APLICAÇÃO DE TÉCNICAS DE ANÁLISE MULTIVARIADAS DE DADOS EM UM ESTUDO EXPLORATÓRIO DA INFLUÊNCIA DO USO DO SOLO NAS ESCOLHAS DE PADRÕES DE ENCADEAMENTO DE VIAGENS

Cira Souza Pitombo
Eiji Kawamoto
Universidade de São Paulo/ Escola de Engenharia de São Carlos

RESUMO
Este estudo é uma tentativa inicial de melhor compreender a relação entre padrões de uso do solo e o comportamento referente a viagens encadeadas. Através da aplicação de técnicas de Análise Multivariadas (AM), este trabalho investiga os fatores que influenciam a escolha dos destinos (zonas de tráfego caracterizadas pelo uso do solo) durante a cadeia de viagens dos indivíduos residentes na Região Metropolitana de São Paulo (RMSP). Trata-se de uma análise exploratória que visa responder algumas indagações. Neste trabalho, foram utilizadas duas técnicas de AM: (1) Análise de *Cluster*, objetivando agrupar e caracterizar as zonas de tráfego da RMSP e, (2) Árvore de Decisão, a fim de encontrar relações entre características socioeconômicas, atributos de uso do solo e escolhas de destinos. Os resultados indicam que o grau de atividade da zona de origem influencia a escolha dos padrões de viagens.

ABSTRACT
This study is an initial attempt to improve the understanding of relationships between land use patterns and the trip-chaining behavior. By using Multivariate Data Analysis (MA), this work investigates the factors that influence the destination choices (Traffic Zones (TZ) characterized by land use attributes) during the trip-chaining of the residents in the São Paulo Metropolitan Area (SPMA). It is an exploratory analysis that aims to answer some research questions. In this work, two techniques of MA had been used: (1) Cluster Analysis, to group and to characterize the TZ of the SPMA and, (2) Decision Tree, to find relationships between socioeconomic characteristics, land use attributes and destination choices. The results indicate that the degree of activity of origin zone influences the trip-chaining patterns.

1. INTRODUÇÃO

Um dos tópicos mais importantes na análise de demanda por transportes é a possível relação entre a estrutura espacial das cidades, as necessidades individuais de realização de atividades espacialmente dispersas, a escolha dos locais de residência e diferenças no comportamento relacionado a viagens (Srinivasan, 2000).

As cidades, ao longo de toda a sua história, têm sido locais de concentração de inúmeras atividades: comerciais, industriais, serviços, etc. Com o objetivo de realizar as diferentes atividades distribuídas no meio urbano, os indivíduos programam o seu itinerário de viagens, ou seja, a seqüência de viagens a serem realizadas durante o dia, levando em conta a localização relativa da residência e um conjunto de oportunidades.

A escolha da seqüência dos diferentes destinos, provavelmente sofre influência de vários fatores: necessidade de realizar atividades obrigatórias (como trabalho, por exemplo, que geralmente possui localização e horários fixos e envolve decisões tomadas a longo prazo), realização de atividades flexíveis e adequação de horários e localizações, distribuição de diversas atividades em determinadas zonas de tráfego, a inter-dependência entre todas as viagens realizadas em determinado período (considerando as distintas origens e destinos da cadeia de viagens), a localização relativa das zonas , o modo de transporte mais conveniente, as distâncias a serem percorridas e a atratividade ou acessibilidade de determinados locais.

As variáveis de uso do solo fazem parte do conjunto de fatores que afetam decisões individuais de realização de viagens. A complexidade da inter-relação entre atributos de uso do solo e a programação de viagens vem sendo estudada ao longo dos anos, havendo divergências

quanto às respostas encontradas (Wee, 2002; Kitamura *et al*., 1997). A incorporação de tais variáveis ao estudo de viagens encadeadas envolve uma série de dificuldades como, por exemplo, representar ou mensurar tais atributos e, de que forma, tais variáveis podem fazer parte dos modelos obtidos, já que, intuitivamente, espera-se que a estrutura urbana influencie a seqüência de viagens realizadas pelos indivíduos durante o dia.

O trabalho em questão apresenta uma análise exploratória a respeito da influência do uso do solo no encadeamento de viagens. O objetivo deste tipo de estudo é buscar uma forma de incluir variáveis que caracterizem o uso do solo em modelos que analisam o tipo da cadeia de viagens realizadas pelos indivíduos. Este trabalho faz uma investigação, com auxílio de técnicas de Análise Multivariadas de dados (AM), dos fatores que influenciam a escolha dos destinos (zonas de tráfego) dos indivíduos residentes da Região Metropolitana de São Paulo.

Através da aplicação conjunta das técnicas de AM, conhecidas como Análise de *Cluster* (AC) e Árvore de Decisão (AD), tenta-se encontrar relações entre escolhas de destinos (zonas de tráfego caracterizadas pelo uso do solo) durante a cadeia de viagens realizadas pelos indivíduos e as características socioeconômicas individuais e domiciliares e atributos das diferentes zonas residenciais.

2. JUSTIFICATIVA

Um dos principais desafios, objeto principal deste trabalho, seria compreender a escolha dos destinos de diferentes grupos de pessoas, com determinadas características individuais e domiciliares e atribuir tais escolhas à atratividade das zonas residenciais, por exemplo. Este tipo de análise visa responder algumas indagações, como:

- Pessoas que residem em zonas de alta atividade comercial, industrial, etc. trocam de zona para realizar diversas atividades? Qual o destino escolhido e as características deste? E o motivo das viagens?
- Estudantes geralmente trocam de zona para realizar atividades de estudo? Ou geralmente freqüentam escolas próximas às suas residência e, portanto, na mesma zona de tráfego? E no caso do ensino superior, onde geralmente não é possível escolher faculdades próximas ao domicílio?
- Indivíduos que residem em zonas de tráfego de alta densidade populacional e baixa atividade (zonas tipicamente de características residenciais), geralmente vão trabalhar em quais regiões? Em zonas próximas ou distantes? Eles tendem ir a zonas mais distantes com maiores ofertas de empregos ou permanecem nas regiões próximas às suas residências?

3. ETAPAS PARA REALIZAÇÃO DO TRABALHO

Para atingir a proposta desse trabalho foram realizadas as etapas representadas na Figura 1, sumariadas nas seções subseqüentes.

