A partir de todos os dados disponíveis: 107.789 indivíduos.

O modelo estimado com os 107.789 indivíduos e partindo de todas as variáveis disponíveis, portanto incluindo a idade e o salário familiar, fornece as seguintes estimativas dos parâmetros:

$$\text{Viagpes} = 0,3768 \times \text{Dumysex} - 0,1035 \times \text{dsitf2} + 0,0117 \times \text{dsitf3} -$$

(35,82) (-6,49) (0,75)

$$- 0,3770 \times \text{dsitf4} + 0,6668 \times \text{dinst2} + 1,0246 \times \text{dinst3} +$$

(-18,99) (59,07) (67,24)

$$+ 1,2130 \times \text{dinst4} + 1,3150 \times \text{dinst5} + 0,6871 \times \text{dfam2} +$$

(72,53) (57,09) (38,33)

$$+ 0,6738 \times \text{dfam3} + 0,7461 \times \text{dfam4} + 0,8474 \times \text{dfam5} +$$

(30,80) (34,10) (29,15)

$$+ 0,8017 \times \text{dfam6} + 0,3260 \times \text{dicv2} - 0,0949 \times \text{dicv3} -$$

(20,15) (27,82) (-2,65)

$$- 0,0030 \times \text{idade} + 0,0285 \times \text{salind} + 0,0049 \times \text{salfam} + 0,0160 \times \text{npesfam}$$

(-8,81) (29,96) (12,59) (5,82)

$$r^2 = 0,5887 \quad r^2_{\text{ajustado}} = 0,5886 \quad F = 8117,80 \quad \alpha = 0,01\% (0,0001)$$

Observe que excetuando-se dsitfam3 todos os t-observados, entre parênteses, são altamente significantes.

6.4 Discussão dos resultados obtidos com os modelos gerados a partir dos dados da O/D-1987 - Metrô

Conforme comentado anteriormente, entre as diversas etapas de análises estatísticas houve um momento em que se dispunha somente de dados agrupados por faixas de renda, de salário, de instrução e etc.

A maneira de tratar estatisticamente estes dados era representar as características de um indivíduo pela média do grupo a que este indivíduo pertencia, ou seja, a um indivíduo com 30 anos, do sexo masculino, chefe de família, com nível de instrução até 4ª série do primeiro grau e renda de 3 salários mínimos, era atribuído uma mobilidade média obtida através do quociente do número de viagens total da categoria à qual pertencia este indivíduo pelo número de habitantes da respectiva categoria.

A perda de informação é muito grande, por exemplo não era conhecido que 34,7% dos indivíduos pesquisados tinham número de viagens igual a zero, nem se podia caracterizá-los. O que fazia com que a média de viagens sempre fosse mais baixa que o esperado. Levando-se em conta que boa parte da população (46,8%) faz no mínimo duas viagens, uma para o trabalho e a outra, o retorno para a residência.

A variação do número médio de viagens de cada categoria fica dividida pelo número de indivíduos da categoria, ou seja, $\text{variancia}(\bar{X}) = \frac{\text{variancia}(X)}{n}$ onde $\bar{X}$ é a média de viagens da categoria e X é o número de viagens de cada indivíduo da categoria.

Isto é bem ilustrado nos Gráficos 2, 3, 4 e 5 onde gráficos para os dados em categorias são bem suavizados em relação aos dados individuais, que oscilam muito.

Para que todas as categorias possam ser representadas adequadamente, sem que se comprometa ainda mais a perda de informação foi necessário gerar cerca de 1960 categorias das quais 783 tiveram representação (2 sexos x 4 situações ocupadas na família x 5 níveis de instrução x 7 faixas de renda x 7 faixas de idade = 1960).

Este deve ter sido um dos motivos pelos quais Supernak, Talvitie & De John (1983) não detectaram a influência das características domiciliares. Convém lembrar que os dados individuais foram agrupados em categorias e a única variável considerada dentro deste contexto era o tipo de estrutura familiar. Os autores abandonaram esta linha de pesquisa no início da análise estatística dos dados.

Na literatura aparece a discussão da não verificação das hipóteses do modelo de regressão quando se procura ajustar um modelo desagregado a nível de domicílios (White, 1976). Foi surpreendente a adequação apresentada pelos modelos ora ajustados, desagregado a nível de indivíduos, mesmo levando em consideração que só a apresentação do coeficiente de determinação não é suficiente para considerar-se um modelo de análise de regressão ajustado satisfatoriamente.

Foram muito satisfatórios os resultados obtidos através de criteriosa análise de resíduos e procedimentos de diagnóstico em regressão (Montgomery & Peck, 1983) com a finalidade de detectar observações influentes ou discrepantes,

bem como violações das suposições do modelo e a presença de multicolinearidade. Estes resultados, além de um valor do coeficiente de determinação (r^2) e da estatística *F-Snedecor* (sempre significativa), mostram que o modelo foi capaz de captar grande parte da variação do número médio de viagens geradas pelos indivíduos na região Metropolitana de São Paulo - S.P.

É interessante observar que os coeficientes estimados das variáveis explicativas do modelo apresentado revelam comportamentos esperados, tais como:

- a) os filhos e a seguir os chefes, são mais móveis (quando comparados as outras opções de papéis dentro da família) se mantidas as demais características constantes.
- b) a mobilidade aumenta com o nível de instrução.
- c) como era de se esperar pelos resultados evidenciados na análise descritiva, a diferença entre os sexos quanto à mobilidade não aparece uma vez que fatores mais relevantes que o sexo por si só, determinam as diferenças de mobilidade.
- d) os coeficientes baixos para a renda parecem não revelar a conhecida variação da mobilidade com a renda. Na verdade esta variação ficou diluída entre as demais variáveis que têm relação com a renda, como a instrução, embora não se tenha detectado grau acentuado de multicolinearidade, provavelmente devido à diversidade da população da RMSP.

Há autores que afirmam que a quantidade de viagens depende predominantemente das necessidades específicas do indivíduo em desenvolver atividades fora do domicílio, a despeito de suas restrições orçamentárias, desde

que sua percepção aponte para um baixo custo, num sentido amplo da palavra custo (Supernak, 1983).

Pesquisar-se os domicílios, no sentido de levar em conta dados individuais de todos os seus membros, parece não interferir na análise estatística de dados se comparados com amostras aleatórias de indivíduos fora do âmbito de seu domicílio, ou seja, amostras em que a família toda ou o grupo todo com o qual convive o indivíduo não são considerados juntos.

Esta afirmação baseia-se no cálculo das correlações de Pearson entre o número de viagens para os membros de 50 famílias extraídas aleatoriamente para a variável número de viagens individual, apresentados na Tabela 16. Nenhum dos dados abaixo é significativo ao nível de significância de 5%.

Tabela 16 - Correlação de Pearson para o número de viagens diárias individuais

	chefe	cônjuge	filhos
chefe			
cônjuge	0,1849		
filhos	0,2008	0,1840	
outros	-0,2172	-0,0178	0,5056

Parece haver correlação em famílias ampliadas (15,7% das famílias pesquisadas) entre o número de viagens de filhos e de outros membros da família, embora também não se rejeite a hipótese de que esta correlação seja zero.

A importância destes fatos consiste na dificuldade de se obter amostras probabilísticas de indivíduos em amostras não domiciliares. Entrevistar pessoas num local conveniente entre os transeuntes, tomados a olho, não constitui amostra probabilística, o que impede o uso das técnicas estatísticas para fins de inferência. Os resultados podem ser úteis mas não são estatísticos. Para uma discussão mais ampla deste assunto veja Rosander (1977) e Kish (1965).

7. A CONTRIBUIÇÃO EFETIVA DESTE TRABALHO

Está muito bem estabelecida na literatura que a demanda por transportes é uma demanda derivada de outras demandas, por exemplo por atividades ligadas ao lazer, a satisfação de necessidades fisiológicas, econômicas e etc.

Uma primeira contribuição deste trabalho é estabelecer um elo de ligação forte entre as características domiciliares e o comportamento do indivíduo, mostrando como tais características influenciam as atividades dos indivíduos e por conseguinte seus deslocamentos. Isto foi bem estabelecido no Capítulo 4 com o auxílio da sociologia e da psicologia social, cujos conceitos orientaram na escolha de quais variáveis selecionar para efetivamente incorporar ao modelo de geração de viagens a influência das características domiciliares sobre o indivíduo.

Uma segunda contribuição relevante é o detalhamento dos métodos estatísticos. Sendo esclarecidas suas limitações e potencialidades no decorrer do Capítulo 3.

As análises de dados descritas nos Capítulos 5 e 6 bem como as notas metodológicas ali inseridas, em muito contribuem para a orientação do uso do método de regressão linear bem como nos cuidados a serem tomados. Destaca também a importância que se deve dar à análise preliminar dos dados. A análise estatística descritiva inicial, por exemplo o exame dos “*box plots*”, direciona a elaboração do modelo ao detectar-se inicialmente quais as fontes de variação no número de viagens individuais.

Também é de importância ter mostrado que desagregar a nível individual não só é viável como efetivamente oferece vantagens, logo uma terceira contribuição é a retomada desta linha de pesquisa há muito negligenciada pela literatura.

Uma quarta contribuição é ter incorporado ao modelo de geração de viagens a influência das características domiciliares sobre o indivíduo do ponto de vista abordado. Sendo que os resultados apontados na literatura nesta direção não são satisfatórios.

O sucesso na detecção da significância estatística do grupo de variáveis domiciliares (Capítulo 6) deve-se à desagregação ao nível do indivíduo e aos resultados obtidos no exame dos '*box plots*' para os diversos cruzamentos dos níveis das variáveis, onde se pode observar haver uma real variação do número de viagens diárias individuais para cada contexto domiciliar onde está inserido o indivíduo.

