BRUNA PIZZOL

**Padrões de atividades de residentes de Paraisópolis:
Análise de dados de múltiplos dias coletados com *smartphones***

São Paulo
2018

BRUNA PIZZOL

**Padrões de atividades de residentes de Paraisópolis:
Análise de dados de múltiplos dias coletados com *smartphones***

Dissertação apresentada à Escola
Politécnica da Universidade de São Paulo
para obtenção do título de Mestre em
Ciências.

São Paulo
2018

BRUNA PIZZOL

**Padrões de atividades de residentes de Paraisópolis:
Análise de dados de múltiplos dias coletados com *smartphones***

Dissertação apresentada à Escola
Politécnica da Universidade de São Paulo
para obtenção do título de Mestre em
Ciências.

Área de concentração: Engenharia de
Transportes

Orientador: Prof. Dr. Orlando Strambi

São Paulo
2018

Autorizo a reprodução e divulgação total ou parcial deste trabalho, por qualquer meio convencional ou eletrônico, para fins de estudo e pesquisa, desde que citada a fonte.

Este exemplar foi revisado e corrigido em relação à versão original, sob responsabilidade única do autor e com a anuência de seu orientador.

São Paulo, _____ de _____ de _____

Assinatura do autor: _____

Assinatura do orientador: _____

Catálogo-na-publicação

Pizzol, Bruna

Padrões de atividades de residentes de Paraisópolis: Análise de dados de múltiplos dias coletados com smartphones / B. Pizzol -- versão corr. -- São Paulo, 2018.
185 p.

Dissertação (Mestrado) - Escola Politécnica da Universidade de São Paulo. Departamento de Engenharia de Transportes.

1. Engenharia de transportes 2. Planejamento de transportes
3. Mobilidade urbana I. Universidade de São Paulo. Escola Politécnica. Departamento de Engenharia de Transportes II. t.

AGRADECIMENTOS

Tenho de admitir que minha trajetória durante o mestrado foi repleta de dúvidas “cabeludas”: qual caminho seguir, como selecionar um bom artigo, como escrever concisa e claramente, qual ferramenta usar, como fazer algo que faça sentido e tenha valor na comunidade científica (e também fora dela), como me organizar para conciliar trabalho e estudo, como decidir entre trabalho e estudo (e sobreviver “apenas” estudando) e quem era a pessoa que eu estava me tornando. Nada disso me assustou, pois estou acostumada a ter dúvidas. E, devo confessar, chego a gostar delas.

No entanto, sem dúvidas, foram as inúmeras certezas que tenho o privilégio de ter na vida que me motivaram a continuar e poder concluir essa etapa:

O apoio incondicional dos meus pais Lucimara e Francisco, sempre tão presentes, responsáveis e carinhosos;

A amizade da minha irmã Gabriela, que me entende e chega a ser quase como eu;

O apadrinhamento do vô Zé e da vó Dalva, que assumiram essa missão com louvor desde o início e cuidam de mim como filha;

O companheirismo do Marcos, que estando perto ou longe, consegue sempre aquecer meu coração;

Poder contar invariavelmente com meus bons amigos de Piracicaba e São Paulo, que me conhecem tão bem há tantos anos;

A proximidade da família Gambi, que domina a arte de dar à vida mais alegria;

Os ensinamentos precisos e poderosos e o carinho imensurável dos mestres yoguis Adriana e Edmilson;

O apoio e companhia do Renato, que me ajudou com os dados e sábios conselhos;

O acolhimento da professora e amiga Mariana, que para mim é uma grande inspiração dentro e fora da vida acadêmica;

A parceria da chefe e amiga Bianca, que me proporcionou oportunidades ímpares de desenvolvimento pessoal e profissional;

A orientação do meu querido professor Orlando, que me acompanhou de perto nesta caminhada, sempre de forma atenciosa (e criteriosa), leve e alegre, o que fez toda a diferença para esta experiência ser enriquecedora e garantiu a satisfação que estou sentindo hoje.