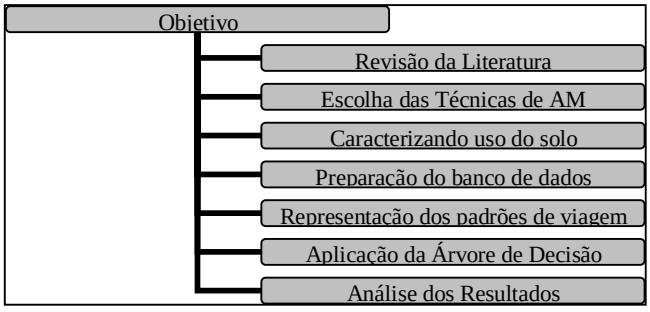

Figura 1: Síntese das etapas seguidas no trabalho

4. ESTRUTURA URBANA, USO DO SOLO E PADRÕES DE VIAGENS

A história do desenvolvimento dos núcleos urbanos está diretamente relacionada ao sistema de transportes e à evolução da sua infra-estrutura. Assim, é notória a relação intrínseca entre uso do solo e demanda de viagens, assim como a necessidade de coordenação entre políticas de uso do solo e planejamento de transportes.

A idéia de que políticas de uso do solo podem influenciar comportamento relacionado a viagens é proveniente da observação da correlação entre tais variáveis. Durante décadas, os modelos tradicionais de previsão de demanda por transportes previam o número de viagens originadas em uma determinada área geográfica (zona de tráfego) ou o número de viagens atraídas à outra área geográfica, considerando variáveis que incluíam população e empregos por zonas de tráfego. No entanto, as decisões individuais acerca das viagens a serem realizadas fazem parte de uma análise mais complexa.

O comportamento referente a viagens é função da necessidade individual de realização de atividades dispersas nas cidades, determinando desta forma o padrão de viagem a ser escolhido pelo indivíduo, dentro de um conjunto de opções e restrições. A disposição dos locais das atividades (comércio, serviço, educação, etc) nos centros urbanos determina a maior ou menor facilidade com que o indivíduo pode realizá-las diariamente (Arruda, 2005).

Reconhecendo a importância da estrutura urbana na formação de padrões de viagens, nos últimos anos, surgiram novos trabalhos que consideram dimensões de uso do solo em busca da representação mais realística do comportamento referente ao encadeamento de viagens. Embora haja divergências entre resultados obtidos na literatura, a tentativa de incorporação de tais variáveis e a investigação da sua possível influência nos padrões de viagem auxiliam a construção de estruturas mais adequadas para previsão de demanda de transportes.

Kitamura (1985) examinou relações entre a tendência de os indivíduos encadearem as viagens e as características de uma cidade hipotética linearmente disposta. A análise mostrou que a tendência de encadear viagens é função da utilidade de um conjunto de oportunidades e do tipo de atividade a ser realizada. Arruda (2005) aplicou um modelo baseado em atividades na cidade de São Carlos (SP) a fim de estudar relações entre uso do solo e comportamento de viagem/agenda de atividades. Entretanto, os resultados gerados pelo modelo não foram suficientes para fornecer evidências de que as características de uso do solo atuam de forma significativa no processo de tomada de decisão individual em relação a viagens.

Srinivan (2000) investigou como as características da vizinhança (uso do solo, rede de transporte e medidas de acessibilidade, quantificadas com auxílio de um SIG) afetam o comportamento de viagem em relação à escolha modal e cadeia de viagens. Dentre os resultados obtidos, observa-se, por exemplo, que indivíduos residentes em zonas de tráfego mistas (com altas densidades comerciais e residenciais), além de realizarem cadeias de viagens com atividades diferentes de trabalho, fazem este tipo de viagem a pé.

5. ANÁLISE MULTIVARIADAS DE DADOS

O progresso tecnológico e o avanço computacional, observados nas últimas décadas, ocasionaram uma crescente habilidade de pesquisadores, profissionais e empresas em manipular e armazenar uma quantidade cada vez maior de dados. Desta forma, viabilizou-se uma vasta compreensão e aplicação de um conjunto de técnicas estatísticas conhecidas como Análises Multivariadas (AM).

De uma maneira genérica, as técnicas de AM podem ser definidas como ferramentas analíticas que auxiliam na investigação acerca de fenômenos complexos envolvendo múltiplas dimensões. Identificam padrões que emergem de uma profusão de variáveis em interação. Hair *et al.* (1998) conceituam AM como um conjunto de técnicas estatísticas utilizadas com o objetivo de explicar e prever o grau de relações entre diversas variáveis independentes (inclusive entre si) e a variável dependente.

A crescente complexidade do estudo de demanda de viagens baseadas nas atividades exige o uso de técnicas estatísticas cada vez mais poderosas. Desta forma, é cada vez mais comum o uso de AM em diferentes trabalhos que investigam relações complexas, envolvendo múltiplas dimensões na área de Transportes (Cobbs *et al.*, 2002).

Existem diferentes técnicas multivariadas que podem ser utilizadas para diversas finalidades específicas, sendo comum a todas elas um elevado grau de complexidade que requer uma matemática relativamente sofisticada. Algumas técnicas podem ser consideradas extensões de técnicas estatísticas tradicionais, conhecidas como univariadas ou bivariadas (regressão linear simples – regressão linear múltipla; ANOVA – MANOVA). A rigor, esse conjunto inclui muitas outras técnicas, tais como Análise Fatorial, Análise de Aglomerados (*Cluster Analysis*) e Escalonamento Multidimensional. Neste trabalho, foram utilizadas duas técnicas em conjunto: (1) Análise de *Cluster*, objetivando encontrar inter-relações entre as zonas de tráfego, agrupando-as e caracterizando-as em relação ao uso do solo, realizando assim um papel complementar para aplicação de demais técnicas; (2) Árvore de Decisão, a fim de encontrar relações entre características socioeconômicas, atributos de uso do solo e escolhas de destinos na cadeia de viagens.

5.1 Análise de *Clusters*

A Análise de *Cluster* (AC), também conhecida como análise de conglomerados, é um conjunto de técnicas estatísticas cujo objetivo é agrupar objetos segundo suas características, formando grupos ou conglomerados homogêneos. Os objetos em cada grupo tendem a ser semelhantes entre si, porém diferentes dos demais objetos dos outros conglomerados. Os conglomerados obtidos devem apresentar tanto uma homogeneidade interna (dentro de cada conglomerado), como uma grande heterogeneidade externa (entre conglomerados). Portanto, se a aglomeração for bem sucedida, quando representados em um gráfico, os objetos dentro dos conglomerados estarão muito próximos, e os conglomerados distintos estarão afastados (Hair *et al.*, 1998). AC é uma técnica do tipo de interdependência, pois não é possível determinar antecipadamente as variáveis dependentes e independentes. Ao contrário, examina relações de interdependência entre todo o conjunto de variáveis.