Uma quinta contribuição é expor resultados estatísticos obtidos com dados de alta credibilidade como os da pesquisa Origem/Destino de 1987, Metrô-SP, cujo rigor desde a elaboração da pesquisa até as etapas finais são dignos de nota.

O procedimento geral usado para analisar os dados consistiu de algumas etapas básicas descritas abaixo:

- Seleção das variáveis de interesse: sexo, idade, renda individual e familiar (em salários mínimos de setembro de 1987), nível de instrução (em 5 categorias), número de pessoas da família, papel desempenhado dentro da estrutura familiar e número de viagens diárias individual.

- Transformação deste arquivo num arquivo de dados do SAS - Statistical Analysis System a partir da leitura do arquivo original em ASCII.
- Correção do arquivo com o intuito de eliminar dados inconsistentes como por exemplo famílias com 2 chefes, com 50 filhos, com valor negativo de renda e etc. Para tal, tabulações foram realizadas construindo-se a distribuição de frequências das diversas variáveis, bem como, de alguns cruzamentos. A partir daí, examina-se todo o domicílio de onde pertence o indivíduo com algum valor inconsistente e decide-se eliminá-lo por completo ou não.
- Transformação das variáveis ou criação de outras a partir das já existentes: foi o caso da atribuição do tipo de estrutura familiar e a construção do índice de ciclo de vida da família, descritos no item 5.2.1.
- Exame criterioso da influência de cada variável bem como de cada cruzamento das variáveis sobre o número de viagens individuais, sobretudo através do exame dos “*box plots*”.
- Construção dos vários modelos apresentados para diversas amostras aleatórias selecionadas (também como auxílio do SAS - Statistical Analysis System) conforme diretrizes delineadas no Capítulo 6, incluindo uma análise dos erros ao se estimar a mobilidade do conjunto total de indivíduos disponíveis usando amostras bem menores.

8. CONCLUSÕES, RECOMENDAÇÕES, SUGESTÕES

• Conclusões gerais

- A verificação estatística mostrou a magnitude da contribuição da inclusão de variáveis domiciliares no modelo desagregado em nível individual, o que não ocorre no modelo desagregado em nível de domicílios conforme conclusões obtidas por Stopher & McDonald, 1983.
- Além das vantagens apontadas por Supernak (1981) a desagregação em nível individual é necessária para se obter um modelo que represente o número de viagens que um indivíduo realiza, num dado intervalo de tempo. Pois, desta forma pode-se viabilizar a inclusão de características específicas do indivíduo dentro de seu contexto domiciliar, como por exemplo o papel que ele ocupa dentro da estrutura do grupo em que convive.
- Ao se desagregar o modelo em nível domiciliar, limitam-se variáveis domiciliares àquelas que caracterizem aspectos não relacionados especificamente à influência dos membros do domicílio sobre o indivíduo diretamente, perdendo-se informação importante para detectar mudanças no padrão de deslocamentos. Excetua-se o número de membros do domicílio.
- Estas variáveis domiciliares devem incluir variáveis que caracterizem não só o papel ocupado pelo indivíduo dentro da estrutura de seus membros, bem como, variáveis que possam diferenciar as formas de organização das atribuições destes papéis. Neste trabalho utilizaram-se as relações de

parentesco, classificando as famílias como nucleares simples, completas, quebradas, etc.

- Tanto as análises estatísticas descritivas como os testes de significância tornaram evidente a contribuição do estágio do ciclo da vida da família para compreender as transições nos papéis desempenhados pelo indivíduo e pela família como um todo, bem como as conseqüentes mudanças no padrão de deslocamentos.
- Pode-se observar que a posição espacial ocupada pelo hiperplano estimado quando se inclui as variáveis que descrevem as características domiciliares, conforme diretrizes acima citadas, minimiza os erros, quando comparado com os pontos observados no espaço todo. Pode-se imaginar que este hiperplano esteja bem centralizado na nuvem de pontos observada. O que não ocorre ao se considerar o hiperplano estimado sem a inclusão do grupo de variáveis domiciliares.
- Verificou-se que pequenas amostras de indivíduos de tamanho médio igual a 265, tomados isoladamente, separando-os dos demais membros do domicílio, forneceram resultados muito satisfatórios para a estimação do total de viagens de um grupo de 107.789 indivíduos.
- Resultados análogos puderam ser obtidos tomando-se amostras de 50 domicílios, pesquisando-se todos os membros em cada um deles. Contudo neste caso deve-se verificar a não existência de correlação entre os membros. Não é desejável que ocorra correlação estatisticamente diferente de zero entre chefes e cônjuges por exemplo. Isto acarretaria correlações nos erros do modelo de regressão.

- As análises estatísticas, sobretudo através dos *box plot's* evidenciaram um grande efeito de interação das variáveis familiares, ou seja, evidenciaram a detecção de variações no número de viagens de indivíduos inseridos nas diversas estruturas familiares, sobretudo quando se leva em conta o papel desempenhado dentro do grupo no qual ele convive.
- De forma geral os modelos individuais acrescidos das características domiciliares, foram capazes de explicar grande parte da variação do padrão de viagens. Isto se deve, em grande parte, à captação de características intrínsecas ao indivíduo no que diz respeito ao comportamento relacionado a decisões de realizar deslocamentos.
- Convém ressaltar que o objetivo principal deste trabalho é diagnosticar a importância do grupo de variáveis domiciliares, destacando sua magnitude mesmo ao incluir-se variáveis tão importantes quanto: número de automóveis da família e ocupação principal do indivíduo. A inclusão destas variáveis, no modelo de regressão, juntas ou separadamente, alteram a magnitude da significância estatística da contribuição do grupo de variáveis domiciliares. Entretanto a significância ainda se mantém muito alta.

• **Sugestões e recomendações para pesquisas futuras**

PESQUISAS FUTURAS

Algumas questões ainda podem ser melhores esclarecidas:

- Este estudo seria muito enriquecido ao se estudar as diferenças da mobilidade por motivos, tais como, trabalho, lazer, estudo, etc.

- Verificar o porquê da predominância da subestimação também seria conveniente e traria grande contribuição.
- Investigar melhor a razão de o coeficiente da renda ser tão baixo.
- Detectar a real diferença entre os sexos, ela existe ou fica diluída nos papéis que os indivíduos ocupam dentro da estrutura familiar?
- A possibilidade de que o modelo individual acrescido das variáveis domiciliares possa ser transferido no tempo e no espaço deve ser bem maior do que os modelos obtidos com a desagregação em nível domiciliar, ou até mesmo individual, quando não se leva em conta as variáveis que caracterizam o domicílio do ponto de vista abordado. Tal hipótese ainda não foi verificada estatisticamente.
- Verificar a magnitude da contribuição do grupo de variáveis domiciliares diante da inclusão no modelo de outras variáveis individuais de relevância na descrição do padrão de viagens, por exemplo como foi feito com a inclusão da variável número de automóveis da família e ocupação principal do indivíduo na seção 6.2.1.

DESENVOLVIMENTO DAS ANÁLISES DE DADOS

- Em qualquer destes estudos é conveniente começar as análises estatísticas pelo estudo das variações que cada variável provoca no número de viagens diárias, por exemplo, efetuar-se a comparação dos “*box plots*” para os diversos níveis das variáveis; para diversas estruturas familiares comparadas aos filhos e etc. Logo não se deve esquecer de coletar variáveis, aqui chamadas de variáveis domiciliares que são responsáveis por uma

contribuição significativa para explicar a variação do número de viagens individuais diárias.

- Para aqueles que desejarem usar o SAS: Statistical Analysis System, outra questão de importância é a necessidade de consulta constante aos manuais do SAS (1989, 1990, 1991). São diversos manuais que somados conduzem a utilização correta das rotinas disponíveis. A ajuda *on line* não substitui de forma alguma tal consulta, a não ser que se conheça em detalhes a rotina a ser utilizada. É comum as Universidades possuírem a licença oficial para uso e instalação do sistema completo tendo adquirido somente os manuais básicos e não o módulo estatístico: SAS/STAT que é imprescindível.
- É de extrema importância a correção do arquivo de dados, visando a eliminação de dados inconsistentes. Por exemplo, zeros que significam não resposta ao invés de significar zero mesmo; famílias com 2 chefes já que só é admitido um chefe; valor negativo de renda; etc. Dos quase 111.000 indivíduos pesquisados, utilizou-se dados de 107.789 e em alguns casos 99061. Isto deve ao fato de que o número de dados inconsistentes varia de acordo com o parâmetro de interesse, por exemplo, nem todos que responderam número de viagens, responderam o valor da renda. Observe que cerca de 10% dos dados foram desprezados em algumas análises.

APLICAÇÃO DO ENFOQUE PROPOSTO

- Para a aplicação de modelos com o enfoque aqui proposto é muito importante a determinação do estágio no ciclo da vida da família. Existe vasta literatura a respeito, recomendo a leitura de Bussab & Wagner (1994) como imprescindível. A aplicação do método ali apresentado constitui um dos pontos altos deste trabalho, em virtude não só da qualidade do método

em sí, bem como da fidedignidade dos dados utilizados nas pesquisas do SEADE: Fundação Sistema Estadual de Análises de Dados.