Com essas certezas, a cada dia eu tive a certeza de que o processo em si já estava valendo a pena. E isso eu devo a todas essas pessoas amadas. Portanto, meu sincero obrigada!

Agradeço, por fim, aos acontecimentos corriqueiros da vida, por vezes tão contraditória, por me mostrar, cada vez mais claramente, como somos pequenos, e na nossa pequenez, o caminho para sermos mais grandiosos com nós mesmos e com os outros, todos os dias.

RESUMO

PIZZOL, B. **Padrões de atividades de residentes de Paraisópolis: Análise de dados de múltiplos dias coletados com *smartphones***. 2018. 185 p. Dissertação (Mestrado em Engenharia de Transportes) – Escola Politécnica da Universidade de São Paulo, São Paulo, 2018.

Esse trabalho propõe articular dois eixos principais de análise: investigar os padrões de atividades de 105 residentes de Paraisópolis, segunda maior favela de São Paulo, e avaliar o processo de coleta e processamento de dados de múltiplos dias obtidos a partir de um aplicativo instalado nos *smartphones*. Os dados foram obtidos por meio de uma entrevista domiciliar realizada 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 de atividades dos moradores. Uma segunda rodada de classificação agrupou os indivíduos que apresentavam padrões semanais de atividades semelhantes, em sete grupos distintos.

A composição de cada grupo de indivíduos foi investigada segundo as características socioeconômicas das pessoas que os formavam, permitindo avaliar se os grupos com comportamento de atividades similar apresentavam também homogeneidade no que diz respeito ao gênero, idade, renda, ocupação, nível de escolaridade, entre outros atributos.

Os resultados indicam que entre os sete grupos de indivíduos identificados, dois grupos realizam a maior parte de suas atividades na residência; outros dois grupos possuem uma atividade principal dentro de Paraisópolis, sendo que um deles apresenta padrão mais regular e o outro, padrão mais variado; e outros três grupos apresentam uma atividade principal fora de Paraisópolis, sendo o primeiro com padrão mais regular, o segundo com padrão mais variado e um terceiro grupo com padrão muito variado e ativo, com alto número de atividades realizadas no período.

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.

Palavras-chave: Padrões de atividades. Dados de múltiplos dias. Classificação de indivíduos. Comportamento individual. Smartphone. GPS. Favela. Paraisópolis.

ABSTRACT

PIZZOL, B. **Activity patterns of residents of Paraisópolis: analysis of multiple days of data collected with smartphones.** 2018. 185 p. Master thesis (Master in Transportation Engineering) – Polytechnic School of University of São Paulo. São Paulo. 2018.

This dissertation explores two main lines of research: to investigate the activity patterns of 105 residents of the Paraisópolis, second largest slum of São Paulo, using data collected over multiple days using smartphones. 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 of individuals was then investigated regarding its socioeconomic and demographic composition, to verify if similarity of behavior was associated with homogeneity concerning gender, age, income, occupation, level of education, among other attributes.

Results suggest that among the seven clusters of individuals, two of them conduct most of their activities at home; individuals of other two groups seem to have a main frequent regular activity outside their home but within the area of Paraisópolis, one with greater diversity of activities than the other; finally, individuals of the three remaining groups participate in a frequent regular activity outside Paraisópolis, differing among themselves concerning variability and number of other activities conducted during one week.

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.

Keywords: Activity patterns. Multiple days data. Classification of individuals. Individual behavior. Smartphone. GPS. Slum. Paraisópolis.