Para aplicação da AC é necessário: (1) Definir o problema de aglomeração e as variáveis a serem tratadas estatisticamente; (2) Escolher uma medida de distância dos conglomerados (distância euclidiana, *Log-likelihood*,etc.); (3) Definir o processo de aglomeração que dependerá das variáveis em estudo e do problema em foco (hierárquico e não-hierárquico); (4) Definir ou não previamente o número de conglomerados; (5) Interpretação dos conglomerados resultantes em termos das variáveis usadas para constituí-los e de outras variáveis adicionais importantes; (6) Avaliar a validade do processo de aglomeração.

Com a finalidade de identificar características similares de uso do solo entre as zonas de tráfego da RMSP, aplica-se neste trabalho a AC a fim de agrupar zonas de tráfego segundo características como taxa de industrialização, densidade populacional, etc.

5.2 Árvore de Decisão

A técnica utilizada neste trabalho, em conjunto com a AC, é Árvore de Decisão, considerada uma forma simples de representação de relações existentes em um conjunto de dados. Os dados são divididos em subgrupos, com base nos valores das variáveis. O resultado é uma hierarquia de declarações do tipo "Se ... então ..." que são utilizadas, principalmente, para classificar dados.

Uma árvore de decisão pode ser definida como um gráfico acíclico e direto que satisfaz as seguintes propriedades: (1) A hierarquia é denominada árvore e cada segmento é denominado nó; (2) Há um nó, chamado raiz, que contém todo o banco de dados; (3) Este nó contém dados que podem ser subdivididos dentro de outros sub-nós, chamados de nós filhos; (4) Existe um único caminho entre o nó raiz e cada nó; (5) Quando os dados do nó não podem ser mais subdivididos dentro de um outro subconjunto ele é considerado um nó terminal ou folha.

Para processamento da árvore foi utilizado o Software S-PLUS 6.1. A árvore contida no S-PLUS é uma variante do algoritmo do CART, que estabelece uma relação entre variáveis independentes e variável dependente. O algoritmo é ajustado mediante sucessivas divisões binárias no conjunto de dados, de modo a tornar os subconjuntos resultantes cada vez mais homogêneos, em relação à variável dependente. Essas divisões são representadas por estrutura de árvore binária, sendo que cada nó corresponde a uma divisão (Breiman *et al*., 1984). O *software* trata a árvore como modelo de probabilidade, empregando o desvio como critério de divisão. As principais razões para sua escolha foram a sua disponibilidade e verificação da sua aplicabilidade em trabalhos anteriores (Ichikawa et al., 2002; Pitombo *et al.*, 2004).

6. CARACTERIZANDO USO DO SOLO: APLICAÇÃO DA AC

As características espaciais de uma região podem ser mensuradas através de diversas variáveis. No trabalho em questão, as variáveis de uso do solo utilizadas consideraram a distribuição de atividades (empregos, matrículas, indústria, serviços, comércio, etc.) por zonas de tráfego. Foram utilizadas algumas variáveis agregadas (características/zona de tráfego) disponibilizadas pelo METRÔ-SP. Outras foram derivadas das variáveis existentes (taxas, densidades populacionais, razão emprego e população, etc.).

Visando realizar uma caracterização das zonas de tráfego que compõem a RMSP segundo o uso do solo, foi realizada a aplicação da técnica de AC. Com o propósito de encontrar grupos homogêneos (zonas de trafego) com características de uso do solo semelhantes entre si, aplicou-se a AC (*software* SPSS 13.0), considerando as variáveis contínuas descritas na Tabela 1, o método de aglomeração *TwoStep Cluster* , a medida de distância *Log-likelihood*, e número fixo de 7 *clusters*.

Tabela 1: Variáveis contínuas usadas para aplicação da AC (Metrô-SP, dados 1997)

Empregos			**Residência**	**Educação**	
Tx de Indústria	Tx de Serviços	Tx de Comércio	Dens Populacional	Tx de Matríc abaixo do 2º gr	Tx de Matrículas acima do 2º gr
Trabalhadores na indústria por população resid	Trabalhadores no setor de serviços por população resid	Trabalhadores no setor de comércio por população resid	População residente por área	Matrículas em creche/pré-escola e 1º grau por pop. residente	Matrículas em 2º grau e ensino superior por pop. residente

Foram obtidos 7 grupos de zonas de tráfego, homogêneos segundo as variáveis indicadas na Tabela 1. As médias das variáveis em cada um dos grupos, uma nomenclatura adotada pelos autores para caracterizar cada um dos grupos (tipo de zonas) e o número de zonas que compõem cada um dos aglomerados encontram-se representados na Tabela 2. A Figura 2 ilustra a distribuição dos 7 *clusters* na RMSP.

Tabela 2: Características gerais dos 7 aglomerados obtidos

Cluster	Nº de zonas	Tx Serviços	Tx Indústria	Tx Comércio	Dens Pop	Tx_ abaixo2ºgr	Tx acima2ºgr	Nomenclatura adotada
1	4	8,30	1,08	2,32	30,51	0,53	2,88	Altísssima atividade e baixa dens. populacional
2	19	1,87	0,38	0,71	74,67	0,65	0,36	Alta atividade e moderada dens. populacional
3	58	0,39	0,06	0,12	182,75	0,19	0,06	Moderada atividade e altíssima dens. populacional
4	126	0,20	0,06	0,08	115,28	0,26	0,03	Baixa atividade e alta dens. populacional
5	51	0,50	0,21	0,17	64,04	0,38	0,08	Moderada atividade e moderada dens. populacional
6	85	0,14	0,08	0,06	17,41	0,20	0,01	Baixa atividade e baixa dens. populacional
7	18	1,00	0,78	0,42	34,31	0,29	0,08	Alta atividade e baixa dens. populacional

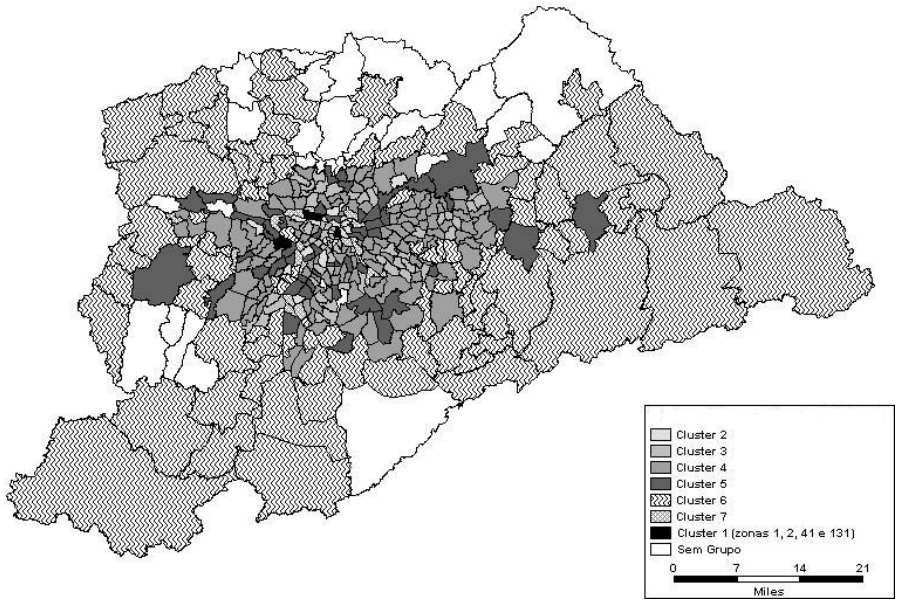

Figura 2: Distribuição dos aglomerados na RMSP.