- Finalmente recomenda-se a leitura de van de Bilt (1997). Trata-se de uma dissertação de mestrado defendida 20 dias antes da defesa desta tese. A importância deste trabalho, que não aparece na revisão bibliográfica apresentada no Capítulo 3, deriva não somente de sua qualidade, bem como da escassez na literatura de trabalhos que focalizem a desagregação em nível individual. Este trabalho não pode deixar de ser aqui comentado também em virtude dos vários pontos em comum com o trabalho ora apresentado. O principal objetivo de van de Bilt (1997) “é a identificação e compreensão das variáveis que determinam os padrões individuais e familiares de atividades e viagens, medidos através das taxas de produção de viagens, classificadas por motivos” (van de Bilt, 1997 p.117). Muitos dos resultados ali apresentados corroboram com a tese aqui proposta. A autora sistematiza o critério de escolha de categorias das variáveis através da modelagem de segmentação, utilizando o procedimento CHAID: Chi Square Automatic Interaction Detection. Também aplica o método para os dados da pesquisa Origem/Destino - 1987 - Metrô - SP. Ela encontra fortes evidências que mostram ser o padrão de atividades e viagens dos indivíduos influenciado por características dos domicílios e vice-versa, ou seja, os “padrões de comportamento das famílias são também em parte, determinados pela participação dos indivíduos que as compõe em atividades compulsórias” (van de Bilt, 1997 p.93). Conclui, portanto, ter trazido luz a questão, dita polêmica, da unidade amostral a ser utilizada na modelagem de geração de viagens de passageiros urbanos: indivíduos ou domicílios? Conclui também que embora “interpretações mais ricas possam ser obtidas com a desagregação em nível individual, o enfoque mais adequado é uma combinação dos dois” (van de Bilt, 1997 p.93), ou seja, modelos individuais

que utilizem características domiciliares ou modelos domiciliares que incluam características individuais. Aponta como “potencialmente útil a exploração mais detalhada dos conceitos de estágio no ciclo da vida e estrutura domiciliar” (van de Bilt, 1997 p.92) do ponto de vista das relações entre os membros, afirmando não se ter chegado a um consenso sobre quais fatores são mais adequados para a definição destas variáveis, nem quanto a real vantagem da utilização delas nos modelos de análise de demanda. Só foi possível apresentar estas últimas considerações neste texto em virtude das correções efetuadas após a defesa desta tese com a finalidade de elaborar o texto definitivo a ser entregue para a EESC - USP - São Carlos, num prazo máximo de 60 dias após a data da defesa, período no qual pude ler, com prazer, o referido trabalho.

ANEXO A - ANÁLISE DE REGRESSÃO LINEAR - UM RESUMO DESCRITIVO E INTERPRETATIVO

A.1 O modelo de regressão linear

O modelo estatístico de uma regressão linear múltipla é:

$$Y_j = a_0 + a_1 X_{1j} + a_2 X_{2j} + \dots + a_{p-1} X_{(p-1)j} + u_j, j=1, \dots, n.$$

onde:

Y_j : é a variável dependente.

a_i : são os parâmetros a serem estimados.

X_{ij} : são as variáveis explicativas.

u_j : é o erro, parte do modelo não explicada pela regressão.

ou em notação matricial:

$$Y = Xa + e$$

onde:

Y : é uma matriz $n \times 1$, cujas linhas são os valores da variável dependente para cada uma das n observações.

X : é uma matriz $n \times p$, de n observações de p variáveis, chamadas de variáveis preditoras ou explicativas.

a : é um vetor $p \times 1$, de p parâmetros a serem estimados.

e : é um vetor $n \times 1$ de erros ou resíduos.

As hipóteses necessárias são:

1 - Aditividade: a variável dependente (Y_j) é função linear das variáveis independentes: X_{ij} , $i=1, \dots, p-1$. (por exemplo, não é multiplicativa: $Y = X_1^{a_1} \cdot X_2^{a_2} + e$).

2 - Os valores das variáveis independentes são fixos.

3 - Os erros são aleatórios, com distribuição de probabilidade normal, são independentes, têm a mesma variância e têm média zero, resumindo:

$$e_j \sim N(0, \sigma^2) \text{ e } E(ee') = I \sigma^2,$$

onde: E é a esperança matemática (média).

Estas hipóteses implicam que:

$$Y/X_i \sim N(Xa, \sigma^2 I)$$

Em outras palavras: uma vez que os erros ou resíduos são variáveis aleatórias independentes com distribuição normal e variância constante assim o é Y dado

X_i , por ser uma função linear de variáveis não aleatórias acrescido de um erro aleatório com tais características.

A.2 Desvios das hipóteses do modelo de regressão: causas e diagnóstico (detecção)

Pode-se resumir as hipóteses básicas do modelo de regressão em quatro palavras chaves, das quais as 3 primeiras descritas abaixo são referentes aos erros aleatórios:

- não correlação,
- normalidade,
- homocedasticidade (igualdade das variâncias)
- aditividade.

A.2.1 Causas dos desvios das suposições

A suposição de não correlação dos erros, em geral, pode ser assegurada através de um esquema de aleatorização apropriado no estágio de planejamento do experimento. A violação desta suposição ocorre mais frequentemente quando as observações são tomadas sequencialmente ao longo do tempo. Dados correlacionados exigem métodos de análise apropriados.

A heterocedasticidade dos erros, desigualdade das variâncias, pode aparecer de várias formas, pode ser causada por condições menos cuidadosas de controle

do experimento ou pela escolha da unidade amostral. Isto é mais comum em experimentos agrícolas, onde uma dosagem intermediária de adubo em solo ácido apresenta boas produções em alguns lotes e produções muito baixas em outros, causando grande variância.

Há um tipo de heterocedasticidade que está associada a não normalidade dos dados, por exemplo, onde a variância é função da média. Pode ainda ser causada pela presença de *outliers*, que são observações atípicas ou muito discrepantes.

A normalidade perfeita raramente ocorre em dados reais, contudo pequenos desvios da normalidade não afetam seriamente o modelo (Montgomery e Peck, 1983) e a não aditividade pode ser causada pela presença de *outliers* ou ser inerente ao modelo construído.

A.2.2 Diagnóstico: detecção dos desvios das suposições

Sob o nome geral de diagnóstico em regressão, estuda-se técnicas que auxiliem a detectar desvios nas suposições do modelo, a presença de multicolinearidade na matriz X e a presença de *outliers* entre os dados usados no processo de estimação dos parâmetros do modelo.

As estratégias apresentadas na literatura para este diagnóstico resumem-se em três tipos:

- A análise descritiva dos dados: quando criteriosamente desenvolvida pode detectar desvios das suposições antes de ajustar o modelo.

- A análise de resíduos: é uma análise descritiva contudo aplicada após o ajuste do modelo.
- Testes de significância ou testes de hipóteses sobre os parâmetros do modelo: são técnicas de inferência estatística, portanto são mais formais.

A.2.2.1 A análise descritiva dos dados

Fornece evidências sobretudo quanto à normalidade dos dados e a heterocedasticidade.

- Quanto à normalidade:

A construção do histograma ou da distribuição de frequências de Y ou dos resíduos pode ajudar a encontrar evidências de não normalidade dos dados. Se aproximadamente 68% dos resíduos padronizados $\left(\frac{e_i}{\sqrt{qme}}\right)$ estiverem entre -1 e 1 ou cerca de 95% estiverem entre -2 e 2 não há evidências para suspeitar-se de ausência de normalidade.

Também o uso do gráfico de probabilidade normal, disponível em alguns softwares, indica não normalidade. Este procedimento consiste em colocar os resíduos num sistema de eixos bidimensional num papel conhecido como papel de probabilidade normal. Este é um papel especial desenhado para que a função de distribuição acumulada da distribuição normal apresente-se como uma reta. Então desenhando no gráfico os valores dos resíduos contra suas

freqüências acumuladas, deve-se encontrar uma curva próxima de uma reta caso esteja diante da distribuição normal.

- **Quanto à heterocedasticidade dos erros:**

Quando se dispõe de várias unidades amostrais, formando grupos com o mesmo valor de Y , há evidências de heterocedasticidade quando: as variâncias amostrais de cada grupo apresentarem-se muito diferentes entre si; os coeficientes de variação nos grupos forem muito próximos, indicando que a variância cresce com a média; os coeficientes de variação nos grupos forem muito diferentes quando as médias são muito próximas.

A.2.2.2 Análise de resíduos

É uma técnica muito eficiente para detectar-se desvios das suposições. Consiste em métodos de análises gráficas, métodos numéricos e ainda outros mistos (Siqueira, 1983).

É recomendada como um procedimento de rotina.

A análise de resíduos não só é válida para resíduos provenientes de ajustes realizados através da regressão linear, como o é para toda a situação onde um modelo é ajustado e medidas de variação não explicada, na forma de resíduos, são disponíveis para o exame (Draper & Smith, 1981).

O i-ésimo resíduo, denotado por e_i , é definido como a diferença entre o i-ésimo valor da variável dependente observado, y_i , e o i-ésimo valor ajustado, $\hat{y}_i$:

$$e_i = y_i - \hat{y}_i$$

Portanto o resíduo é a parte que o modelo não consegue explicar. Também são interpretados como erros observados sob a hipótese de que o modelo esteja correto (Draper & Smith, 1981; Siqueira, 1983). Sendo assim, os resíduos devem apresentar tendências que confirmem estas hipóteses.

Os resíduos têm média zero e sua variância média é:

$$\frac{\sum_{i=1}^n (e_i - \bar{e})^2}{n-2} = \frac{\sum_{i=1}^n e_i^2}{n-2} = \frac{sqe}{n-2} = qme$$

onde:

sqe: é a soma de quadrados dos erros ou residual: $\sum_{i=1}^n (\hat{y}_i - \bar{y})^2$

qme: é o quadrado médio dos erros: $sqe/(n-p-1)$. É um estimador não viesado para σ^2

Observe que a expressão acima é para a variância média, a variância de cada e_i é dada por:

$$V(e_i) = (1-h_{ii}) \sigma^2$$

onde,

h_{ii} é o valor do i -ésimo elemento da diagonal da matriz H dada por:

$$H = X(X'X)^{-1}X'$$

Os valores estimados de y , variável dependente no modelo de regressão, são obtidos fazendo-se:

$$\hat{y} = Hy$$

Portanto H é a matriz que “põe o chapéu” em y , por este motivo seu nome é H , de *hat* em inglês que significa chapéu. Na verdade ela é a matriz de projeção ortogonal no espaço gerado pelos vetores colunas da matriz X (Montgomery & Peck, 1983; Dachs & Carvalho, 1984).

