# PADRÕES DE ATIVIDADES DE RESIDENTES DE PARAISÓPOLIS: ANÁLISE DE DADOS DE MÚLTIPLOS DIAS COLETADOS COM SMARTPHONES

**Bruna Pizzol**
**Orlando Strambi**
**Mariana Abrantes Giannotti**
Universidade de São Paulo
Escola Politécnica
**Bianca Bianchi Alves**
Banco Mundial

**RESUMO**
Este trabalho propõe investigar os padrões de atividades de 105 residentes de Paraisópolis, segunda maior favela de São Paulo. Os dados foram obtidos por meio de uma entrevista domiciliar em campo e pela coleta de dados de localização de GPS a cada cinco minutos, pelo período de uma semana. A partir dos dados de localização, foram identificadas atividades e locais de interesse para cada indivíduo. As atividades foram classificadas em seis categorias, usadas para descrever os padrões semanais de atividades dos moradores. Uma segunda rodada de classificação agrupou os indivíduos que apresentavam padrões semanais de atividades semelhantes, em sete grupos distintos. Cada grupo foi então descrito em termos de suas características demográficas e socioeconômicas. Destaca-se o fato de que, entre todos os indivíduos, 56% foram classificados em grupos de comportamento variado, apontando para a necessidade de considerar outros tipos de padrões de atividades nos processos de modelagem de transportes, além dos padrões regulares tipicamente considerados, como casa-trabalho.

**ABSTRACT**
This paper investigates the activity patterns of 105 residents of Paraisópolis, the second largest slum of São Paulo. Data was collected with a home interview and tracking GPS location every 5 minutes for one week. Based on location data, individual stays and points of interest we inferred. Based on their characteristics, stays were clustered into 6 classes, later used to describe individuals' weekly activity patterns. Individuals were then clustered into 7 categories, based on similarity of their activity patterns. Each group was then described in terms of its demographic and socioeconomic characteristics. It should be highlighted that 56% of the people in the sample was classified in groups with diversified behavior, indicating the need to consider other activity patterns beyond the more usual simple commute considered in modelling efforts.

## 1. INTRODUÇÃO

Estudos que buscam compreender como os indivíduos se movem no tempo e espaço podem ser diretamente relevantes para o desenvolvimento de políticas de transporte, servindo de instrumento para os tomadores de decisão na implantação de políticas a curto e longo prazo. A análise dos padrões de atividades e viagens apresenta-se como um aspecto importante nesse sentido, já que é necessário entender o comportamento das pessoas para poder oferecer soluções de transporte que atendam adequadamente às suas demandas diárias.

A maioria dos modelos e análises de comportamento relativo a viagens objetiva explicar as variações no comportamento dos indivíduos em relação às viagens com base em suas características pessoais e do ambiente que os cerca, o que constitui a chamada *variabilidade interpessoal*. Para tanto, são tipicamente baseados nos registros de viagem dos indivíduos para um único dia.

Dados de um dia, no entanto, revelam apenas aspectos limitados sobre o comportamento de viagens (Hanson e Huff, 1981; Pas e Koppelman,1987), visto que as pessoas não repetem o mesmo padrão todos os dias. Apesar da chamada *variabilidade intrapessoal* ser reconhecida conceitualmente, avaliar sua magnitude requer a análise de dados de múltiplos dias de um

mesmo indivíduo. Contudo, até recentemente, a coleta de dados de viagens por múltiplos dias não era comum entre as pesquisas de transporte.

A posse de dados de múltiplos dias, que permitem avaliar a existência da variabilidade intrapessoal, aumenta, porém, a complexidade das análises. A fim de reduzi-la, é importante classificar diferentes grupos da população, visando identificar comportamentos similares ao longo de múltiplos dias. Essas categorias podem auxiliar a aprimorar os modelos desenvolvidos para descrever o comportamento de atividades e viagens.
Ao considerar especificamente indivíduos residentes em favelas e assentamentos informais, as condições geográficas, sociais e econômicas a quais estão submetidos podem apresentar considerável influência nas atividades e viagens que realizam. Todavia, os estudos de transporte em favelas em geral referem-se a locais na África e Ásia, havendo poucas referências da América do Sul ou, em particular, do Brasil (Koch *et al*., 2013).

Considerando a escassa literatura sobre o uso de *smartphones* para a coleta de dados de transporte no Brasil, assim como sobre padrões de viagens de população residente em favelas, este artigo investiga o comportamento de moradores de Paraisópolis, usando dados coletados por múltiplos dias a partir de um aplicativo de *smartphone* e de uma entrevista que aborda aspectos socioeconômicos e de hábitos de transporte.

Após esta introdução, o artigo está organizado em mais cinco seções. É apresentada, no item 2, uma breve revisão de literatura a respeito dos temas envolvidos. O item 3 apresenta a coleta de dados da pesquisa e, o item 4, o processamento dos dados, incluindo limpeza dos dados e seleção dos indivíduos para análise, método de identificação de atividades e *Points of Interest* (POI's), e métodos de agrupamento. No item 5 são apresentados e analisados os resultados obtidos e, por fim, no item 6, são apresentadas as conclusões e considerações finais do estudo.