7. DADOS

O estudo foi baseado nos dados da Pesquisa Origem-Destino (O-D) da Região Metropolitana de São Paulo (RMSP), realizada em 1997, por meio de entrevista domiciliar, pela Companhia do Metropolitano de São Paulo (METRÔ-SP). Na época, a RMSP contava com uma população de aproximadamente 17 milhões de habitantes, distribuídos em 39 municípios. O banco de dados da RMSP é composto originalmente de 98.780 indivíduos e 26.278 domicílios entrevistados. A região foi subdividida em 389 zonas de tráfego e foram coletados dados referentes aos deslocamentos das pessoas entrevistadas e suas características socioeconômicas.

Na etapa de tratamento dos dados procurou-se separar a amostra, eliminando dados incompletos ou aqueles que não faziam parte dos objetivos da análise. Separou-se a amostra final, seguindo as etapas descritas: (1) eliminação de dados incompletos, (2) eliminação de indivíduos que tenham realizado uma ou mais de quatro viagens, com intuito de limitar a complexidade da análise, (3) eliminação dos indivíduos que não tenham tido como origem e destino final a residência, (4) eliminação daqueles indivíduos que não viajaram, já que se investigou a escolha dos destinos, (5) eliminação das zonas de tráfego que não foram agrupadas pela AC, (6) eliminação dos padrões de viagem menos freqüentes devido à limitação de 128 categorias de variáveis dependentes do *software* utilizado.

Nesta etapa foram incorporados também o número e nome das zonas de tráfego residenciais de cada indivíduo, bem como características relacionadas a estas. Assim, a amostra final constituía características individuais e domiciliares e características das zonas de tráfego da resi-

dência dos indivíduos que realizaram duas, três ou quatro viagens. Finalmente, para a aplicação da Árvore de Decisão, foi analisada uma amostra final composta de 42.766 indivíduos.

8. REPRESENTAÇÃO DOS PADRÕES DE VIAGEM

Os padrões de viagens, termo que se refere à cadeia de viagens associada às origens e destinos dos diferentes deslocamentos realizados pelos indivíduos, são as categorias da variável dependente/resposta criada pelos autores. Estes representavam, através de uma codificação numérica, características dos destinos escolhidos e das zonas da residência de cada um dos indivíduos analisados (7 tipos de zonas agrupadas pela AC). A Tabela 3 apresenta a caracterização das zonas de tráfego proposta com auxílio da técnica de AC (7 *clusters*) e as zonas que compõem cada um dos grupos obtidos.

Tabela 3: Caracterização dos padrões de viagem

Grupo	Tipo ou classificação proposta da zona de tráfego	Zonas de tráfego que constituem cada grupo
1	Altíssima atividade e baixa densidade populacional	1,2,41,131
2	Alta atividade e moderada densidade populacional	5,7,8,16,20,21,22,39,40,43,44,66,70,77,128,158,191,204
3	Moderada atividade e altíssima densidade populacional	3,4,6,11,12,14,15,17,18,31,32,33,34,35,38,65,71,74,76,81,86,87,89,97,99,104 107,115,116,119,142,144,145,146,147,148,167,171,173,174,175,176,182,183 184,224,225,233,252,253,254,280,282,283,286,289,306
4	Baixa atividade e alta densidade populacional	10,26,29,30,37,45,47,49,53,54,55,56,58,60,61,63,67,73,75,78,82,83,84,85,92,93,94 95,96,98,100,103,105,106,109,111,113,114,118,121,123,137,139,140,141,143,152 153,154,155,156,157,160,161,162,163,164,165,166,168,169,172,177,178,181,185 186,188,197,199,200,201,202,203,206,210,217,239,241,242,245,247,255,256,257,258 259,261,262,263,266,267,279,281,284,285,287,288,290,291,293,294,297,299,302,303 307,308,309,315,316,331,336
5	Moderada atividade e moderada densidade populacional	19,24,25,36,42,48,50,51,52,59,62,64,68,69,90,91,101,110,117,120,124,127,129,132,134 135,136,149,140,151,179,180,190,193,198,205,222,234,248,260,264,269,278,310 319,334,338,355,372
6	Baixa atividade e baixa densidade populacional	125,130,138,159,170,187,194,195,196,207,208,209,215,216,230,231,238,244,249 251,270,271,272,276,292,295,296,298,300,301,304,305,311,312,313,314,317,321,322 324,325,326,328,329,332,333,335,337,339,340,341,342,343,344,345,346,347,348,349 350,351,356,357,358,361,368,369,370,371,373,374,375,376,377,378,379,380,382,383 384,385,387,388,389
7	Alta atividade e baixa densidade populacional	9,23,27,28,46,57,72,79,80,102,122,126,133,189,192,250,265,318
s/grupo	108,112,211,212,213,214,268,273,274,275,277,320,323,327,330,352,353,354,359,360,362,363,364,365,366,367,381,386	

Assim, o primeiro e último algarismos representam sempre a primeira origem e destino final, que devem ser a residência. Os padrões representavam o tipo de zona de origem (residência) e dos diversos destinos. Por exemplo, o padrão “121” indica a realização de duas viagens: (1) A primeira com origem (residência) em uma zona tipo 1 (altíssima atividade e baixa densidade populacional) e destino em uma zona tipo 2 (alta atividade e moderada densidade populacional); (2) A segunda viagem tem como origem, o destino anterior (zona tipo 2) e destino final em uma zona tipo 1 (residência). O padrão “333”, por exemplo, indica que não houve troca de zona, ou então ocorreu troca de zonas do mesmo tipo (moderada atividade e altíssima densidade populacional) durante as duas viagens realizadas.

Para o caso de três e quatro viagens, há uma maior diversidade de padrões de viagem. O padrão “61316” representa indivíduos que realizaram quatro viagens: (1) A primeira viagem com origem (residência) em uma zona de tráfego tipo 6 (baixa atividade e baixa densidade populacional) e destino em uma zona tipo 1 (altíssima atividade e baixa densidade populacional); (2) A segunda viagem com origem em uma zona tipo 1 e destino em uma zona tipo 3 (moderada atividade e altíssima densidade populacional); (3) A terceira com origem em uma zona tipo 3 e destino em uma zona tipo 1; (4) Finalmente, a última viagem com origem em uma zona tipo 1 e retorno final ao domicílio (zona tipo 6). A Figura 3, em seguida, ilustra

padrões de viagens para indivíduos que realizaram duas (Padrão “333”), três (Padrão “6166”) e quatro viagens (Padrão “72177”), respectivamente, considerando zoneamentos fictícios.