H é conhecida como *hat matrix* ou matriz chapéu.

Os valores h_{ii} são tais que $0 < h_{ii} < 1$, logo superestima-se as variâncias dos resíduos quando se utiliza o quadrado médio do erro, qme , como estimador ao invés de $(1-h_{ii}) \sigma^2$.

Mesmo assim é comum trabalhar-se com os resíduos padronizados segundo a transformação abaixo:

$$d_i = \frac{e_i}{\sqrt{qme}}$$

Resíduos assim padronizados mantêm a média igual a zero e têm a variância aproximadamente igual a 1 e são úteis para detectar-se *outliers*.

Geralmente pontos próximos da origem têm variância maior do que resíduos mais distantes da origem. Violações das suposições dos modelos são mais suaves em pontos mais distantes da origem e portanto mais difíceis de serem detectadas pelo exame dos resíduos e_i ou pelos resíduos padronizados como d_i . Por isto alguns autores sugerem a padronização conhecida como resíduo “studentizado”.

Os resíduos *studentizados* são calculados fazendo-se:

$$\frac{e_i}{\sqrt{s_i^2(1 - h_{ii})}}$$

onde s_i^2 é o estimador da variância do erro (qme). Existem outras formas de *studentizar* os resíduos descritas em Dachs & Carvalho (1984).

Os resíduos não são independentes, a covariância entre dois resíduos é dada por:

$$\text{Cov}(e_i, e_j) = -\sigma^2 h_{ij}$$

Esta não independência tem poucos efeitos sobre seu uso para investigar a adequabilidade do modelo (Montgomery & Peck, 1983).

A suposição de não correlação dos erros às vezes pode ser verificada através dos resíduos, mas nem sempre é fácil, uma evidência pode ser seqüências grandes de resíduos com o mesmo sinal. *Outliers* podem ser detectados pois em muitos casos estas observações apresentam resíduos muito grande.

Existem vários procedimentos gráficos envolvendo os resíduos que potencialmente revelam não normalidade e heterocedasticidade dos erros. Destacam-se alguns gráficos comentados a seguir.

- O gráfico dos resíduos versus cada regressor X_j .
- O gráfico dos resíduos versus $\hat{y}_i$, o valor estimado da variável dependente.
- O gráfico dos resíduos ao longo do tempo.

Todos estes gráficos devem apresentar uma variação aleatória em torno do zero (eixo x). Padrões diferentes deste merecem análise.

A.2.2.3 Testes de hipóteses

Estes testes são apropriados para cada suposição do modelo. Há vários testes para se detectar a não normalidade, são os testes de aderência que usam a estatística qui-quadrado ou os testes não paramétricos como o Kolmogorov-Smirnov.

Existem também vários testes de homogeneidade de variâncias, o teste de Cochran, o de Hartley, o de Bartlett e vários para se detectar a correlação dos erros, teste do sinal, teste de Durbin Watson.

O leitor interessado nestes testes pode consultar Draper & Smith (1981), Montgomery & Peck (1983) entre outros.

A.3 Multicolinearidade: definição, fontes, efeitos e detecção

Freitas (1988) descreve com detalhes o conceito de multicolinearidade, as fontes, os efeitos, várias medidas de detecção e várias alternativas de estimação diante de forte multicolinearidade. Os principais aspectos são descritos a seguir.

A.3.1 O conceito de multicolinearidade

No estudo da estimação dos parâmetros do modelo de regressão, merece destaque o comportamento das variáveis preditoras.

Os estimadores de mínimos quadrados são obtidos calculando-se: $\hat{a} = (X'X)^{-1} X'Y$, quando algumas ou todas as variáveis são perfeitamente colineares, diz-se estar diante da multicolinearidade perfeita, situação em que não se pode calcular os estimadores acima porque $X'X$ é singular, portanto não inversível.

Encontra-se na literatura definições de multicolinearidade que descrevem a situação encontrada na prática onde não se observa a multicolinearidade perfeita em virtude de coletar-se amostras da população:

Diz-se existir multicolinearidade, se para algum $\eta \geq 0$ existir um vetor c , $c = (c_0, c_1, \dots, c_p)$ de elementos não todos nulos, tal que:

$$\sum_{j=1}^p c_j X_j = \delta \text{ com } \|\delta\| \leq \eta \|c\| \quad (\text{Freitas, 1988})$$

O termo multicolinearidade também é utilizado quando as variáveis são correlacionadas, ou seja, $X'X$ é quase singular. Nesta situação pode-se calcular o estimador de mínimos quadrados, contudo com sérios prejuízos das propriedades estatísticas.

Pode-se dizer que todo conjunto de dados possui multicolinearidade, a diferença está na intensidade. Usando a decomposição espectral, ou decomposição em valores singulares de X verifica-se facilmente os efeitos sobre os estimadores e demais estatísticas envolvidas na análise de regressão e pode-se concluir facilmente que a multicolinearidade é um problema de grau e não uma questão de presença ou ausência (Freitas, 1988).

A.3.2 As fontes de multicolinearidade

É importante compreender as diferenças entre as várias causas da presença da multicolinearidade, uma vez que a interpretação do modelo resultante está extremamente ligada a elas.

A multicolinearidade pode surgir por várias razões:

- A especificação do modelo.
- Propriedades inerentes à população.
- Deficiências no método de amostragem.

- Modelos mal definidos, ou seja, o número de observações é menor que o número de variáveis.
- A presença de *outliers*.

Freitas (1988) apresenta exemplos para cada uma destas situações.

A.3.3 Efeitos da presença de multicolinearidade

A presença de grau forte de multicolinearidade traz efeitos sérios sobre as propriedades estatísticas dos estimadores:

- os estimadores de mínimos quadrados tornam-se viesados,
- suas variâncias e covariâncias são aumentadas,
- a amplitude dos intervalos de confiança são aumentadas,
- o poder dos testes de hipóteses fica reduzido e
- alterados os valores de previsões.

Existem alguns meios de quantificar o grau de multicolinearidade, estas medidas servem como instrumentos de detecção e algumas fornecem informações sobre quais os regressores, ou variáveis explicativas estão envolvidos.

A.3.4 Detecção e medidas do grau de multicolinearidade

Freitas (1988) apresenta diversas medidas para o grau de multicolinearidade. Destaca-se aqui as mais comuns e de simples interpretação.

- **A matriz $X'X$**

Uma investigação simples e bastante esclarecedora é obtida através do exame dos elementos de $X'X$, quando X está padronizada (cada variável encontra-se subtraída da média e dividida pelo respectivo desvio padrão) esta matriz coincide com matriz de correlação amostral, portanto um valor próximo de 1 indica forte evidência de relação linear entre duas variáveis explicativas.

É interessante notar que se 3 ou mais regressores estiverem relacionados linearmente, na mesma equação, não significa necessariamente que as correlações duas a duas sejam próximas de 1 e portanto este tipo de multicolinearidade não é detectada pelas correlações simples.

- **Os autovalores e autovetores de $X'X$**

O exame dos autovalores e autovetores de $X'X$ descreve eficazmente a estrutura de multicolinearidade no modelo de regressão linear.

Na ausência total de multicolinearidade $X'X$ é ortogonal, ou seja, $X'X = \theta I$ onde I é a matriz identidade. Sendo assim os autovalores são todos iguais. Mas não há conjunto de dados que reproduza exatamente esta situação, uma vez que, na prática, utiliza-se amostras no lugar de toda a população. Contudo esta

informação é valiosa para detectar-se o grau de multicolinearidade tendo-a como parâmetro de comparação ao examinar-se os autovalores de $X'X$.

Seja Γ_i um autovetor de $X'X$ correspondente ao autovalor λ_i , então $X\Gamma_i = \lambda_i\Gamma_i$. Se λ_i for próximo de zero, então $X\Gamma_i \cong 0$, indicando a existência de uma combinação linear entre os regressores.

O interessante é que o autovetor correspondente ao autovalor próximo de zero descreve completamente a relação entre as variáveis explicativas, os elementos de Γ_i mostram a estrutura de multicolinearidade existente em X , através da relação $X\Gamma_i \neq 0$.

- **O número da condição espectral**

É comum utilizar-se a raiz quadrada da razão entre o maior e o menor autovalor como indicadora do grau de multicolinearidade. Esta razão é conhecida como o número da condição espectral e é denotada por K :

$$K = \left(\frac{\max(\lambda_i)}{\min(\lambda_i)} \right)^{1/2}$$

Esta é a medida conhecida como a mais estável, deve ser igual a 1 na total ausência de multicolinearidade e valores próximos de 100 (cem) são considerados como evidências convincentes da existência de forte grau de multicolinearidade.

- O fator de inflação de variância - VIF

A matriz de variância e covariância do vetor de estimadores de mínimos quadrados de a em $Y = Xa + e$ é denotada pela letra *sigma* e é dada por:

$$\sum \hat{a} = \sigma^2 (X'X)^{-1}$$

Os elementos da diagonal da matriz são as variâncias dos estimadores e podem ser escritos como:

$$\text{var}(a_j) = \sigma^2 \frac{1}{1 - r_j^2}$$

onde:

r_j^2 é o coeficiente de determinação obtido da regressão múltipla de X_j sobre os demais X_k , $k \neq j$.

$\frac{1}{1 - r_j^2}$ é o VIF_j (*Variance Inflation Factor*) fator de inflação da variância.

Quando r_j^2 é próximo de 1, existe forte evidência de relação linear entre os regressores descrita por este modelo de regressão e conseqüentemente o VIF_j torna-se muito grande. Vifs maiores que 10 merecem atenção, tais valores causam um aumento muito grande na $\text{var}(a_j)$ daí o nome de fator de inflação da variância.