LISTA DE FIGURAS

Figura 1. Etapas do método. Fonte: Elaboração própria.	21
Figura 2. Processo de caracterização dos padrões de atividades e viagens dos indivíduos. Fonte: Elaboração própria.....	32
Figura 3. Exemplo de cálculo do espaço de atividades pelo segundo momento dos locais de atividades. Fonte: Susilo e Kitamura, 2005.	35
Figura 4. Localização de Paraisópolis. Fonte: Elaboração própria, com dados do Metrô de São Paulo e CEM – Centro de Estudos da Metrópole.....	43
Figura 5. Vista aérea de Paraisópolis. Fonte: Google Earth, 2016.	44
Figura 6. Acidentes de trânsito 2010 e 2017, em Paraisópolis e região. Fonte: Elaboração própria, com dados da Companhia de Engenharia de Tráfego (CET-SP).	45
Figura 7. Linhas de ônibus que passam por Paraisópolis e pontos de ônibus localizados no entorno. Fonte: Elaboração própria com dados do GeoSampa.....	46
Figura 8. Linhas de ônibus, metrô e trem na região de Paraisópolis. Fonte: Elaboração própria com dados do GeoSampa.	47
Figura 9. Aplicação da primeira fase da pesquisa em Paraisópolis (outubro de 2015). Foto: Irene Quintáns.	49
Figura 10. Treinamento de pesquisadores de campo no LabGeo. Foto: Cobrape.	49
Figura 11. Subáreas de amostragem dos domicílios. Fonte: Elaboração própria.....	51
Figura 12. Formulário de acompanhamento de campo. Foto: Irene Quintáns.	52
Figura 13. <i>Website</i> de acompanhamento da pesquisa, com dados e gráficos em tempo real sobre o progresso da coleta. Fonte: Cobrape.	53
Figura 14. Mapa georreferenciado para identificação dos domicílios visitados. Fonte: Cobrape.	54
Figura 15. Polígono de frequências de dias de coleta por indivíduo. Fonte: Elaboração própria.	58
Figura 16. Etapas do método desenvolvidas no Capítulo 4. Fonte: Elaboração própria.....	59
Figura 17. Distribuição de intervalos entre os registros de GPS. Fonte: Elaboração própria.	61

Figura 18. Relação entre o número de registros por indivíduo e a porcentagem de intervalos maiores que 30 minutos. Fonte: Elaboração própria.	62
Figura 19. Porcentagem de intervalos maiores que 30 minutos por indivíduo. Fonte: Elaboração própria.	62
Figura 20. Processo de seleção dos indivíduos para análise. Fonte: Elaboração própria. ...	63
Figura 21. Ilustração de clusters identificados pelo DBSCAN. Fonte: Ester et al., 1996.....	65
Figura 22. Exemplo de ilustração da distância equivalente usada para identificação das atividades. Fonte: Adaptação de esquema proposto por Renato Arbex.	65
Figura 23. Inspeção visual da distribuição espacial dos Pontos de Interesse (POI's) para um indivíduo selecionado aleatoriamente. Fonte: Elaboração própria, com imagens do <i>Google Maps</i>	69
Figura 24. Ilustração dos conceitos de <i>Duração Curta</i> e <i>Duração Longa</i> das atividades. Fonte: Elaboração própria.	79
Figura 25. Etapas do método desenvolvidas no Capítulo 5. Fonte: Elaboração própria.....	85
Figura 26. Lógica para elaboração da análise de agrupamentos. Fonte: Favero e Fávero, 2017.	86
Figura 27. Esquemas de aglomeração em análise de agrupamentos. Fonte: Favero e Fávero (2017).	88
Figura 28. Exemplo de dendrograma para o método de encadeamento completo, com distância euclidiana. Fonte: <i>Minitab Support</i>	89
Figura 29. Exemplo de ilustração do procedimento <i>k-means</i> . Fonte: <i>Maciej Pacula</i>	90
Figura 30. Porcentagem de atividades por cluster. Fonte: Elaboração própria.....	99
Figura 31. Clusters de atividades realizadas em sua maioria na residência.....	101
Figura 32. Clusters de atividades realizadas em sua maioria dentro e fora de Paraisópolis.	101
Figura 33. Distribuição da variável <i>Local</i> para cada cluster.....	107
Figura 34. Distribuição da variável <i>Local_Antes</i> para cada cluster.....	108
Figura 35. Distribuição da variável <i>Local_Depois</i> para cada cluster.....	109
Figura 36. Distribuição da variável <i>Dist_Resid</i> para cada cluster.....	110
Figura 37. Distribuição da variável <i>Dist_POI_Depois</i> para cada cluster.....	111