## 2. REVISÃO DA LITERATURA

Coletar dados de transporte de pessoas que vivem em favelas é um desafio. Segundo Ampt e Hickman (2015), elas podem ser consideradas um *grupo de difícil acesso* por alguns motivos: (i) dificuldade de enumerar seus habitantes, já que os dados de censos demográficos podem ser incompletos ou faltantes para essa parcela da população; (ii) dificuldade de escolher uma amostra, dada a lacuna de endereços formais e números de telefone; (iii) dificuldade de atingir a população alvo da pesquisa, dados os horários de trabalho normalmente atípicos; iv) necessidade de respeitar o crime local organizado, quando existente. Pode haver também barreiras de educação ou tecnológicas e as pessoas podem ser menos motivadas a participar da pesquisa, dada a possível falta de confiança de que a mesma produzirá benefícios locais. Devido a isso, certos incentivos para redução de taxas de não resposta podem ser necessários (Behrens *et al.*, 2009).

Novos métodos de coleta de dados podem contribuir nesse sentido, já que permitem a coleta de um grande volume de dados a baixo custo, de forma mais fácil que pesquisas tradicionais. Por exemplo, o uso de *smartphones* com receptor de GPS (*Global Positioning System*) embutido permitem registrar as trajetórias espaço-temporais diárias dos indivíduos rastreados (Chen *et al.*, 2016). Esse tipo de coleta de dados é associado a níveis maiores de precisão da localização e redução da subestimativa de viagens, aspectos que são particularmente úteis em comunidades densas como favelas, nas quais pode-se esperar altos níveis de atividades locais e deslocamentos curtos, normalmente sub representados em métodos tradicionais de coleta de

dados de atividades e viagens (Koch *et al.*, 2013; Maia *et al.*, 2016).

O principal problema do uso de receptores de GPS para a coleta de dados está relacionado a possíveis perdas do sinal de satélite. Essa situação pode ser bastante comum em edifícios, veículos, passagens subterrâneas ou até mesmo em áreas densamente construídas, causando o chamado efeito urbano de *canyon* (Stopher *et al.*, 2008). Além disso, Anda *et al.* (2017) apontam uma desvantagem a ser considerada ao utilizar esses tipos de dados: a sua natureza bruta, sendo necessário um esforço analítico adicional para identificação de atividades, viagens e suas características. Hoje o principal desafio imposto nesse cenário é o desenvolvimento de algoritmos robustos que possam extrair agendas individuais diárias a partir de dados de mobilidade, por vezes bastante esparsos.

Ainda assim, esse método facilita a coleta de dados de múltiplos dias, permitindo o conhecimento da localização contínua das pessoas. Além disso, destaca-se atualmente a crescente posse de *smartphones* entre os mais diversos extratos da população. Até mesmo em uma comunidade de baixa renda como Paraisópolis, *smartphones* são comuns e a internet 3G é amplamente disponível.

## 3. COLETA DE DADOS

A pesquisa de campo foi realizada como parte de um projeto desenvolvido pelo Banco Mundial, em parceria com o Laboratório de Geoprocessamento (LabGEO) da Escola Politécnica da Universidade de São Paulo. A pesquisa foi aprovada pelo sistema CEP/CONEP em um processo submetido pela Plataforma Brasil, que regula a questão ética de realização de pesquisas envolvendo humanos no país.

Paraisópolis é uma favela famosa em São Paulo: a maior favela existente por área (pouco menor que 1 km²) e a segunda maior por população (aproximadamente 70.000 residentes e 17.700 domicílios, segundo o Censo 2010 do IBGE), além de apresentar densidade construída 12 vezes maior que a média da cidade (Carvalho, 2009). A distribuição desigual do espaço público dentro da comunidade prioriza os automóveis particulares em detrimento do transporte coletivo, o que contribui para aumentar o congestionamento e reduzir a velocidade dos ônibus. Além disso, muitas vezes os pedestres são forçados a caminhar nas ruas, devido à presença de vendedores e carros estacionados nas calçadas, onde existentes.

A coleta de dados em Paraisópolis foi realizada em duas fases. Após uma fase piloto, a pesquisa foi realizada entre maio e agosto de 2016, por uma empresa com experiência na área; previa a coleta de dados de 381 indivíduos, sendo que um total de 1.585 visitas foram realizadas. Cada participante teve um aplicativo instalado em seu *smartphone*, programado para coletar dados de localização do GPS a cada 5 minutos, por até 15 dias. Para essa atividade, foram necessários alguns recursos especiais, como compra de créditos para os celulares dos participantes e compra de pacote de dados de internet 3G.

Os participantes responderam também a uma entrevista domiciliar, realizada em formulário impresso, sobre seus dados demográficos e socioeconômicos, como a composição familiar, idade, gênero, faixa de renda, condição de trabalho, nível de instrução e posse de automóvel, e informações complementares sobre seus hábitos de transporte, incluindo os modos de transporte que usam, dificuldades que enfrentam e outras informações de natureza qualitativa.

O aplicativo usado para a coleta de dados foi o *FollowMee*, que grava continuamente, em intervalos fixos, a localização estimada do dispositivo móvel por meio de um receptor de GPS, carregando os dados no servidor. O intervalo adotado de 5 minutos permitiu preservar a bateria do dispositivo de uma forma que não afeta o uso diário do *smartphone*. Algumas vantagens que favoreceram a escolha do *FollowMee* foram: instalação fácil e gratuita, baixo consumo de bateria, o fato de não ser intrusivo para o usuário, além da interface limpa e direta, que facilita sua utilização.

A amostragem foi realizada em dois estágios: primeiro com sorteio de domicílios e, em seguida, entre os indivíduos dos domicílios sorteados. A amostra de domicílios foi selecionada a partir do cadastro de residências do local. Para garantir uma cobertura geográfica mínima, com base nos setores censitários, considerando características construtivas e de ocupação do terreno, a área de Paraisópolis foi segmentada em 13 subáreas. Em cada subárea, foram selecionados no mínimo 20 e no máximo 40 domicílios. Para cada domicílio visitado, foram identificados os indivíduos aptos a participar da pesquisa, sendo necessário possuir um *smartphone* e ter idade maior que 16 anos. Entre os indivíduos aptos, foi sorteado um único indivíduo do domicílio para participação da pesquisa.