- **Decomposição espectral (valores singulares) de X e de $X'X$**

Uma matriz arbitrária A , de dimensão $n \times p$, pode ser decomposta em seus valores singulares. Esta decomposição é chamada de decomposição espectral e é dada por:

$$A = USV'$$

onde:

U , uma matriz $n \times r$, ortogonal;

V , matriz $p \times r$, ortogonal;

S é uma matriz $r \times r$, diagonal, de elementos não negativos.

Se $n \geq p$ e $r \leq p$ pode-se escrever:

$$X = HD \Gamma'$$

onde:

H : é uma matriz $n \times p$, tal que suas colunas são os autovetores normalizados de XX' , associados aos p autovalores diferentes de zero. Portanto $H'H = I$.
Convém ressaltar que estes p autovalores são os mesmos de $X'X$.

Γ : é $p \times p$, tal que suas colunas são os autovetores normalizados de $X'X$, associados aos p autovalores diferentes de zero: $\Gamma'\Gamma = I$

D : é uma matriz diagonal $p \times p$, cujos elementos são as raízes quadradas dos autovalores de $X'X$, diferentes de zero, chamados de valores singulares de X ,

em alguns casos denota-se D por $\Delta^{1/2}$, onde Δ é a matriz diagonal dos autovalores de $X'X$.

Decomposição espectral de $X'X$ na descrição da matriz de variância e covariância dos estimadores

$(X'X)^{-1}$ pode ser decomposta como:

$$(X'X)^{-1} = \sum_{i=1}^p \frac{\Gamma_i \Gamma_i'}{\lambda_i}$$

A matriz de var e covar dos estimadores de mínimos quadrados do modelo de regressão pode ser escrita como:

$$\sum \hat{a} = \sigma^2 (X'X)^{-1}$$

$$\text{então: } \text{var}(\hat{a}_j) = \sigma^2 \left(\sum_{i=1}^p \frac{\Gamma_{ji}^2}{\lambda_i} \right) = \sigma^2 \left(\frac{\Gamma_{j1}^2}{\lambda_1} + \dots + \frac{\Gamma_{jp}^2}{\lambda_p} \right)$$

Subdividindo-se ambos os membros por $\sigma^2 \text{VIF}_j$, obtém-se:

$$1 = \frac{\Gamma_{j1}^2}{\lambda_1 \text{VIF}_j} + \dots + \frac{\Gamma_{jp}^2}{\lambda_p \text{VIF}_j}$$

Portanto cada parcela desta soma reflete a contribuição percentual de cada autovalor na variância dos estimadores, quanto menor o autovalor em relação aos demais maior é a sua contribuição.

Decomposição espectral de X na descrição dos estimadores de mínimos quadrados.

No modelo de regressão linear, a decomposição espectral de X é muito conveniente pois ao expressarmos os estimadores de mínimos quadrados é fácil ver os termos inflacionados pela presença de multicolinearidade. Freitas (1988) descreve em detalhes a obtenção da expressão para os estimadores a partir da decomposição espectral de X. Obtém-se:

$$\hat{a} = \sum_{i=1}^p c_i \frac{\Gamma_i}{\lambda_i}$$

onde c_i é um escalar.

Observe que na expressão acima identifica-se facilmente as parcelas grandes quando os λ_i são próximos de zero.

A.4 Outliers: definição, causas, detecção

Várias são as definições e terminologia para descrever-se pontos que diferem dos demais, sendo considerados como observações atípicas.

Draper & Smith (1981) sugerem que se considere como *outlier* ou observação influente ou ainda discrepante, aquela cujo resíduo esteja 3 ou 4 desvios padrões distante da média dos demais resíduos. Outros autores fazem distinção entre observações discrepantes e muito discrepantes (SEADE).

A presença de observações discrepantes podem ser provenientes de erros humanos ou de aparelhos, como por exemplo, na leitura ou digitação dos dados, na calibração de aparelhos, defeitos de aparelhos, ou ainda podem ser observações mais raras, contudo passíveis de ocorrerem.

- **Exame da matriz H**

A matriz chapéu, $H = X(X'X)^{-1}X'$ desempenha um papel importante na detecção de *outliers*.

O posto da matriz H é igual ao posto da matriz X e é dado por:

$$\text{posto}(X) = \text{posto}(H) = \sum_{i=1}^n h_{ii} = p$$

Espera-se que em média cada elemento da diagonal seja p/n . Então observações com $h_{ii} > 2p/n$ merecem atenção.

- **Exame dos resíduos *studentizados***

Resíduos *studentizados* onde na estimação da variância elimina-se a i -ésima observação, com valor absoluto maior que 2 devem ser examinados.

- **A matriz de var e covar dos estimadores dos parâmetros**

A razão entre os determinantes das matrizes de variância e covariância dos estimadores ($\hat{\sigma}^2 = \sigma^2(X'X)^{-1}$) é obtida fazendo-se no numerador o cálculo

sem a i -ésima observação. No processamento através do SAS esta estatística recebe o nome de covratio e merecem atenção valores tais que:

$$|\text{covratio} - 1| \geq 3p / n$$

- **DFFITS**

A diferença padronizada entre os valores estimados da variável dependente com e sem a i -ésima observação é conhecida como dffits e é dada por:

$$\frac{\hat{y}_i - \hat{y}(i)}{s(i)\sqrt{h_i}}$$

onde:

$$h_i = \mathbf{x}_i'(\mathbf{X}'\mathbf{X})^{-1}\mathbf{x}_i,$$

$s(i)$ é o desvio padrão dos erros estimado sem a i -ésima observação,

$\hat{y}(i)$ é o estimador de y sem a i -ésima observação.

dffits maiores que 2 merecem atenção.

- **DFBETAS**

A diferença padronizada entre as estimativas dos parâmetros com e sem a i -ésima é conhecida como Dfbetas e é dada por:

$$\frac{(a_j - a_j(i))}{s(i)\sqrt{(X'X)^{jj}}}$$

onde:

$(X'X)^{jj}$ é o elemento j, j de $(X'X)^{-1}$,

$s(i)$ é o desvio padrão dos erros, estimado sem a i -ésima observação,

$a_j(i)$ é a estimativa do j -ésimo parâmetro sem a i -ésima observação.

Valores de $dfbetas$ maiores que 2 merecem atenção.

A.5 Comparação e escolha entre modelos com e sem intercepto

É comum escolher-se modelos de regressão com ou sem o termo constante, o intercepto, baseando-se no valor do coeficiente de determinação, o r^2 , ou seja, ajusta-se os dois modelos e escolhe-se o de maior r^2 .

Estes valores de r^2 não são comparáveis, uma vez que o r^2 é dado pelo quociente entre a soma de quadrados da regressão e a soma de quadrados total, e estas somas de quadrados são calculadas de forma diferente para os modelos com e sem intercepto. As principais diferenças estão resumidas na tabela a seguir (Searle, 1971; Uyar & Erdem, 1990; Kvalseth, 1985; Hawkins, 1980).

Comparação entre os modelos de regressão com e sem intercepto.

soma de quadrados:	COM intercepto	SEM intercepto (4)
sqr ₍₁₎	$\alpha' X'Y - n\bar{y}^2 = \sum_{i=1}^n (\bar{y}_{ii} - \hat{y}_i)^2$	$\alpha' X'Y = \sum_{i=1}^n (\hat{y}_i)^2$
sqe ₍₂₎	$Y'Y - \alpha' X'Y = \sum_{i=1}^n (y_i - \hat{y}_i)^2$	$Y'Y - \alpha' X'Y = \sum_{i=1}^n (y_i - \hat{y}_i)^2$
sqt ₍₃₎	$Y'Y - n\bar{y}^2 = \sum_{i=1}^n (y_i - \bar{y})^2$	$Y'Y = \sum_{i=1}^n (y_i)^2$
$r^2 = 1 - \frac{sqe}{sqt} = \frac{sqr}{sqt}$	$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2}{\sum_{i=1}^n (y_i - \bar{y})^2}$	$\frac{\sum_{i=1}^n (\hat{y}_i)^2}{\sum_{i=1}^n (y_i)^2}$
$F = \frac{sqr / p}{sqe / (n - p - 1)}$	$\frac{\sum_{i=1}^n (\hat{y}_i - \bar{y})^2 / p}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / n - p - 1}$	$\frac{\sum_{i=1}^n (\hat{y}_i)^2 / p}{\sum_{i=1}^n (y_i - \hat{y}_i)^2 / n - p - 1}$
$\sum_{i=1}^n e_i$	$= 0$	$\neq 0, \sum_{i=1}^n X_i e_i = 0(5)$

(1) sqr = soma de quadrados da regressão.

(2) sqe = soma de quadrados do erro ou resíduo.

(3) sqt = soma de quadrados total.

(4) Searle, 1971 e Kvalseth, 1985.

(5) Neter, Wasserman e Kutner, 1983.

Observa-se na tabela acima que tanto na soma de quadrados da regressão como na soma de quadrados total, somente no modelo com intercepto, existe a subtração do termo $\bar{y}$, por este motivo a literatura cita estas somas de quadrados como sendo “corrigidas pela média”.

Tanto os valores de r^2 como os do teste F-Snedecor são calculados de forma diferente, não se pode basear a escolha de um ou outro modelo por nenhuma destas estatísticas. Kvalseth (1985) e Montgomery & Peck (1983) ressaltam que se esta redefinição das somas de quadrados não for feita, ou seja mantendo-se as somas de quadrados corrigidas pela média, é comum obter-se algumas contradições do tipo: um valor de r^2 negativo para o modelo sem intercepto; o quadrado médio do resíduo (soma de quadrados do resíduo dividida por $n-p-1$) do modelo com intercepto menor que o quadrado médio do resíduo do modelo sem intercepto e contudo um valor de r^2 do modelo com intercepto menor que o r^2 do modelo sem intercepto.