Figura 38. Distribuição da variável <i>Uso_Solo</i> para cada cluster.....	112
Figura 39. Distribuição da variável <i>Dia_Útil_FDS</i> para cada cluster.....	113
Figura 40. Distribuição da variável <i>Alocação_Hora_Início</i> para cada cluster.	114
Figura 41. Distribuição da variável <i>Dur_Curta_Aloc</i> para cada cluster.	115
Figura 42. Distribuição da variável <i>Dur_Curta_Arred</i> para cada cluster.	116
Figura 43. Distribuição da variável <i>Dur_Longa_Arred</i> para cada cluster.	117
Figura 44. Distribuição da variável <i>Freq_POI_Vezes</i> para cada cluster.	118
Figura 45. Distribuição da variável <i>Freq_POI_Dias</i> para cada cluster.....	119
Figura 46. Distribuição da variável <i>Ordem_Atividade</i> para cada cluster.....	120
Figura 47. Porcentagem de indivíduos por cluster. Fonte: Elaboração própria.....	127
Figura 48. Comparação entre os clusters de indivíduos. Fonte: Elaboração própria.	130
Figura 49. Distribuição das variáveis socioeconômicas para cada grupo de indivíduos e amostras (Parte 1). Fonte: Elaboração própria.	140
Figura 50. Distribuição das variáveis socioeconômicas para cada grupo de indivíduos e amostras (Parte 2). Fonte: Elaboração própria.	141
Figura 51. Distribuição das variáveis socioeconômicas para cada grupo de indivíduos e amostras (Parte 3). Fonte: Elaboração própria.	142
Figura 52. Distribuição das variáveis socioeconômicas para cada grupo de indivíduos e amostras (Parte 4). Fonte: Elaboração própria.	143
Figura 53. Dendrograma para o método de encadeamento completo – Método Hierárquico.	161
Figura 54. Média das variáveis para cada cluster – Método Hierárquico.....	163
Figura 55. Distribuição da variável <i>NumAtividades</i> para cada cluster.	169
Figura 56. Distribuição da variável <i>NumClusters</i> para cada cluster.....	170
Figura 57. Distribuição da variável <i>MaxRepClusters</i> para cada cluster.....	171
Figura 58. Distribuição da variável <i>%Cluster 1 (Residência – Longa)</i> para cada cluster. ...	172
Figura 59. Distribuição da variável <i>%Cluster 2 (Residência – Curta-média)</i> para cada cluster.	173
Figura 60. Distribuição da variável <i>%Cluster 3 (Fora – Média-longa)</i> para cada cluster.....	174

Figura 61. Distribuição da variável %Cluster 4 (<i>Dentro – Média-longa</i>) para cada cluster.	175
Figura 62. Distribuição da variável %Cluster 5 (<i>Dentro – Curta</i>) para cada cluster.	176
Figura 63. Distribuição da variável %Cluster 6 (<i>Fora – Curta</i>) para cada cluster.....	177