Foram realizadas 1.585 tentativas de contato, em até 4 visitas a um domicílio sorteado. Nas visitas, em 92 residências não havia ninguém em casa, 70 domicílios não foram encontrados e outros dois endereços abrigavam imóveis de uso institucional. Dos 799 indivíduos contatados, 271 (33,9%) se recusaram a participar e 147 (18,4%) não possuíam *smartphone*. Em 381 domicílios, foram selecionadas pessoas qualificadas, segundo os critérios preestabelecidos, e que aceitaram participar da pesquisa. Contudo, 21 *smartphones* apresentaram problemas durante a instalação do aplicativo e 88 pessoas desinstalaram o aplicativo logo após concluir a entrevista (2,6% e 11% do total de indivíduos contatados, respectivamente). Portanto, 272 pesquisas foram concluídas com coleta de dados de localização via GPS (34,1% do total de indivíduos contatados). Mais informações sobre a pesquisa de campo podem ser consultadas em Pizzol (2017).

## 4. PROCESSAMENTO DOS DADOS

O método empregado para o processamento dos dados consistiu nas seguintes etapas: a partir dos dados de GPS (*Global Positioning System*) dos *smartphones*, foram identificados os pontos de interesse (POI's – *Points of Interest*) de cada indivíduo e as atividades realizadas nesses POI's, caracterizadas por diversos atributos. Todas as atividades identificadas, independentemente dos indivíduos que a realizaram, foram agrupadas em diversas categorias, usadas então para descrição de um padrão de atividades para cada indivíduo. Uma segunda rodada de clusterização agrupou os indivíduos que apresentavam padrões de atividades semelhantes. Por fim, a composição de cada grupo de indivíduos foi investigada segundo as características demográficas e socioeconômicas das pessoas que os formavam.

### 4.1. Limpeza dos dados e seleção dos indivíduos para análise

Após a coleta dos dados de localização, foi necessário realizar uma limpeza dos dados brutos, para que fosse possível usá-los como entrada no processo de agrupamento das atividades, e em seguida, dos indivíduos.

O processo de limpeza dos dados se iniciou com o conjunto dos 272 indivíduos que não desinstalaram o aplicativo logo após concluir a entrevista. Apesar do aplicativo ter sido

configurado para coletar dados a cada 5 minutos, frequentemente os dados foram registrados a intervalos maiores, principalmente devido à falta ou falha de sinal. Ademais, os dados de GPS foram usados para identificação de atividades, que normalmente acontecem em lugares fechados, nos quais se espera um sinal mais fraco.

Os altos intervalos de tempo observados entre registros de GPS foram tratados de acordo com a seguinte abordagem: indivíduos que tiveram 6% ou mais dos intervalos entre registros de GPS maiores que 30 minutos foram excluídos; partindo dos 272 participantes, a amostra foi reduzida para 189 indivíduos. Grandes porcentagens de altos intervalos estão mais associadas a indivíduos que coletaram menor volume de dados; 6% foi um limite identificado a partir do qual a porcentagem de intervalos maiores que 30 minutos aumentou consideravelmente.

A abordagem adotada para análise do comportamento dos indivíduos se baseou nas atividades que realizaram no período de uma semana completa. Com isso, foram considerados apenas os indivíduos com pelo menos 7 dias consecutivos de dados de GPS, o que reduziu a amostra de 189 para 114 indivíduos. Foram selecionados exatos 7 dias de dados dos indivíduos que coletaram dados por um período maior, a fim de garantir a comparabilidade entre indivíduos. Para metade desses indivíduos, sorteados aleatoriamente, foram selecionados os 7 primeiros dias da coleta e para a outra metade, os 7 últimos dias.

Em seguida, foi realizada uma verificação da compatibilidade entre os dados de localização coletados pelo GPS, visualizados em *software* GIS, e as informações obtidas na entrevista domiciliar. Por exemplo, a partir do georreferenciamento dos endereços de residência indicados no questionário, pode-se observar que alguns indivíduos não apresentavam nuvens de pontos de GPS em sua própria residência, o que pode indicar um possível erro de associação entre a entrevista e o conjunto de dados de GPS. Com isso, mais 5 indivíduos foram eliminados da amostra, a fim de garantir a consistência da base de dados.

Por fim, durante a fase de identificação e caracterização das atividades, foram eliminadas 4 pessoas cujos celulares enviaram dados sempre da mesma localização, isto é, não registraram movimento algum durante o período de análise. Esse procedimento foi realizado a fim de evitar casos possivelmente associados a problemas técnicos dos *smartphones*. Portanto, a amostra final contém 105 indivíduos, cujos dados de GPS e informações coletadas pela entrevista apresentam as características necessárias para identificação de padrões semanais de atividades e análise do perfil demográfico e socioeconômico associado.

### 4.2. Método de identificação de atividades e *Points of Interest* (POI's)

Para identificação das atividades e POI's foi utilizado um algoritmo de clusterização denominado DBSCAN (*Density-Based Spatial Clustering of Application with Noise*). Antes da aplicação do algoritmo, os dados de localização coletados pelo GPS foram interpolados temporalmente a cada 2 minutos. Esse procedimento foi realizado visando melhorar o desempenho do algoritmo em identificar as atividades, dada a maior frequência de dados espaciais resultante, em relação aos registros originais. Por exemplo, se uma dada localização foi registrada no instante 0', e a próxima localização foi registrada no instante 6', registros "artificiais" foram criados a cada 2 minutos, nos instantes 2' e 4', com longitude e latitude iguais às do registro do instante 0'.