Entre as recomendações para os devidos cuidados ao se utilizar o modelo sem intercepto, destaca-se o fato de previamente escolher-se se haverá ou não o termo constante no modelo linear. O modelo sem intercepto só deve ser escolhido se houver razões teóricas suficientemente fortes, advindas da experiência do pesquisador sobre o fenômeno (Kvalseth, 1985; Montgomery & Peck, 1983).

Em caso de dúvidas sobre o fenômeno em estudo, quanto ao embasamento teórico para justificar-se a ausência do intercepto, Montgomery & Peck (1983) sugerem:

- Fazer um diagrama de dispersão dos dados, quando possível, para verificar a existência de evidências que os pontos passem pela origem.
- Ajustar ambos os modelos com e sem intercepto, e escolher o que ofereça melhor ajuste, identificando-se como o modelo melhor ajustado o que oferecer menor quadrado médio do resíduo (soma de quadrados do resíduo dividido pelo número de graus de liberdade, $n-p-1$).
- Testar a hipótese de que o intercepto seja zero no modelo ajustado com intercepto. Caso esta hipótese não seja rejeitada há fortes evidências para admitir-se o ajuste do modelo sem intercepto.

Geralmente quando o *software* oferece a opção de escolha da presença ou não do intercepto no modelo ele redefine a forma de calcular-se estas somas de quadrados, o que passa despercebido pelo usuário, gerando a comparação indevida dos resultados obtidos nos modelos com e sem intercepto. O SAS, Statistical Analysis Systems, imprime uma mensagem junto à tabela de análise de variância da regressão informando a redefinição do cálculo do r^2 (Hawkins, 1980; SAS v.1 e v. 2, 1989).

A.6 Intervalo de confiança para $E(Y/X=x)$ e intervalo de previsão de Y

É muito importante saber se deseja-se estimar o valor médio de Y dado a ocorrência de $X=x$ ($E(Y/X=x)$) ou o valor de Y para uma única unidade amostral onde $X=x$.

Por exemplo: imagine que se utilizou o modelo $\hat{Y} = a + bX$ para estimar o tempo de reação a um determinado estímulo (Y) em função da idade (X).

Há diferença entre desejar estimar o tempo médio de reação aos 28 anos e o tempo de reação de uma pessoa de 28 anos.

No primeiro caso, utiliza-se o intervalo de confiança para $E(Y/X=x)$ e no segundo caso o intervalo de previsão para Y quando $X=x$ (Bussab, 1986).

Nos intervalos abaixo X é toda a matriz de planejamento portanto tem dimensão $n \times p$, n observações originais em p variáveis e x é um vetor de uma observação em p variáveis, portanto de dimensão $p \times 1$.

- **Intervalo de Confiança para $E(Y_i/X=x_i)$**

A variância de $\hat{y}$ é dada por: $x'(X'X)^{-1}x\sigma^2$

Portanto o intervalo é: $\hat{y}_i \pm \sqrt{x'(X'X)^{-1}x(qme)} \cdot t_{n-p-2}$

onde:

X: é a matriz de planejamento e x é o vetor específico da i-ésima observação.

qme: é o quadrado médio do erro para estimar a variância de Y.

$X=x$: pode ou não pertencer a amostra original através da qual ajustou-se o modelo.

Prova-se que $\sum_{i=1}^n \frac{V(\hat{y}_i)}{n} = \text{traço} \{X(X'X)^{-1}X'\} \frac{\sigma^2}{n} = p \cdot \frac{\sigma^2}{n}$ (Draper & Smith, 1981).

- **Intervalo de Previsão de Y_i para $X_i=x$**

Normalmente é utilizado quando o valor de $X=x$ não faz parte da amostra que originou o modelo e deseja-se prever o valor da variável dependente nesta nova observação.

O que muda em relação ao intervalo anterior é o erro da previsão. O intervalo passa a ser:

$$\hat{y} \pm \sqrt{(x'(X'X)^{-1}x + 1)\sigma^2} \cdot t_{n-p-2}$$

Para entender esta alteração, usaremos o modelo onde $p=2$: $\hat{y} - Y = (a + bX)$ para estimar $Y = \alpha + \beta X + e$.

O erro da previsão é: $\hat{y} - Y = (a + bX) - (\alpha + \beta X + e) = (a - \alpha) + (b - \beta)X - e$.

Observe que: $E(\hat{y}) = E(Y)$ já que $e \sim N(0, \sigma^2)$, no entanto $E(\hat{y}) \neq Y$.

$\hat{y}$ e Y são independentes em função da hipótese de independência entre os erros. Sendo assim:

$$\text{Var}(\hat{y} - Y) = \text{Var}(\hat{y}) + \text{Var}(Y) = x'(X'X)^{-1}x\sigma^2 + \sigma^2 = (x'(X'X)^{-1}x + 1)\sigma^2.$$

ANEXO B - BOX PLOT

Bloco A

chefes

Observa-se a variação da distribuição do número de viagens, como um todo, para chefes de acordo com as diversas estruturas familiares, destacando-se como de maior mobilidade os chefes de famílias compostas por casais com filhos e agregados (famílias nucleares completas e ampliadas).

Bloco B-1

chefs

Nesta série de *box plot* observa-se a variação proveniente da diferença entre os estágios no ciclo da vida das famílias de chefes de famílias nucleares simples (casais sem filhos): chefes de famílias jovens tem maior mobilidade que chefes de famílias adultas ou velhas.

Bloco B-2

chefes

Nº de viagens individual

Nesta série de *box plot* observa-se a variação proveniente da diferença entre os estágios no ciclo da vida das famílias de chefes de famílias nucleares completas (casais com filhos): chefes de famílias jovens tem maior mobilidade seguidos por chefes de famílias adultas e depois velhas.

Bloco B-3

chefes

Nº de viagens individual

Nesta série de *box plot* observa-se a variação proveniente da diferença entre os estágios no ciclo da vida das famílias de chefes de famílias nucleares completas e ampliadas: chefes de famílias jovens e adultas tem maior mobilidade que chefes de famílias velhas.

Bloco B-4

chefes

Nesta série de *box plot* observa-se a variação proveniente da diferença entre os estágios no ciclo da vida das famílias de chefes de famílias quebradas (sem um dos cônjuges): chefes de famílias jovens tem maior mobilidade que chefes de famílias adultas seguidos por chefes de famílias velhas.

Bloco B-5

chefes

Nº de viagens individual

Nesta série de *box plot* observa-se a variação proveniente da diferença entre os estágios no ciclo da vida das famílias de chefes de grupos conviventes, chefes de famílias jovens tem maior mobilidade que chefes de famílias adultas ou velhas.

Bloco B-6

chefes

Nº de viagens individual

Nesta série de *box plot* observa-se a variação proveniente da diferença entre os estágios no ciclo da vida das famílias de pessoas que moram só, se forem jovens têm maior mobilidade.

Bloco C

chefs

Nesta série de *box plot* observa-se a variação proveniente somente da diferença entre os estágios no ciclo da vida das famílias dos chefes, observando-se novamente que chefes de famílias jovens tem maior mobilidade que chefes de famílias adultas ou velhas.

Bloco D

cônjuges

Nesta série de *box plot* observa-se a variação proveniente somente da diferença entre os estágios no ciclo da vida das famílias dos cônjuges, cônjuges de famílias jovens e adultas tem maior mobilidade que os de famílias velhas.

Bloco E

filhos

Nesta série de *box plot* observa-se a variação proveniente somente da diferença entre os estágios no ciclo da vida das famílias dos filhos, filhos de famílias adultas tem maior mobilidade que os de famílias jovens ou velhas.

Bloco F-1

salário individual < 3 salários mínimos

Quando se leva em conta esta faixa de renda individual, a mais baixa, não se pode determinar diferenças na mobilidade devido a variação no estágio do ciclo da vida da família.

Bloco F-2

$3 < \text{salário individual} \leq 5$
(salários mínimos)

Nº de viagens individual

Quando se leva em conta esta faixa de renda individual, pode-se determinar diferenças na mobilidade devido a variação no estágio do ciclo da vida da família, indivíduos de famílias adultas tem maior mobilidade.

Bloco F-3

$5 < \text{salário individual} \leq 10$
(salários mínimos)

Nº de viagens individual

Quando se leva em conta esta faixa de renda individual, pode-se determinar diferenças na mobilidade devido a variação no estágio do ciclo da vida da família, indivíduos de famílias adultas e jovens tem maior mobilidade.

Bloco F-4

10 < salário individual ≤ 15
(salários mínimos)

Nº de viagens individual

Quando se leva em conta esta faixa de renda individual, pode-se determinar diferenças na mobilidade devido a variação no estágio do ciclo da vida da família, indivíduos de famílias adultas e jovens tem maior mobilidade.

Bloco F-5

15 < salário individual ≤ 20
(salários mínimos)

Nº de viagens individual

Quando se leva em conta esta faixa de renda individual, pode-se determinar diferenças na mobilidade devido a variação no estágio do ciclo da vida da família, indivíduos de famílias adultas e jovens tem maior mobilidade.

Bloco F-6

salário individual > 20 salários mínimos

Nº de viagens individual

Quando se leva em conta esta faixa de renda individual, pode-se determinar diferenças na mobilidade devido a variação no estágio do ciclo da vida da família, indivíduos de famílias adultas e jovens tem maior mobilidade e note o aumento na mobilidade de indivíduos de famílias velhas em relação as demais faixas de renda.

Bloco G-1

nível de instrução <4ª série do 1º. grau

Nº de viagens individual

Quando se leva em conta este nível de instrução individual, pode-se determinar diferenças na mobilidade devido a variação no estágio do ciclo da vida da família, indivíduos de famílias adultas e jovens tem maior mobilidade.