LISTA DE TABELAS

Tabela 1. Descrição dos tipos de padrões diários de viagens e atividades	37
Tabela 2. Descrição dos tipos de padrões semanais de viagens e atividades	38
Tabela 3. Amostra de indivíduos calculada por erro amostral	50
Tabela 4. Principais barreiras para a coleta de dados e ações propostas (Pizzol et al., 2017)	54
Tabela 5. Indivíduos contatados na pesquisa	56
Tabela 6. Exemplo de registros de localização para um indivíduo, obtida a partir do <i>website</i> do aplicativo <i>FollowMee</i>	56
Tabela 7. Número de dias de coleta	57
Tabela 8. Análise de sensibilidade dos parâmetros para o algoritmo DBSCAN	67
Tabela 9. Exemplo de identificação de Atividades e POI's após o processamento do algoritmo DBSCAN.....	70
Tabela 10. Número de registros de GPS, Atividades e POI's para a amostra de 105 indivíduos	72
Tabela 11. Distribuição das primeiras e últimas atividades entre residências e demais POI's	72
Tabela 12. Número de Atividades e POI's finais para a amostra de 105 indivíduos.....	73
Tabela 13. Atributos das atividades	73
Tabela 14. Descrição de Tipologias de Uso do Solo.....	77
Tabela 15. Exemplo dos Campos de Identificação e Variáveis Espaciais para as atividades de um indivíduo	82
Tabela 16. Exemplo das Variáveis Temporais para as atividades de um indivíduo (continua)	83
Tabela 17. Exemplo das Variáveis Temporais (continuação) e Variáveis de Repetição e Sequência para as atividades de um indivíduo	84
Tabela 18. Variáveis usadas para o agrupamento das Atividades e respectivo nível de medição adotado	92
Tabela 19. Categorias adotadas para as Variáveis Espaciais Ordinais.....	93

Tabela 20. Categorias adotadas para a Variável de Repetição Ordinal	93
Tabela 21. Categorias adotadas para as Variáveis Temporais Ordinais	94
Tabela 22. Categorias adotadas para a Variável Espacial Nominal <i>Uso_Solo</i>	94
Tabela 23. Variáveis e resultados dos testes conduzidos no processo de agrupamento das atividades	97
Tabela 24. Variáveis e resultados dos testes conduzidos na etapa final de agrupamento das atividades	98
Tabela 25. Alocação das atividades por cluster	99
Tabela 26. Identificação dos grupos de atividades.....	105
Tabela 27. Variáveis usadas para a clusterização dos Indivíduos.....	121
Tabela 28. Variáveis e resultados dos testes conduzidos no processo de clusterização dos indivíduos – Método <i>TwoStep</i>	124
Tabela 29. Análise de variância (ANOVA) dos clusters de indivíduos – Método <i>TwoStep</i> .	125
Tabela 30. Identificação dos padrões de atividades dos clusters de indivíduos	126
Tabela 31. Alocação dos indivíduos por cluster	126
Tabela 32. Medianas e quartis das variáveis para cada cluster de indivíduo	128
Tabela 33. Comparação entre as características socioeconômicas da população de Paraisópolis e das amostras inicial e final.....	133
Tabela 34. Variáveis socioeconômicas para cada grupo de indivíduos e amostras - Pessoal e família (Parte 1)	135
Tabela 35. Variáveis socioeconômicas para cada grupo de indivíduos e amostras - Pessoal e família (Parte 2)	136
Tabela 36. Variáveis socioeconômicas para cada grupo de indivíduos e amostras - Estudo e trabalho.....	137
Tabela 37. Variáveis socioeconômicas para cada grupo de indivíduos e amostras - Estudo e trabalho e Transporte.....	138
Tabela 38. Variáveis socioeconômicas para cada grupo de indivíduos e amostras - Transporte e exercício físico	139
Tabela 39. Identificação dos padrões de atividades e caracterização socioeconômica dos clusters de indivíduos	148

Tabela 40. Estatística descritiva dos clusters de indivíduos – Método Hierárquico	162
Tabela 41. Análise de variância (ANOVA) dos clusters de indivíduos – Método Hierárquico	164
Tabela 42. Estatística descritiva dos clusters de indivíduos – Método K-means	166
Tabela 43. Análise de variância (ANOVA) dos clusters de indivíduos – Método K-means .	167