Proposto por Ester et al. (1996), o DBSCAN é um método de clusterização baseado na

densidade espacial de pontos, efetivo para identificar clusters de formato arbitrário e de diferentes tamanhos, separar os ruídos dos dados e detectar clusters "naturais" e seus arranjos dentro do espaço de dados, sem qualquer informação preliminar sobre os grupos, ao contrário de algoritmos não hierárquicos de clusterização tradicionais (por exemplo, *K-means*).

O DBSCAN baseia-se em dois parâmetros para definição de um cluster. O primeiro parâmetro define a distância máxima (*Eps*) entre dois pontos a serem incluídos no mesmo cluster e o segundo determina o número mínimo de pontos (*MinPts*) para que um cluster seja formado. O processo foi realizado em dois estágios, sendo a identificação de atividades o primeiro estágio, seguido pela identificação de POI's.

*4.2.1. Identificando atividades*

No primeiro estágio, uma distância equivalente foi considerada, combinando distância e tempo entre cada par de pontos do GPS. Assim, visitas diversas ao mesmo local (por exemplo, em diferentes períodos do dia ou dias distintos) foram identificadas como atividades distintas. Tais atividades são formadas por clusters espaço-temporais de pontos do GPS, sendo a localização geográfica da atividade o centroide de todos os pontos do cluster.

Os parâmetros considerados para o processamento do DBSCAN neste primeiro estágio foram distância máxima de 100 metros e tempo máximo de 30 minutos entre dois pontos consecutivos, e mínimo de 15 pontos para formação de um cluster. Os valores foram definidos a partir de uma análise de sensibilidade, seguida por inspeção visual.

*4.2.2. Identificando POI´s*

A localização geográfica da atividade, estimada no primeiro estágio, pode variar mesmo para atividades que ocorreram no mesmo local, em função da imprecisão do GPS. Devido a isso, o segundo estágio consistiu na identificação de uma única localização geográfica para representar esse conjunto de atividades que ocorreram no mesmo local, porém em períodos ou dias distintos.

Nesta etapa, as diferentes atividades foram dados de entrada para o algoritmo de clusterização, que considerou apenas a distância euclidiana entre elas. Dessa forma, atividades próximas umas das outras foram agrupadas em um único cluster, representando um local significativo para o indivíduo, ou seja, um *Point of Interest*. A localização geográfica do POI foi configurada como o centroide das atividades que o formaram. Os parâmetros usados para o processamento do DBSCAN foram distância máxima de 100 metros e número mínimo de pontos igual a 1, uma vez que mesmos locais visitados apenas uma vez ao longo da semana devem ser considerados um POI do indivíduo.

**4.3. Caracterização das atividades**

A identificação de categorias de atividades passa pela descrição de alguns de seus atributos, uma vez que não é conhecida a natureza das atividades realizadas (trabalho, estudo, compras, saúde, etc.). Para tanto, uma abordagem multidimensional foi usada, incluindo todos os atributos disponíveis e relevantes para compor o perfil de uma atividade. As variáveis consideradas foram divididas em três categorias, apresentadas na Tabela 1.

**4.4. Métodos de agrupamento**

A análise de agrupamentos, como conjunto de técnicas mais estruturado, teve origem na

década de 30 do século passado, nos campos da Antropologia e Psicologia, conforme discutido por Favero e Fávero (2017). Segundo os autores, a análise de agrupamentos representa um conjunto de técnicas exploratórias que podem ser aplicadas quando há a intenção de se verificar a existência de comportamentos semelhantes entre observações em relação a determinadas variáveis, com o objetivo de identificar grupos, ou *clusters*, homogêneos internamente e heterogêneos entre si. Os métodos de agrupamentos foram usados neste estudo para formar grupos de atividades e, posteriormente, de indivíduos.

**Tabela 1:** Variáveis usadas para a caracterização das atividades

| Categoria da variável | Nome da variável | Descrição da variável |
|---|---|---|
| Espacial | *Local* | Indicação se a atividade ocorreu na residência ou em outro local, dentro ou fora de Paraisópolis. |
| | *Local_antes* | Indicação se a atividade anterior ocorreu na residência, dentro ou fora de Paraisópolis. |
| | *Local_depois* | Indicação se a atividade posterior ocorreu na residência, dentro ou fora de Paraisópolis. |
| | *Dist_resid* | Distância linear do POI da atividade à residência do indivíduo, em metros. |
| | *Dist_antes* | Distância linear do POI da atividade ao POI da atividade anterior, em metros. |
| | *Dist_depois* | Distância linear do POI da atividade ao POI da atividade posterior, em metros. |
| | *Uso_solo* | Uso predominante do solo na quadra do POI da atividade. |
| Temporal | *Dia_útil_fds* | Indicação se a atividade ocorreu em um dia útil ou no fim de semana (“fds”). |
| | *Hora_início* | Horário de início da atividade, no formato HHMM. |
| | *Duração*[1] | Duração da atividade, no formato HHMM. |
| Repetição e Sequência | *Freq_vezes* | Número de vezes que o POI da atividade foi visitado pelo indivíduo no período. |
| | *Freq_dias* | Número de dias em que o POI da atividade foi visitado pelo indivíduo no período. |
| | *Ordem_atividade* | Ordem da atividade do indivíduo no dia. |

[1] Em função da imprecisão dos horários de início e fim da atividade, em decorrência de intervalos de tempo superiores a 5 minutos entre registros consecutivos do GPS, formulações alternativas foram incluídas para representar a duração da atividade.