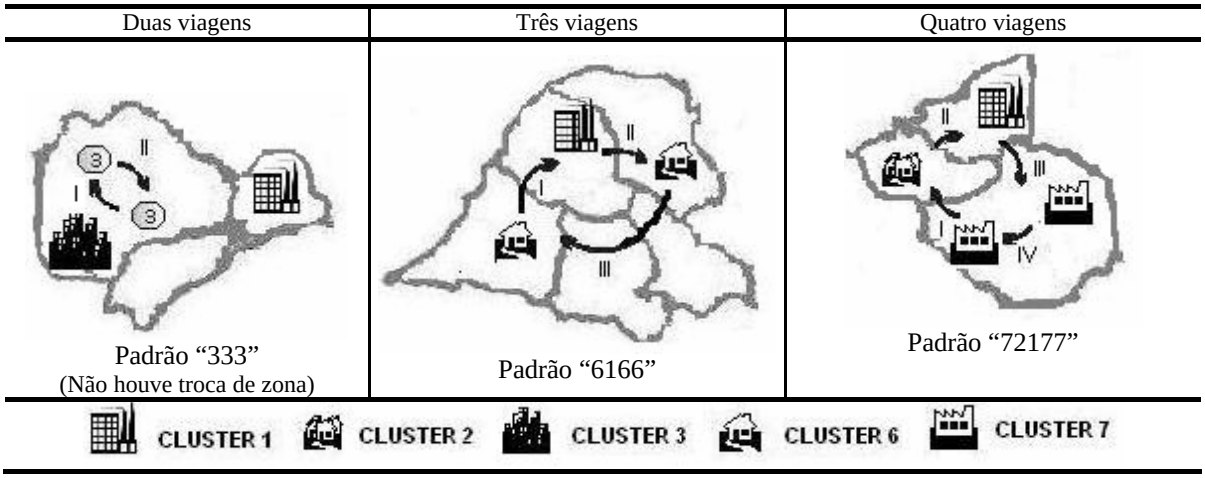

Figura 3 : Exemplo de Padrões de Viagem (2, 3 e 4 viagens)

A Tabela 4 representa as freqüências dos padrões de viagem mais observados para os casos de duas, três e quatro viagens respectivamente, além da sua distribuição por motivo de viagem em cada uma das viagens realizadas (E – Escola, T – Trabalho, A – Outras atividades, R – Residência). O motivo da última viagem (em todos os casos: 2, 3 e 4 viagens) foi desconsiderado já que sempre será “Residência”.

Tabela 4 : Padrões de viagem mais freqüentes e sua distribuição por motivo de viagem

	2 vgs (%)		Motiv(1ª vg)	3 vgs (%)		Motiv(1ª vg)	Motiv(2ª vg)	4 vgs (%)		Motiv(1ª vg)	Motiv(2ª vg)	Motiv(3ª vg)
Padrões de viagens	**444**	22,40	E (55%) T (30,9%) A (14,1%)	**4444**	5,68	E (19,44%) T (69,44%) A (11,1%)	E (33,33%) T (19,44%) A (47,22%)	**44444**	26,36	E (36,95%) T (63,05%) A (0,00%)	R (98,47%) A (1,52%)	E (50,48%) T (47,81%) A (1,71%)
	666	12,62	E (58,1%) T (28%) A (13,9%)	**4554**	3,31	E (28,57%) T (71,43%) A (0,00%)	E (42,85%) T (28,57%) A (28,57%)	**66666**	14,16	E (30,14%) T (69,86%) A (0,00%)	R (99,65%) A (0,35%)	E (44,68%) T (53,19%) A (2,13%)
	333	7,80	E (50,2%) T (30,2%) A(15,7%)	**4454**	3,00	E (26,32%) T (52,63%) A (21,05%)	E (47,36%) T (26,31%) A (26,31%)	**55555**	9,79	E (29,23%) T (70,77%) A (0,00%)	R (100,00%) A (0,00%)	E (42,05%) T (56,41%) A (1,54%)
	454	6,53	E (24,1%) T (55,7%) A (20,2%)	**4544**	3,00	E (21,05%) T (73,68%) A (5,26%)	E (42,11%) T (21,05%) A (36,84%)	**33333**	9,19	E (41,53%) T (58,47%) A (0,00%)	R (98,91%) A (1,09%)	E (54,09%) T (45,91%) A (0,00%)

9. APLICAÇÃO DA ÁRVORE DE DECISÃO E RESULTADOS

O S-Plus 6.1 gera resultados na forma gráfica e tabular. A forma gráfica representa a árvore gerada e apresenta em cada um dos nós terminais (folhas) o padrão que ocorre com maior freqüência. Cada nó terminal representa um “grupo homogêneo” de indivíduos segundo dois critérios fixados exogenamente: número mínimo de observações em nós terminais e desvio mínimo. A forma tabular consiste num relatório que apresenta detalhadamente os resultados da árvore: número do nó, total de observações no nó, desvio, padrão predominante e as probabilidades de ocorrência de todos os padrões em análise. A árvore foi gerada a partir da amostra final com 42.766 indivíduos, variável dependente (padrões de viagem segundo caracterização proposta de uso do solo), adotando-se o mínimo de 50 observações por nó terminal e desvio global de 0,1. As variáveis independentes (variáveis socioeconômicas, participação em atividades, e características agregadas por zona de tráfego da residência) estão representadas na Tabela 5.