Bloco G-2

8^a. série ≤ instrução < 2^o. grau

Nº de viagens individual

Quando se leva em conta este nível de instrução individual, pode-se determinar diferenças na mobilidade devido a variação no estágio do ciclo da vida da família, indivíduos de famílias adultas tem maior mobilidade.

Bloco G-3

2º. grau $\leq$ instrução < superior

Quando se leva em conta este nível de instrução individual, pode-se determinar diferenças na mobilidade devido a variação no estágio do ciclo da vida da família, indivíduos de famílias adultas e jovens tem maior mobilidade.

Bloco G-4

nível de instrução $\geq$ superior

Quando se leva em conta este nível de instrução individual, pode-se determinar diferenças na mobilidade devido a variação no estágio do ciclo da vida da família, indivíduos de famílias adultas e jovens tem maior mobilidade. Observe o aumento geral da mobilidade em relação aos demais níveis considerados.

REFERÊNCIAS BIBLIOGRÁFICAS

ASH, SOLOMON E. *Psicologia social*. Trad. de Dante Moreira Leite e Miriam Moreira Leite. São Paulo, Cia. Ed. Nacional, 4ª edição, 1977.

BALDWIN, Alfred L. *Teorias de desenvolvimento da criança*. Trad. de Dante Moreira Leite. São Paulo, Livraria Pioneira Editora, 2ª edição, 1980.

BEVERIDGE, William I. B. *Sementes da descoberta científica*. Trad. de S. R. Barreto. São Paulo, T. A. Queiroz Editor Ltda e Editora da Universidade de São Paulo, 1981.

BODMER, Milena. *Análise dos elementos comportamentais da demanda de transportes*. Rio de Janeiro, 1984. 170p. Tese (Doutorado) - Universidade Federal do Rio de Janeiro.

BRUSCHINI, Cristina. Estrutura familiar e trabalho na grande São Paulo. *Cadernos de Pesquisa São Paulo - SEADE*, n.72, p.39-57, fev 1990.

BRUSCHINI, Cristina. Mulher e trabalho: uma avaliação da década da mulher (1975-1985). *Rev. Fund. SEADE / SP em Perspectiva*, v.2, n.1, p.41-48, jan/abr 1986.

BRUTON, Michael J. *Introduction to transportation planning*. London, UCL Press: University College London, 1993.

BUSSAB, Wilton de Oliveira *Análise de variância e de regressão*, Atual, 1986.

BUSSAB, Wilton de Oliveira; WAGNER, Eugênia Sales Indicador de ciclo de vida familiar: uma construção alternativa. *São Paulo em Perspectiva: Revista de Fundação SEADE*, v.8, n.2, p.92-97, abr/jun 1994.

CARPENTER, Susan; JONES, Peter *Recent advances in travel demand analysis*. Great Britain Gower, 1983.

CHATTERJEE, Samprit; PRICE, Bertran. *Regression analysis by example*. New York, John Wiley & Sons, 1977.

CHINOY, Ely. *Sociedade - uma introdução à Sociologia*. Trad. de Octavio Mendes Cajado. São Paulo, Ed. Cultrix, 4ª edição, 1975.

CIA DO METROPOLITANO DE SÃO PAULO. *Pesquisa OD/87 Síntese das Informações*. São Paulo, Cia do Metropolitano São Paulo - Metrô, 1990.

CLARKE, M. I.; DIX, M. C.; JONES, P. M.; HEGGIE, I. G. Some recent developments in activity travel analysis and modelling. *Transportation Research Record*, n.794, p.1-8, 1981.

DACHS, José Roberto W.; CARVALHO, José Ferreira de *Diagnóstico em regressão*. In: SIMPÓSIO NACIONAL DE PROBABILIDADE E ESTATÍSTICA, 6. Rio de Janeiro, 1984.

DALY, Andrew Applicability of disaggregate models of behaviour: a question of methodology. *Transportation Research A*, v.16A, n.5-6, p.363-370, sept./nov. 1982.

DALY, A. J.; ORTÚZAR, J. De D. Forecasting and data aggregation: theory and practice. *Traffic Engineering and Control*, v. 31, n. 12, p. 632-643, dec. 1990.

DEY, Soumya; FRICKER, Jon D. Bayesian updating of trip generation data: combining national trip generation rates with local data. *Transportation*, v.21, n. 4, p. 394-403, 1994.

DOUGLAS, A. A. Home based trip end models-a comparison between category analysis and regression analysis procedures. *Transportation*, v. 2, p. 53-70, 1973.

DOUGLAS, A. A.; LEWIS, R. J. Trip generation techniques: category analysis and summary of trip generation techniques. *Traffic Engineering and Control*, v. 12, n. 10, p. 532-535, fev. 1971-a.

_____ Trip generation techniques: household least squares regression analysis. *Traffic Engineering and Control*, v. 12, n. 9, p. 477-479, jan. 1971-b.

_____ Trip generation techniques: zonal least squares regression analysis. *Traffic Engineering and Control*, v. 12, n. 8, p. 428-431, dec. 1970-a.

_____ Trip generation techniques:introduction. *Traffic Engineering and Control*, v. 12, n. 7, p. 362-365, nov. 1970-b.

DRAPER, N.R.; SMITH, H. *Applied regression analysis*. New York, John Wiley & Sons, 1981.

FLACSO: Faculdade Latino Americana de Ciencias Sociales. *Brasil - Mulheres Latinas Americanas em Dados*. Instituto de la Mujer - Ministerio de Asuntos Sociales de Espana, 1993.

FOX, Michael. Transport planning and the human activite approach. *Journal of Transport Geography*, v.3, n.2, p.105-116, 1995.

FREITAS, Maria da Conceição Farias. *Estimação de β no modelo de regressão linear múltipla na presença de multicolinearidade*. São Paulo, 1988. 119 p. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo.

GEORGE, Linda K. Sociological perspectives on life transitions. *Annual Review of Sociology*, v. 19, p. 353-373, 1993.

GUIMARÃES, Iracema Brandão. Características de um mercado de trabalho e participação do grupo familiar. *Caderno do CRH: Centro de Recursos Humanos da Universidade Federal da Bahia*, Fator Ed., n.14, p.31-46, jan/jul 1991.

HAGA, Atsuko. A participação atual da mulher no mercado de trabalho. *São Paulo em Perspectiva*. São Paulo - SEADE, v.2, n.1, p.5-16, jan/abr 1986.

HAITOVSKY, Y. *Regression estimation from grouped observations*. London, Charles Griffin & Company Limited, 1973.

HANSON, S.; HANSON, P. The travel activity patterns of urban residents; dimensions and relations to sociodemographic characteristics. *Economic Geography*, v.57, n.4, p.332-347, oct. 1991.

HAWKINS, Douglas M. A note on fitting a regression without an intercept term. *The American Statistician*, v. 34, n.4, p. 233, nov. 1980.

HEGEMBERG, Leônidas. *Etapas da investigação científica*, v. 2. São Paulo, Ed. Pedagógica e Universitária e Ed. Da Universidade de São Paulo, 1976.

HEMPEL, Carl G. *Filosofia da ciência natural*. Trad. de P. S. Rocha. Rio de Janeiro, Zahar Editores, terceira edição, 1981.

HERENDEEN, James H.; MEMBER, Jr.; SCHELLINGER, David J.; SUNG, Myung Hak. Intercity travel demand forecasting for the Windsor-Quebec city corridor. *Proceedings of the First International Conference on High Speed Ground Transportation Systems*. New York, p.276-288, 1993.

HOCKING, R. R. The analysis and selection of variables in linear regression. *Biometrics*, v. 32, n. 1, p. 1-49, mar. 1976.

HOROWITZ, Joel L. Evaluation of discrete choice random utility models as practical tools of transportation systems analysis. *Transportation Research Board*, SR, 201, 1983.

KANAFANI, Adib. *Transportation demand analysis*. New York, Mc Graw-Hill book Company, 1983.

KAWAMOTO, Eiji *Um novo enfoque do processo de escolha em transporte com tratamento baseado na psicofísica multidimensional*. São Paulo, 1987. 126p. Tese (Doutorado) - Escola de Engenharia de São Carlos, Universidade de São Paulo.

KAWAMOTO, Eiji. *Análise de Sistemas de Transportes*. São Carlos, EESC-USP, 1993.

KAWANO, Masaya; CHISHAKI, Takeshi; YOSHITAKE, Tetsunobu; TSUTSUMI, Masafumi. Analysis of trip generation and trip production. *Memoirs of the Faculty of Engineering*, v. 50, n.3 p.201-215, sept. 1990.

KHISTY, C. Jotin. *Transportation engineering: an introduction*. New Jersey, Prentice Hall, 1990.

KISH, Leslie. *Survey Sampling*. New York, John Wiley & Sons, 1965.

KITAMURA, R. Life style and travel demand In: *Transportation Research Board*, sr 220, p.149-189, 1988.

KLAUSNER, Samuel Z. Social sciences as conceptual resource for transportation research. *Transportation Research Board - Special Report 201 - Travel Analysis Methods for the 1980s*, p.74-82, 1983.

KOLLO, Hanna P. H.; PURVIS, Charles L. Regional travel forecasting model system for the San Francisco bay area. *Transportation Research Record* 1220, p. 58-65, 1989.

KUMARAGE, Amal S.; WIRASINGHE, S. C. Transferability of aggregate intercity total demand models. *Canadian Journal of Civil Engineering*, v. 22, n. 2, p. 283-291, apr. 1995.

KUNERT, Uwe. Weekly mobility of life cycle groups *Transportation*, v.21, n. 3, p. 271-288, aug. 1994.

KUTTER, Eckhard. A model for individual travel behaviour. *Urban Studies*. v. 10, n. 2, p. 235-258, jun. 1973.

KVALSETH, Tarald O. Cautionary note about r^2 . *The American Statistician*, v. 39, n. 4, p. 279-285, nov. 1985.