### *4.4.1. Método Hierárquico*

Os esquemas de aglomeração hierárquicos podem ser aglomerativos ou divisivos, de acordo com a forma que é iniciado o processo. Neste estudo foi utilizado apenas o processo aglomerativo, no qual todas as observações são inicialmente consideradas separadas e, a partir de suas distâncias, são formados os grupos até que se chegue a um estágio final com apenas um agrupamento (Favero e Fávero, 2017). Entre os esquemas hierárquicos aglomerativos, foi usado o método de encadeamento completo, que considera as maiores distâncias entre observações pertencentes a clusters diferentes, para que sejam formados novos clusters a cada estágio de agrupamento, pela incorporação de observações ou grupos.

### *4.4.2. K-means*

Os esquemas de aglomeração não hierárquicos, entre os quais o procedimento mais popular é o *k-means*, ou k-médias, consistem em processos em que são definidos centros dos clusters e as observações são alocadas segundo sua proximidade a eles. O método requer a definição do número de clusters *a priori*, a partir do qual são definidos os centros dos agrupamentos e em seguida alocadas as observações. Por essa razão, quando não há uma estimativa razoável da quantidade de clusters que podem ser formados a partir dos dados disponíveis, recomenda-se usar um esquema de aglomeração hierárquico para estimativa do número de clusters, e então usá-lo como entrada para o esquema de aglomeração não hierárquico (Favero e Fávero, 2017).

*4.4.3. TwoStep Clustering*

O procedimento *TwoStep Cluster Analysis* do *IBM SPSS Statistics* foi desenvolvido para atender duas situações principais: lidar com grandes bancos de dados e formar clusters a partir de variáveis categóricas (ou qualitativas) e contínuas (ou quantitativas) usadas em conjunto, o que não é possível com outros métodos. Conforme indica o próprio nome, o algoritmo é processado em duas etapas (IBM, 2011). O primeiro passo consiste na formação de pré-clusters, com o objetivo de reduzir a dimensão da matriz de distâncias entre todos os possíveis pares de observações. Os pré-clusters são usados no lugar dos dados originais, na segunda etapa do método, que consiste na aplicação do método hierárquico padrão. Como o número de subgrupos é muito menor do que o número de registros originais, o método tradicional de agrupamento pode ser usado de forma eficiente.

## 5. APRESENTAÇÃO E ANÁLISE DOS RESULTADOS

O método aplicado neste estudo consistiu em classificar todas as atividades de todos os indivíduos em um número de categorias reduzido, a partir das quais foram compostos padrões semanais de atividades individuais. Os resultados obtidos são apresentados a seguir. Mais informações sobre o processo de classificação podem ser obtidas em Pizzol (2018).

### 5.1. Classificação de atividades

Para a classificação das atividades, o método de agrupamentos usado foi o *TwoStep Clustering*, pela possibilidade de se trabalhar com variáveis qualitativas. Em um processo de tentativas e análises sucessivas, foram analisados 9 cenários de agrupamentos distintos, variando o número de clusters e variáveis consideradas. O número de clusters ótimo foi obtido a partir de análise semântica dos resultados do algoritmo e também com base no valor da estatística *silhouette* (Rousseeuw, 1987), resultando em 6 grupos diferentes de atividades. Entre os 6 clusters de atividades, foram identificados 2 clusters com a maioria das atividades realizadas em casa, 2 clusters com a maioria das atividades realizadas em Paraisópolis e 2 clusters com a maioria das atividades realizadas fora de Paraisópolis. A Tabela 2 apresenta a alocação das atividades entre os clusters resultantes, os quais são analisados em seguida.

**Tabela 2:** Alocação das atividades por cluster

| Cluster de atividades | Número de atividades | % |
|---|---|---|
| Cluster 1 | 254 | 18,3 |
| Cluster 2 | 260 | 18,7 |
| Cluster 3 | 189 | 13,6 |
| Cluster 4 | 130 | 9,4 |
| Cluster 5 | 238 | 17,1 |
| Cluster 6 | 318 | 22,9 |
| Total | 1.389 | 100,0 |

*5.1.1. Clusters 1 e 2 - Atividades na Residência*

Os clusters de atividades realizadas na residência se diferenciam principalmente quanto à duração da atividade, sendo o Cluster 1 composto, majoritariamente, por atividades mais longas e o Cluster 2 por atividades de duração curta a média. Quanto ao horário de início da atividade, em ambos os clusters a maioria das atividades se inicia no fim da tarde ou à noite, a partir das 17h30, o que provavelmente caracteriza a chegada em casa para passar a noite, no caso das atividades longas. Em ambos os clusters, ainda que em menor proporção, pode-se observar a presença de atividades que se iniciam em diversos horários, que podem representar

voltas para casa ao longo do dia, como para almoço ou outras situações pontuais.

No que se refere ao uso do solo dos POI's referentes às atividades desses clusters, a grande maioria dos registros foram identificados como terrenos vagos, acentuando o caráter de assentamento informal da favela, sem registros oficiais sobre os domicílios. A maioria das atividades, em ambos os clusters, foram identificadas como a segunda atividade realizada pelos indivíduos no dia, o que pode indicar o retorno para casa depois de atividades externas.