Tabela 5: Variáveis independentes inseridas na análise

Variáveis independentes						
Uso do solo (variáveis/zonas de tráfego)	Número absoluto/zona de tráfego	Domicílios Famílias População		**Uso do solo**	Taxas	Emprego/Pop: Indústria/Pop, Serviços/Pop, Comércio/Pop
		Matrículas	Creche/ Pré Escola 1º grau 2º grau Superior Outros Cursos			Matrículas/Pop: Abaixo do 2o grau, Acima do 2o grau
		Empregos	Indústria Comércio Serviços Outros		Cluster da Origem (Residência: 1,2,3,4,5,6 e 7	
		Automóveis		**Atividades**	**TRABALHA** (1: Assalariado; 2 : Autônomo; 3 Não) **ESTUDA** (1: Não; 2: Creche/Pré-escola; 3 : 1º,2º,3º Graus; 4: Outros)	
	Taxas	Densidade Populacional		**Socioeconômicas**	**Posição do indivíduo na família** (1: Chefe; 2: Cônjuge; 3 : Filho; 4: Parente/Agregado; 5 : Empregado residente; 6 : Visitante) **Grau de instrução** (1: Não-alfabetizado; 2: Pré-escola; 3 : 1º grau incompleto; 4: 1º grau completo; 5 : 2º grau incompleto; 6 : 2º grau completo,7: Superior incompleto, 8: Superior completo) **Renda Individual (R$), Renda Familiar (R$)** **Sexo** (1:Homens, 2: Mulheres), **Idade**, **Nº de autos no domicílios** **Total de Pessoas na família, Nº de crianças no domicílio** **Nº de carteiras de habilitação no domicílio (provável)** **Usa Vale Transporte** (1-Sim, 2 - Não) **Setor de Atividade** (1- Agrícola, 2 - Construção Civil 3- Indústria, 4 Comércio, 5 - Serviços, 6 - Outros, 7 - N se aplica)	
		Densidade de empregos	Densidade indústria Densidade Comércio Densidade Serviços			
		Densidade de matrículas	Densidade creche Densidade 1º grau Densidade 2º grau Densidade superior			
		Taxa de motorização*				

*Número de automóveis particulares por 1.000 habitantes

A variável de maior importância é *Cluster da origem*. A partir da raiz, a árvore se ramifica em dois grupos principais: (1) *Clusters* (3,5,6) e (2) *Clusters* (1,2,4,7). Posteriormente ocorrem novas segmentações do conjunto de dados considerando novamente a variável *Cluster da origem, Idade, Estuda, Usa Vale Transporte* e *Grau de Instrução*. Ao final da segregação dos dados foi encontrado um total de 10 folhas. A Figura 4 representa a árvore gerada pelo S-PLUS 6.1 com desvio mínimo de 0,1. Nas folhas encontram-se ilustrados os padrões de viagem mais freqüentes em cada um dos casos. A Tabela 6, em seguida, apresenta uma síntese rearranjada dos resultados emitidos pelo relatório do *software*, onde a primeira coluna indica o número do nó e o percentual de cada nó em relação à amostra total, enquanto que a segunda coluna mostra as variáveis consideradas pelo modelo CART em cada uma das folhas. As colunas seguintes indicam, em ordem decrescente, os padrões mais freqüentes na população da RMSP caracterizada pelas condições indicadas na segunda coluna, bem como a freqüência da seqüência dos motivos de viagem para cada um dos padrões mais observados.

10. ANÁLISE DE RESULTADOS

Foram identificados 10 grupos de indivíduos caracterizados como "homogêneos" em relação às variáveis socioeconômicas (*Idade, Usa Vale Transporte, Grau de Instrução*), atividades (*Estuda*) e de uso do solo (*Cluster da origem*). Observa-se que a variável *Cluster da origem*, além de ser a mais importante para segmentação dos dados, englobou todas as outras variáveis de uso do solo (que não apareceram no modelo de CART). Através da divisão dos dados, foram separados grupos que se diferenciavam pelo *cluster de origem*: (nó 8) *Cluster* 3; (nó 9); *Cluster* 5; (nó 10 e 11) *Cluster* 6; (nó 12, 52, 106, 107 e 27) *Cluster* 4 e (nó 7) *Clusters* 1, 2 e 7. Verificou-se que a Árvore de Decisão separou os *clusters* de atividade moderada (3,5) do *cluster* 6 (de baixa atividade e baixa densidade demográfica), já o nó 7 foi composto pelos *clusters* de alta ou altíssima atividade (*clusters* 1, 2 e 7).

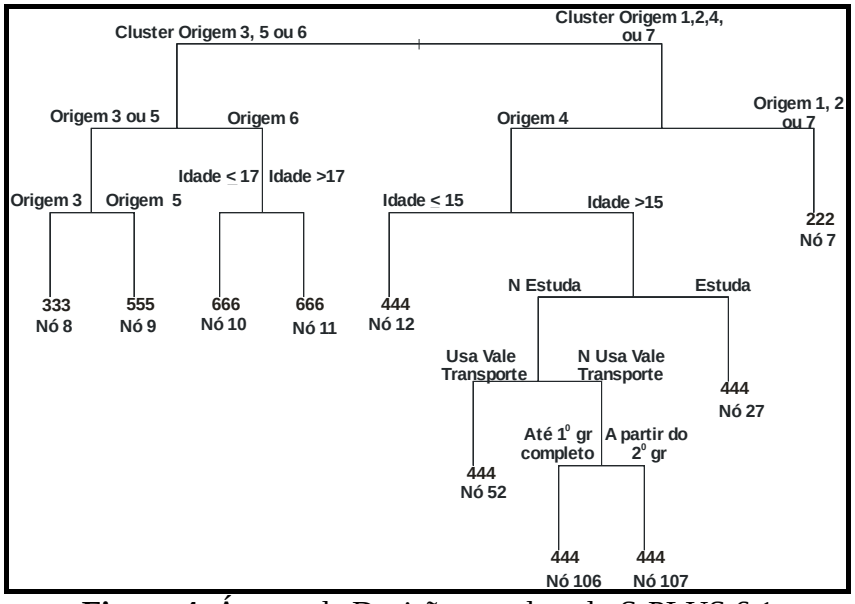

Figura 4: Árvore de Decisão gerada pelo S-PLUS 6.1

Tabela 6: Síntese do relatório (R-Residência, A-Outras Atividade, E-Escola, T-Trabalho)