LEWIS, R. J. Comments on regression techniques in trip generation analysis. *Proceedings of PTRC: Planning and Transport Research and Computation Seminar on Urban Traffic. Model Research*, jun./jul., 1970.

McDONALD, Kathie G.; STOPHER, Peter R. Some contrary indications for the use of household structure in trip generation analysis. *Transportation Research Record 944*, p.92-100, 1983.

MONTALI, Lília. Arranjos familiares: o esforço coletivo para viver na grande São Paulo. *Caderno de Pesquisa, SEADE*. São Paulo n.72, p. 58-69, fev. 1990.

MONTALI, Lília. Família e trabalho na conjuntura recessiva. *São Paulo em Perspectiva*. São Paulo - SEADE, v.5, n.1, p. 72-84, jan/mar 1991.

MONTGOMERY, Douglas C.; PECK, Elizabeth A. *Introduction to linear regression analysis*. New York, John Wiley & Sons, 1983.

MUNSHI, Kaivan. Urban passenger travel demand estimation: a household activity approach. *Transportation Research A*, n. 6, p. 423-432, 1993.

NETER, John; WASSERMAN, William *Applied linear statistical models*. Illinois, Richard D. Irwin Inc. 1974.

NETER, John; WASSERMAN, William; KUTNER, Michael H. *Applied linear regression models*. Illinois, Richard D. Irwin Inc. 1983.

OPPENHEIM, Norbert. *Urban travel demand modeling: from individual choices to general equilibrium*. New York, John Wiley & Sons, 1995.

ORTÚZAR, Juan de Dios; WILLUMSEN, Luis G. *Modelling transport*. New York, John Wiley & Sons, 1994.

PAPACOSTAS, C. S. *Fundamentals of transportation engineering*. New Jersey, Prentice Hall International Inc., 1987.

PAS, Eric I.; KOPPELMAN, Frank S. An examination of the determinants of day to day variability in individual's urban travel behavior. *Transportation*, v. 14, p. 3-20, 1987.

POPPER, Karl R. *Conjecturas e refutações*. Trad. de Sérgio Bath. Brasília, Editora Universidade de Brasília, segunda edição, 1972.

_____. *Conhecimento objetivo*. Trad. de Milton Amado. Belo Horizonte, Ed. Itatiaia; São Paulo, Ed. da Universidade de São Paulo, 1975.

PSACHAROPOULOS, George; TZANNATOS, Zafiris. Female labor force participation: an international perspective. *Research Observer*. v.4, n.3, p.187-201, july 1989. The International Bank for Reconstruction and Development / The World Bank.

- RENGARAJU, V. R.; SATYAKUMAR, M. Structuring category analysis using statistical technique. *Journal of Transportation Engineering*, v. 120, n.6, p. 930-939, nov./dec. 1994.
- RICHARDSON, A. J. Transport Planning & Modelling: a twenty year perspective. *Australian Road Research*, v.20, n.1, p.9-21, mar. 1990.
- ROSANDER, A. C. *Case studies in sample design*. New York, Marcel Dekker Inc., 1977.
- SAID, Galal M.; YOUNG, David H. A general linear model framework for estimating work trip rates for households in Kuwait. *Transportation Research A*. v.24a, n.3, p.187-200, may 1990.
- SAID, M. Galal; YOUNG, David H.; IBRAHIM, Hassank Trip generation procedure for areas with structurally different socioeconomic groups. *Transportation Research Record 1328*, p.1-9, 1991.
- SAKA, Anthony A. Post calibration adjustment of travel demand models. *ITE Journal*, v.63, n.9, p.13-18, sept. 1993.
- SALOMON, I. Life styles - *A broader perspective on travel behavior*. In: CARPENTER, Susan; JONES, Peter. Recent advances in travel demand analysis. England, Gower, 1983.
- SAS Institute Inc. *SAS Language and Procedures*, Usage 1, Version 6, 1.ed., Cary NC: SAS Institute Inc., 1989.

SAS Institute Inc. *SAS Language and Procedures*, Usage 2, Version 6, 1.ed., Cary NC: SAS Institute Inc., 1991.

SAS Institute Inc. *SAS Language: reference*, Version 6, 1.ed., Cary NC: SAS Institute Inc., 1990.

SAS Institute Inc. *SAS Procedures Guide*, Version 6, 3.ed., Cary NC: SAS Institute Inc., 1990.

SAS Institute Inc. *SAS/STAT User's Guide*, Version 6, v.1, 4.ed., Cary NC: SAS Institute Inc., 1989.

SAS Institute Inc. *SAS/STAT User's Guide*, Version 6, v.2, 4.ed., Cary NC: SAS Institute Inc., 1989.

SEADE, Fundação Sistema Estadual de Análise de Dados *Pesquisa de condições de vida na região metropolitana de São Paulo - principais resultados*. São Paulo, 1992a

SEADE, Fundação Sistema Estadual de Análise de Dados *Pesquisa de condições de vida na região metropolitana de São Paulo - uma abordagem multissetorial*. São Paulo, 1992b

SEARLE, S. R. *Linear Models*. John Willey & Sons, 1971.

SEMPLA: Secretaria Municipal do Planejamento. Base de dados para o planejamento, São Paulo, 1992.

SHAO, Jun. Linear model selection by cross-validation. *Journal of the American Statistical Association - Theory and Methods*, v. 88, n. 442, p. 486-494, jun. 1993.

SIQUEIRA, Arminda Lúcia. *Uso de transformação em análise de variância e análise de regressão*. São Paulo, 1983. 154 p. Dissertação (Mestrado) - Instituto de Matemática e Estatística, Universidade de São Paulo.

SOUMYA, S. Dey; FRICKER, Jon D. Bayesian updating of trip generation data: combining national trip generation rates with local data. *Transportation*, v. 21, n. 4, p. 393-403, dec. 1994.

STARKIE, D., N., M.. The treatment of curvi-linearities in the calibration of trip end models. *Proceedings of PTRC: Planning and Transport Research and Computation Seminar on Urban Traffic Model Research*, p. 37-41, jun./jul., 1970.

STOPHER, Peter R.; McDONALD, Kathie G. Trip generation by cross-classification: an alternative methodology. *Transportation Research Record 944*, p.84-91, 1983.

SUPERNAK, Janusz; TALVITIE, Antti; De JOHN, Anthony. Person category trip generation model. *Transportation Research Record 944*, p.74-83, 1983.

SUPERNAK, Janusz; Transportation modelling: lessons from the past and tasks for the future *Transportation 12*, p.79-90, 1983.

SUPERNAK, Janusz; Transferability of person category trip generation model
*Proceedings of 9th PTRC: Planning and Transport Research and
Computation Summer Annual Meeting*, jul., 1981.

THE 1985 INTERNATIONAL CONFERENCE ON TRAVEL BEHAVIOUR,
Noordwijk, the Netherlands-Utrecht., 16-19 april 1985. *Behavioural
research for transport policy*. Utrecht, The Netherlands, VNU Science
Press, 1986.

TELLES, Vera da Silva. Família e trabalho: precariedade e pauperismo na
grande São Paulo. In: *VIII Encontro Nacional de Estudos Populacionais*,
1992, Brasília. *Anais...* Brasília: ABEP - Associação Brasileira de Estudos
Populacionais. São Paulo, v.2, Mulher e Saúde.

TURNER, Jeff; FOURACRE, Philip. Women and transport in developing
countries. *Transport Review*, v. 15, n. 1, p. 77-96, 1995.

TYE, N. B.; SHERMAN, L.; KINNUCAN, M.; NELSON, D.; TARDIFF, T.
Application of disaggregate travel demand models, *Transportation
Research Board-National Cooperative Highway Research Program Report*
253, dec. 1982.

UYAR, Bulent; EEDEM, Orhan. Regression procedures in SAS: problems?
The American Statistician, v. 44, n. 4, p. 296-310, nov. 1990.

van de BILT, Karin Anne. *Análises de taxas de produção de viagens urbanas
utilizando modelagem de segmentação*. São Paulo, 1997. 131 p.
Dissertação (mestrado) - Escola Politécnica, Universidade de São Paulo.

VUREN, Tom Van; GUNN, Hugh; DALY, Andrew. Disaggregate travel demand models: their applicability for british transport planning practice. *Traffic Engineering and Control*, v. 36, n.6, n. 336-344, jun. 1995.

WEISS, Robert E. The influence of variable selection: a bayesian diagnostic perspective. *Journal of the American Statistical Association - Theory and Methods*, v. 90, n. 430, p. 619-625, jun. 1995.

WHITE, Michael T. An examination of residual distributions in ordinary least squares (ols) household based trip generation models. *Transportation Research*, v. 10, p. 249-254, 1976.

WICHERN, Dean W.; JOHNSON, Richard A. *Applied Multivariate Statistical Analysis*. London, Prentice Hall International Inc., 1992.

WILMOT, Chester. Evidence on transferability of trip generation models. *Journal of Transportation Engineering*, n.5, p. 405-410, sep./oct. 1995.

WILSON, R. R.; MEGEED, A. E. T. ABL EL. Stability of province/state wide transport planning model. *Canadian Journal of Civil Engineering*. v. 17, n. 2, p. 192-197, apri 1990.

WOOTTON, H.J.; PICK, G. W. A model for trips generated by households. *Journal of Transportation Economy and Policy*, v. 1, p. 137-153, may 1967.

YANAGUAYA, Waldo. The use of U. K. Transport models in developing countries. *Traffic Engineering and Control*, v. 34, n. 10, p. 476-479, oct. 1993.

YU, Jason C. *Transportation engineering: introduction to planning, design, and operations*. New York, Elsevier North Holland Inc., 1982.

ZHENG, Xiaodong; LOH, Wei-Yin. Consistent variable selection in linear models. *Journal of the American Statistical Association - Theory and Methods*, v. 90, n. 429, p. 151-156, mar. 1995.