*5.1.2. Clusters 4 e 5 - Atividades em Paraisópolis*

Em relação aos clusters de atividades que ocorrem em sua maioria em Paraisópolis, podem ser identificados dois tipos razoavelmente bem definidos: Cluster 4 com atividades de duração média a longa mais frequentes com início pela manhã, e o Cluster 5 com atividades de duração curta pouco frequentes, em sua maioria, uma visita por semana, com início no começo de tarde. As atividades curtas apresentam duração de até 2 horas, enquanto as atividades médias a longas duram de 2 a 6 horas, principalmente.

A caracterização do uso do solo dos POI's das atividades realizadas na região de Paraisópolis se divide entre terrenos vagos e residencial de médio ou alto padrão. Isto pode indicar atividades de trabalho realizadas pelos residentes de Paraisópolis, na prestação de serviços às residências de bairros ricos adjacentes, como o Morumbi, que foram incluídas em um buffer de 100 metros delimitado a partir dos setores censitários que compõem Paraisópolis.

*5.1.3. Clusters 3 e 6 - Atividades fora de Paraisópolis*

Também foram identificados dois tipos razoavelmente bem definidos de clusters com atividades que ocorrem fora de Paraisópolis: Cluster 3 com atividades de duração média a longa mais frequentes e o Cluster 6 com atividades de duração curta pouco frequentes. As atividades mais longas do Cluster 3 podem indicar atividades principais realizadas fora de Paraisópolis, como trabalho ou estudo, com duração entre 6 e 9 horas em sua maioria. Em relação às atividades curtas do Cluster 6, a maior parte apresenta duração de até 1 hora, mas há também atividades com duração de até 3 horas, o que pode representar compras, refeições, tarefas, ou pequenos serviços realizados ocasionalmente fora de Paraisópolis.

A maioria das atividades do Cluster 3 ocorreu em POI's frequentados em 5 dias da semana, indicando um padrão regular, enquanto a maioria das atividades do Cluster 6 ocorreu em POI's visitados apenas 1 vez na semana, indicando um padrão de menor regularidade. O horário das atividades desses clusters reforçam essa conclusão: a maioria das atividades do Cluster 3 teve início pela manhã, entre 6h30 e 10h30, enquanto as atividades do Cluster 6 apresentaram horário de início bastante diverso, com a maioria das atividades se iniciando das 12h30 às 15h30, apesar de ocorrerem ao longo de todo o dia. Ademais, a maior parte das atividades do Cluster 3 é a primeira atividade realizada no dia, o que reforça a possibilidade de ser uma atividade principal como estudo ou trabalho.

A partir da análise do uso do solo, é possível observar maioria de comércios e serviços para as atividades do Cluster 6 e algum uso misto. Para o Cluster 3, por outro lado, a maioria do uso do solo é residencial de médio ou alto padrão, com alguma presença de comércio e serviços e uso misto. A predominância de residencial de médio ou alto padrão, para os POI's onde são realizadas as atividades longas do Cluster 3, pode indicar prestação de serviços a residências de alta renda em outros bairros, ao mesmo tempo que a presença de comércios e serviços pode

indicar atividades de trabalho em lojas, escritórios ou outros estabelecimentos comerciais.

### 5.2. Classificação de indivíduos

A escolha das variáveis para a clusterização dos indivíduos baseou-se em três diferentes critérios para descrição do padrão de atividades individual: intensidade, variação e repetição. A intensidade do padrão de atividades expressa se o indivíduo é muito ou pouco ativo, ou seja, está relacionado ao número de atividades realizadas. A variação do padrão de atividades expressa se o indivíduo realiza atividades de tipos variados, enquanto a repetição do padrão de atividades expressa se o indivíduo realiza muitas atividades de um mesmo tipo, segundo as 6 categorias de atividades previamente definidas. As variáveis são apresentadas na Tabela 3.

**Tabela 3:** Variáveis usadas para o agrupamento dos indivíduos

| Variável por indivíduo | Critério de caracterização do padrão de atividade | Descrição da variável |
|---|---|---|
| *Num_atividades* | Intensidade | Número total de atividades realizadas no período. |
| *Num_clusters* | Variação | Número de clusters de atividades em que foi identificada pelo menos uma atividade do indivíduo. |
| *Max_rep* | Repetição | Máxima repetição de atividades em um mesmo cluster. |
| *Max_rep_ativ* | Repetição | *Max_rep* / *Num_atividades* |
| *Clusters* | Intensidade / Variação / Repetição | Número de atividades classificadas em cada um dos seis clusters. |
| *%Clusters* | Variação / Repetição | Porcentagem de atividades por cluster de atividade. |
| *Num_pois* | Intensidade / Variação | Número total de POIs visitados no período. |
| *Num_atividades_poi* | Repetição | Número de atividades por POI = *Num_atividades* / *Num_pois* |

Devido ao fato de as variáveis consideradas serem escalares, os três métodos de agrupamento puderam ser aplicados, visando a comparação dos resultados. Inicialmente, foi aplicado o método Hierárquico, cuja análise do dendrograma possibilitou estimar a quantidade de clusters, a qual pode ser então utilizada como entrada para aplicação e avaliação dos métodos *K-means* e *TwoStep*. Após realizar diversos testes, os resultados foram semelhantes para os dois métodos, tendo sido selecionados 7 grupos de indivíduos para representação de padrões semanais de atividades distintos, apresentados na Tabela 4.