		Padrões mais observados							
Nó	Características	Padrão1	Seq. de motivos	Padrão2	Seq. de motivos	Padrão3	Seq. de motivos	Padrão4	Seq. de motivos
8	Cluster	333	R-A-R (15,7%)	343	R-A-R (16,95%)	353	R-A-R (17,42%)	323	R-A-R (22,15%)
			R-E-R (50,2%)		R-E-R (34,60%)		R-E-R (19,95%)		R-E-R (19,23%)
16,64%	de origem = 3	44,63%	R-T-R (34,2%)	19,92%	R-T-R (48,45%)	10,79%	R-T-R (62,58%)	10,61%	R-T-R (58,62%)
9	Cluster	555	R-A-R (16,03%)	545	R-A-R (22,53%)	535	R-A-R (20,00%)	525	R-A-R (25,88%)
			R-E-R (48,72%)		R-E-R (30,12%)		R-E-R (24,00%)		R-E-R (22,59%)
12,19%	de origem = 5	50,92%	R-T-R (35,26%)	15,94%	R-T-R (47,11%)	10,08%	R-T-R (56,00%)	8,76%	R-T-R (51,54%)
10	Cluster	666	R-A-R (4,68%)	646	R-A-R (7,57%)	656	R-A-R (7,81%)	626	R-A-R (26,09%)
	de origem = 6		R-E-R (93,32%)		R-E-R (88,55%)		R-E-R (85,31%)		R-E-R (63,04%)
9,40%	Idade ≤ 17	73,42%	R-T-R (2,00%)	12,28%	R-T-R (3,89%)	7,97%	R-T-R (6,88%)	2,29%	R-T-R (10,87%)
11	Cluster	666	R-A-R (26,28%)	656	R-A-R (23,77%)	646	R-A-R (23,85%)	626	R-A-R (30,52%)
	de origem = 6		R-E-R (10,62%)		R-E-R (6,10%)		R-E-R (6,16%)		R-E-R (4,11%)
13,94%	Idade > 17	36,66%	R-T-R (63,09%)	18,43%	R-T-R (70,13%)	17,17%	R-T-R (69,99%)	7,76%	R-T-R (65,37%)
12	Cluster	444	R-A-R (4,67%)	454	R-A-R (10,11%)	434	R-A-R (16,55%)	424	R-A-R (26,67%)
	de origem = 4		R-E-R (94,43%)		R-E-R (86,88%)		R-E-R (80,94%)		R-E-R (69,17%)
12,63%	Idade ≤ 15	80,14%	R-T-R (0,90%)	8,62%	R-T-R (3,01%)	5,15%	R-T-R (2,52%)	2,22%	R-T-R (4,17%)
52	Cluster de orig=4	444	R-A-R (5,77%)	454	R-A-R (5,50%)	424	R-A-R (5,63%)	434	R-A-R (3,66%)
	Idade > 15, N Est		R-E-R (1,15%)		R-E-R (0,63%)		R-E-R (0,28%)		R-E-R (0,56%)
5,06%	Usa Vale Transp	28,06%	R-T-R (93,08%)	21,87%	R-T-R (93,87%)	16,41%	R-T-R (94,08%)	16,41%	R-T-R (95,77%)
106	Clus de orig=4, Idade >15	444	R-A-R (14,22%)	454	R-A-R (24,26%)	434	R-A-R (25,49%)	424	R-A-R (33,17%)
	N Est, N usa VTRA		R-E-R (57,55%)		R-E-R (22,52%)		R-E-R (21,05%)		R-E-R (10,63%)
11,10%	Até 1º gr completo	45,74%	R-T-R (28,23%)	17,82%	R-T-R (53,22%)	11,43%	R-T-R (53,46%)	7,51%	R-T-R (56,19%)
107	Clus de orig=4, Idade >15	444	R-A-R (19,56%)	454	R-A-R (16,63%)	434	R-A-R (16,37%)	424	R-A-R (18,45%)
	N Est, N usa VTRA		R-E-R (9,78%)		R-E-R (10,65%)		R-E-R (8,55%)		R-E-R (8,61%)
7,27%	A partir do 2º gr incomp.	34,59%	R-T-R (70,66%)	18,15%	R-T-R (72,73%)	14,70%	R-T-R (75,07%)	11,55%	R-T-R (72,93%)
27	Cluster de orig=4	444	R-A-R (5,12%)	454	R-A-R (6,11%)	434	R-A-R (6,74%)	424	R-A-R (5,48%)
	Idade > 15		R-E-R (79,83%)		R-E-R (69,77%)		R-E-R (64,04%)		R-E-R (60,96%)
4,60%	Estuda	47,00%	R-T-R (15,05%)	16,00%	R-T-R (24,12%)	9,00%	R-T-R (29,21%)	7,00%	R-T-R (33,56%)
7	Cluster	222	R-A-R (20,50%)	777	R-A-R (11,89%)	747	R-A-R (14,47%)	232	R-A-R (21,20%)
			R-E-R (40,83%)		R-E-R (48,54%)		R-E-R (43,86%)		R-E-R (35,60%)
7,17%	de origem = 1, 2 ou 7	18,00%	R-T-R (38,70%)	13,00%	R-T-R (39,56%)	9,00%	R-T-R (41,40%)	8,00%	R-T-R (43,20%)

Como se pode ver, a variável independente *Cluster da Origem* possui forte correlação com a variável dependente (padrão de viagem), pois todos os padrões de viagem têm início nas zonas com as características descritas pelo *Cluster de Origem*. Embora ela não contribua para explicar o comportamento dos viajantes acerca da formação do encadeamento de viagem, a manutenção dessa variável permite que se agrupem, em cada ramo da árvore, todas as viagens originadas nas zonas com determinadas características, facilitando consideravelmente a visualização e análise dos resultados.

Observando a Figura 4, verifica-se que os padrões predominantes são aqueles correspondentes a indivíduos que realizaram duas viagens (que constituem aproximadamente 95% da amostra total) e que não trocaram de zona, ou então trocaram de zona de mesma característica de uso do solo (por exemplo: “333”, “555”, “444”). Na maioria dos casos não há troca de zonas, com exceção do grupo correspondente ao nó 52 (havendo uso de Vale Transporte), onde permanecem na mesma zona de tráfego apenas 22,08% daqueles indivíduos que realizam o padrão “444”. Analisando a Tabela 6, observa-se que quando há troca de *cluster* ou tipo de zona durante as viagens, aumenta-se o percentual das viagens com motivo “Trabalho”. Pode-se concluir que a necessidade de realizar atividades relacionadas ao trabalho faz com que indivíduos possivelmente tenham que se deslocar maiores distâncias, principalmente quando estes residem em zonas de baixa atividade. O mesmo não ocorre quando se considera as viagens com motivo “Escola”, já que se podem escolher escolas (pelo menos de 1^o ou 2^o grau) próximas às residências, não havendo, assim, necessidade de troca de zona ou tipo de zona.

Analisando as folhas 10 e 11, pode-se concluir a respeito da influência da idade em grupos de indivíduos que residem em zonas de baixa atividade e baixa densidade demográfica. Observa-se, como esperado intuitivamente, que indivíduos de menor idade (nó 10, Idade menor ou igual a 17 anos) tendem a fazer viagens mais curtas, ou seja, possuem uma menor tendência em fazer troca de zonas, considerando ainda que, neste caso, o principal motivo de tais viagens é “Escola”, e, na maioria das vezes, estas localizam-se próximas às residências. Corrobora-se esta afirmação verificando-se também o grupo de indivíduos correspondentes ao nó 12, os quais residem em uma zona de baixa atividade e alta densidade demográfica e têm uma idade inferior ou igual a 15 anos.

Espera-se que indivíduos que residam em zonas de alta atividade, geralmente não troquem de zona para fazer qualquer atividade (seja esta estudo, trabalho ou uma outra atividade). Observando a folha 7, a qual constitui indivíduos que moram nas zonas tipo 1, 2 e 7, observa-se predominância do padrão “222”, já que a zona tipo 2 possui maior densidade demográfica que as zonas tipo 1 e 7, havendo na amostra estudada um maior número de indivíduos residentes nas zonas tipo 2. Analisando mais detalhadamente este nó, verifica-se (ainda que não representado na Tabela 6) que o padrão “111” ocorre em 1,93% do grupo. Desta forma, foi gerada uma matriz O/D (Tabela 7), considerando apenas o subgrupo das pessoas pertencentes ao nó 7 e que realizam o padrão “111”. Nota-se que a grande maioria do subgrupo permanece na mesma zona de tráfego, como esperado previamente. Foi escolhido tal subgrupo para análise já que as zonas agrupadas do *cluster* 1 possuem altíssima atividade. Verificou-se, por exemplo, que todas as viagens da zona 131(Cidade Universitária) têm motivo “Escola”.

Tabela 7: Matriz O/D dos indivíduos que pertencem ao nó 7 e realizam o padrão “111”

	1	2	41	131	**Produzidas**
1 (Sé)	**13**	0	0	0	13
2 (Parque Dom Pedro)	3	**18**	0	0	21
41(Água Branca)	1	0	**15**	0	16
131(Cidade Universitária)	0	0	0	**9**	9
Atraídas	17	18	15	9	**59**

11. CONCLUSÕES

Um dos princípios básicos da análise de viagens baseada em atividades é que características de uso do solo, assim como variáveis socioeconômicas e aspectos do sistema de transportes, influenciam as decisões individuais relacionadas a viagens. No entanto, nem sempre se encontram resultados convergentes na literatura quando se investiga a relação uso do solo e padrões de viagem. Uma das principais dificuldades seria como mensurar ou representar através de

variáveis aspectos da estrutura urbana, e, desta forma, como incorporá-los nos modelos de previsão de viagens encadeadas.

Esta pesquisa, ainda em fase incipiente, apresentou uma análise exploratória a respeito de como caracterizar uso do solo na RMSP, através de AC, e investigar, através da aplicação de Árvore de Decisão, a influência de variáveis socioeconômicas e atributos das zonas de tráfego residenciais nos padrões de viagens caracterizados pelo uso do solo (altíssima atividade, baixa densidade, etc.). Conclui-se que, considerando uma variável dependente relacionada ao uso do solo (Padrão "111", "323", "666", etc.), observa-se a influência da variável característica da zona de origem (ou *cluster da origem*) e variáveis socioeconômicas como idade, grau de instrução, usa Vale Transporte e exercício de atividade estudantil ("ESTUDA"). Verificou-se que, quando há troca de tipo de zona durante a cadeia de viagens, o motivo predominante é "TRABALHO", pois as decisões de viagens envolvendo atividades de trabalho são tomadas a longo prazo e, portanto, nem sempre é possível residir em regiões próximas ao local de trabalho, ou "escolher" em qual zona trabalhar, o que não ocorre quando o motivo de viagem é "ESCOLA". Observando ainda os resultados, conclui-se que indivíduos residentes em zonas de alta atividade, geralmente não trocam de zona para realizar qualquer atividade, já que há alta taxa de serviços, comércio, indústria e escolas próximas às suas residências. Já aqueles indivíduos que residem em zonas de baixa atividade têm maiores necessidades de trocar de zona para realizar atividades diversas, inclusive aquelas relacionadas ao trabalho.

Espera-se, no futuro, que a recém proposta caracterização do uso do solo possa ser útil principalmente na investigação de variáveis dependentes não estritamente relacionadas ao uso do solo, mas a atributos associados à cadeia de viagens (como motivos de viagem, modos de transporte, período do dia, etc.). Pretende-se, desta forma, concluir qual tipo de indivíduo (características socioeconômicas), que reside em determinado tipo de zona (característica de uso do solo), realiza determinado padrão de viagem (seqüência de atividades, modos de transporte e período do dia). Deste modo, abre-se um leque para a compreensão de como o comportamento referente a viagens encadeadas relaciona-se aos padrões de uso do solo nos locais de residência e trabalho dos indivíduos, bem como nas suas vizinhanças.

AGRADECIMENTO: A FAPESP, pelo apoio financeiro à pesquisa.

REFERÊNCIAS BIBLIOGRÁFICAS

Arruda, F. S. (2005) *Aplicação de um Modelo Baseado em Atividades para Análise da Relação Uso do Solo e Transportes no Contexto Brasileiro.* 145 p. Tese (doutorado). Universidade de São Paulo.

Breiman, L.; J.H Friedman; R.A. Olshen E C.J. Stone (1984) *Classification and Regression Trees*. Wadsworth International Group, Califórnia.

Coobs, L.; M. Cunningham,.; C. Gerde, (2002) *Multivariate Analysis of Vehicle Safety.* Capturado em 20/06/05. Disponível na *internet.* http://www.users.edu/porterbm/sumj/2002/VehicleSafety.pdf

Hair, J.F.; R.E Anderson.; R.L Tatham.; W.C. Black (1998). *Multivariate Data Analysis*. 5ª ed. Prentice-Hall. Upper Saddle River, New Jersey, 730p.

Ichikawa, S. M.; C. S. Pitombo; E. Kawamoto (2002) Aplicação de Minerador de Dados na Obtenção de Relações entre Padrões de Viagens Encadeadas e Características Socioeconômicas. *Panorama Nacional de Pesquisa em Transportes*, XVI ANPET, v.2, p.175-186.

Kitamura, R.; P.L Mokhtarian,.; L. Laidet (1997). A micro-analysis of land use and travel in five neighborhoods in the San Francisco Bay Area. *Transportation***.** n. 24, p. 125-158.

Kitamura, R. (1985) Trip Chaining in a Linear City. *Transportation Research*, v.19A, N^0 2, p. 115 – 167.

Pitombo, C. S. ; P. B. Sousa ; E. Kawamoto (2004) A influência de mudanças contextuais nos padrões de encadeamento de viagens urbanas. *Panorama Nacional da Pesquisa em Transportes XVIII Congresso de Pesquisa e Ensino em Transportes (XIII Anpet)*, Florianópolis, v. 1. p. 687-698.

Srinivasan, S. (2000) *Liking Land Use and Transportation: Measuring the Impact of Neighborhood-scale Spatial Patterns on Travel Behavior.* 248 p. Tese (doutorado). Harvard University.

Wee, B. van (2002). Land use and transport: research and policy challenges. *Journal of Transport Geography*, v. 10, n. 4, p. 259-271.

Autores:	Universidade de São Paulo
Cira Souza Pitombo: cira@sc.usp.br Eiji Kawamoto: eiji@usp.br	Escola de Engenharia de São Carlos Avenida Trabalhador SãoCarlense, 400, Centro. CEP: 13566-590