**Tabela 4:** Padrões semanais de atividades dos clusters de indivíduos

| Cluster de indivíduos | Padrão de atividade | Descrição | Número de indivíduos | % |
|---|---|---|---|---|
| Cluster 1 | Casa | Indivíduos que realizam a maior parte de suas atividades em suas residências. | 11 | 10,5 |
| Cluster 2 | Casa – Fora | Indivíduos que realizam parte de suas atividades em suas residências e parte de suas atividades fora de Paraisópolis, sem apresentar, no entanto, um padrão característico de uma atividade regular principal, como trabalho ou estudo. | 15 | 14,3 |
| Cluster 3 | Atividade Principal / Dentro | Indivíduos que apresentam um padrão característico de uma atividade regular principal, como trabalho ou estudo, dentro de Paraisópolis. | 10 | 9,5 |
| Cluster 4 | Atividade Principal / Dentro / Variadas | Indivíduos que apresentam um padrão característico de uma atividade regular principal, como trabalho ou estudo, dentro de Paraisópolis, e também realizam outros tipos de atividades de forma variada. | 22 | 21,0 |
| Cluster 5 | Atividade Principal / Fora | Indivíduos que apresentam um padrão característico de uma atividade regular principal, como trabalho ou estudo, fora de Paraisópolis. | 10 | 9,5 |
| Cluster 6 | Atividade Principal / Fora / Variadas | Indivíduos que apresentam um padrão característico de uma atividade regular principal, como trabalho ou estudo, fora de Paraisópolis, e também realizam outros tipos de atividades de forma variada. | 25 | 23,8 |
| Cluster 7 | Atividade Principal / | Indivíduos que apresentam um padrão característico de uma atividade | 12 | 11,4 |

| | | | | |
|---|---|---|---|---|
| | Fora / Ativos | regular principal, como trabalho ou estudo, fora de Paraisópolis, e também realizam outros tipos de atividades de forma variada e bastante ativa, com alto número de atividades observadas no período. | | |
| Total | | | 105 | 100,0 |

### 5.3. Caracterização demográfica e socioeconômica dos grupos de indivíduos

Os indivíduos de cada grupo foram analisados segundo três eixos principais, a partir dos dados obtidos pela entrevista domiciliar: pessoal e família, estudo e trabalho, e transporte e exercício físico. A Tabela 5 apresenta a caracterização dos grupos de indivíduos.

**Tabela 5:** Caracterização demográfica e socioeconômica dos clusters de indivíduos

| Cluster de indivíduos | Padrão de atividade | Caracterização demográfica e socioeconômica |
|---|---|---|
| Cluster 1 | Casa | Grupo que contém mais mulheres do que homens; em sua maioria os indivíduos são mais velhos e as famílias maiores. São pessoas que moram há mais tempo em Paraisópolis e apresentam rendas mais baixas. Não trabalham ou são autônomos, que trabalham no geral 6 ou 7 dias por semana, acima de 8 horas por dia. Alguns deles tem bicicleta ou praticam exercício físico. |
| Cluster 2 | Casa – Fora | Grupo em que grande parte de indivíduos mora com filhos ou enteados, sendo que desses, a maioria é mulher. São em sua maioria pessoas mais velhas e casadas, com famílias maiores. Moram há menos tempo em Paraisópolis e apresentam rendas mais baixas. Não trabalham ou são autônomos, que trabalham no geral todos os dias, acima de 8 horas por dia. É o grupo com maior posse de automóvel. |
| Cluster 3 | Atividade Principal / Dentro | Grupo com maioria de mulheres, pessoas mais jovens e solteiras. Moram há mais tempo em Paraisópolis, apresentam rendas mais baixas e caracterizam-se por trabalhar na favela, com ou sem carteira assinada. Trabalham no geral 5 ou 6 dias por semana, usualmente acima de 7 horas por dia. É o grupo com maior posse de moto e maior uso de transporte individual para sair ou fora da favela. |
| Cluster 4 | Atividade Principal / Dentro / Variadas | Grupo com notável maioria de mulheres com filhos, e também indivíduos mais jovens e solteiros. Moram há mais tempo em Paraisópolis e apresentam rendas mais baixas. É composto em grande parte por estudantes, mas também desempregados e trabalhadores com e sem carteira assinada. Trabalham 5 ou 6 dias por semana, acima de 8 horas por dia. Usam transporte coletivo e individual para sair ou fora da favela. |
| Cluster 5 | Atividade Principal / Fora | Grupo em que a maioria dos indivíduos mora com filhos ou enteados, apresentando, contudo, famílias menores. Moram há menos tempo em Paraisópolis e possuem rendas pessoais mais altas. Caracterizam-se por trabalhar fora de Paraisópolis com carteira assinada e trabalham, em média, 5 ou 6 dias por semana e 8 horas por dia. Ninguém desse grupo possui carro ou moto, mas são os que mais possuem bicicleta e praticam exercício físico. São os mais dependentes de transporte coletivo para sair ou fora da favela. |
| Cluster 6 | Atividade Principal / Fora / Variadas | Grupo com maior porcentagem de pessoas que vivem sozinhas, caracterizando-se, portanto, por famílias menores. Moram há mais tempo em Paraisópolis, apresentam rendas mais altas e maior grau de escolaridade. Caracterizam-se por trabalhar fora de Paraisópolis, principalmente com carteira assinada, apesar de alguns trabalharem sem carteira assinada. No geral, trabalham 5 ou 6 dias por semana e 8 horas por dia. Também são dependentes de transporte coletivo para sair ou fora da favela. |
| Cluster 7 | Atividade Principal / Fora / Ativos | Grupo de famílias pequenas, que moram há mais tempo em Paraisópolis. Apresentam rendas mais altas e maior grau de escolaridade. Caracterizam-se por trabalhar fora de Paraisópolis, principalmente com carteira assinada, apesar de também haver alguns autônomos. Em sua maioria, trabalham 5 ou 6 dias por semana, 8 horas por dia. Alguns têm moto ou bicicleta e praticam exercício físico. |

Os resultados são coerentes com aqueles obtidos por Pas e Koppelman (1987), que mostraram que indivíduos com menos restrições econômicas, profissionais e familiares apresentam níveis mais altos de variabilidade intrapessoal em seus padrões de atividades. Além disso, segundo Susilo e Axhausen (2014), a repetitividade dos locais e atividades é altamente influenciada pelos compromissos dos indivíduos fora de casa e pelas condições internas do domicílio.

## 6. CONSIDERAÇÕES FINAIS

Este estudo investigou o comportamento de atividades de 105 residentes de Paraisópolis, usando dados de GPS coletados por uma semana, a partir de um aplicativo instalado nos *smartphones*. A partir da classificação das atividades em seis categorias diferentes, os indivíduos puderam ser classificados em sete grupos distintos quanto ao seu padrão semanal de atividades. Por fim, foram analisadas as características socioeconômicas de cada grupo. Foi constatado que cerca de 56% dos indivíduos apresentam padrão de atividades variado, o que reflete a importância de se considerar outros tipos de padrões de atividades, além dos padrões regulares casa-trabalho, nos modelos de demanda por viagens.

Pode-se dizer que o método empregado facilitou a coleta de dados de múltiplos dias, permitindo o conhecimento da localização das pessoas continuamente. Com isso, é ressaltada a importância do uso da tecnologia para obter melhores informações sobre os padrões de atividade dos indivíduos, apesar dos desafios impostos por grupos de difícil acesso. Embora haja dificuldades em empregar novos métodos de coleta de dados, os potenciais benefícios que esses métodos trazem podem superar os riscos existentes. O trabalho atinge, assim, os objetivos de validação de uma metodologia inovadora e obtenção de resultados relevantes para a análise do comportamento relacionado à demanda por transportes.

**Agradecimentos**

Os autores agradecem ao CNPq, pela Bolsa de Produtividade em Pesquisa (PQ) concedida a dois coautores, e ao pesquisador Renato Arbex, pelo apoio no desenvolvimento e implementação do algoritmo DBSCAN.

**REFERÊNCIAS BIBLIOGRÁFICAS**

Ampt, E. e Hickman, M. (2015) Workshop synthesis: Survey methods for hard-to-reach groups and modes. *Transportation Research Procedia*, v. 11, p. 475-480.

Anda, C.; Erath, A. e Fourie, P. J. (2017) Transport modelling in the age of big data. *International Journal of Urban Sciences*, v. 21, n. 1, p. 19-42.

Behrens, R.; Freedman, M. e McGuckin, N. (2009) The challenges of surveying 'hard to reach' groups: Synthesis of a workshop. *In:* Bonnel, P.; Lee-Gosselin, M.; Zmud, J. e Madre, J. L. (*eds.*) *Transport Survey Methods: Keeping up with a Changing World*. Emerald Group Publishing Limited, Bingley.

Carvalho, E. T. (2009) *Plano de Urbanização da Comunidade Paraisópolis – Primeira Etapa: Resultados no Setor de Transporte Público*. São Paulo Transporte S.A., São Paulo.

Chen, C.; Ma, J.; Susilo, Y.; Liu, Y. e Wang, M. (2016) The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation Research Part C: Emerging Technologies*, v. 68, p. 285-299.

Ester, M.; Kriegel, H. P.; Sander, J. e Xu, X. (1996) A density-based algorithm for discovering clusters in large spatial databases with noise. *KDD Proceedings*, v. 96, n. 34, p. 226-231.

Favero, L. e Fávero, P. (2017) *Análise de Dados: Técnicas Multivariadas Exploratórias com SPSS e Stata*. Elsevier Brasil, São Paulo.

Hanson, S. e Huff, J. O. (1981) Assessing day-to-day variability in complex travel patterns. *Transportation Research Record*, v. 891, p. 18-24.

IBM (2011) *IBM SPSS Statistics 20 Algorithms*. IBM Corporation, Armonk.

Koch, J.; Lindau, L. A. e Nassi, C. D. (2013) Transportation in the Favelas of Rio de Janeiro. Lincoln Institute of Land Policy Working Paper.

Maia, M. L.; Lucas, K.; Marinho, G.; Santos, E. e de Lima, J. H. (2016) Access to the Brazilian City – from the perspectives of low-income residents in Recife. *Journal of Transport Geography*, v. 55, p. 132-141.

Pas, E. I. e Koppelman, F. S. (1987) An examination of the determinants of day-to-day variability in individuals' urban travel behavior. *Transportation*, v. 13, n. 2, p. 183-200.

Pizzol, B.; Alves, B. B.; Giannotti, M. A.; Strambi, O.; Arbex, R. e Bruni, L. R. (2017) Travel survey tools and methods: challenges on surveys at slums. Artigo apresentado em *11th International Conference on Transport Survey Methods*, Esterel, Canada.

Pizzol, B. (2018) Padrões de atividades de residentes de Paraisópolis: análise de dados de múltiplos dias coletados com smartphones. Dissertação de Mestrado. Escola Politécnica, Universidade de São Paulo.

Rousseeuw, P. J. (1987) Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of Computational and Applied Mathematics*, v. 20, p. 53-65.

Stopher, P.; FitzGerald, C. e Zhang, J. (2008) Search for a global positioning system device to measure person travel. *Transportation Research Part C: Emerging Technologies*, v. 16, n. 3, p. 350-369.

Susilo, Y. O. e Axhausen, K. W. (2014) Repetitions in individual daily activity–travel–location patterns: a study using the Herfindahl–Hirschman Index. *Transportation*, v. 41, n. 5, p. 995-1011.