SUMÁRIO

1. Introdução.....	18
1.1. Objetivo	19
1.2. Etapas do Método.....	20
1.3. Estrutura do Documento	21
2. Revisão Bibliográfica	22
2.1. Favelas e População de Baixa Renda	22
2.2. Métodos de Coleta e Processamento de Dados de Viagens e Atividades	25
2.2.1. Novos Métodos de Coleta de Dados	25
2.2.2. Uso de Receptores de GPS para Coleta de Dados	26
2.2.3. Identificação de Atividades e Viagens	30
2.3. Estudos com Dados de Múltiplos Dias	31
2.3.1. Métricas Usadas para Identificação de Padrões de Atividades e Viagens em Múltiplos Dias	32
1. Variância da Taxa de Geração de Viagens	33
2. Variabilidade do Espaço de Atividades	34
3. Herfindahl-Hirschman Index (HHI)	35
4. Caracterização de Padrões Semanais a Partir de Padrões Diários.....	36
2.3.2. Resultados Obtidos para Diferentes Grupos de Indivíduos	38
3. Coleta de Dados	42
3.1. Área de Estudo.....	42
3.2. Desenho da Pesquisa.....	47
3.2.1. Planejamento Amostral.....	49
3.2.2. Desafios da Pesquisa	53
3.2.3. Taxas de Resposta.....	55
3.3. Dados de GPS dos <i>Smartphones</i>	56
4. Processamento dos Dados	59

4.1.	Definições.....	59
4.2.	Limpeza dos Dados e Seleção dos Indivíduos para Análise	60
4.3.	Método de Identificação de Atividades e <i>Points of Interest</i> (POI's)	64
4.3.1.	Density-Based Clustering Algorithm (DBSCAN)	64
4.3.2.	Análise de Sensibilidade e Inspeção Visual.....	66
4.3.3.	Identificação de Atividades e POI's.....	69
4.4.	Caracterização das Atividades.....	73
5.	Padrões de Atividades em Múltiplos Dias	85
5.1.	Métodos de Agrupamento.....	85
5.1.1.	Método Hierárquico	87
5.1.2.	<i>K-means</i>	89
5.1.3.	<i>TwoStep Clustering</i>	91
5.2.	Classificação de Atividades	91
5.2.1.	Variáveis Consideradas.....	92
5.2.2.	Processo de Agrupamento das Atividades	94
5.2.3.	Análise dos Grupos de Atividades	99
5.3.	Classificação de Indivíduos	121
5.3.1.	Variáveis Consideradas.....	121
5.3.2.	Processo de Classificação.....	122
5.3.3.	Análise dos Grupos de Indivíduos	126
5.4.	Caracterização Socioeconômica dos Grupos de Indivíduos	133
6.	Conclusões.....	151
6.1.	Recomendações futuras.....	152
	Referências Bibliográficas	153
	Apêndices.....	159
1.	Processo de Classificação dos Indivíduos.....	159
2.	Distribuição de Variáveis por Grupo de Indivíduo.....	168

Anexos	178
1. Formulário da entrevista domiciliar	178

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.

Motivados pela necessidade de prever a demanda futura por viagens para melhor direcionar os investimentos em projetos de infraestrutura e serviços de transporte, os pesquisadores têm procurado, há tempos, desenvolver modelos para estimar como as pessoas se locomovem no tempo e espaço e procuram entender os fatores que afetam as escolhas relacionadas às suas viagens.

A maioria dos modelos e análises de comportamento relativo a viagens objetiva explicar as variações no comportamento dos indivíduos em relação às viagens em termos de suas características e o ambiente que os cerca, o que constitui a chamada *variabilidade interpessoal*. Para tanto, são tipicamente baseados nos registros de viagem dos indivíduos ou domicílios para um único dia, que normalmente é o tipo de dado disponível nas pesquisas realizadas.

Dados de um dia, no entanto, revelam apenas aspectos limitados sobre o comportamento de viagens (Hanson e Huff, 1982; Pas e Koppelman, 1987; Pas, 1988), visto que as pessoas não repetem o mesmo padrão todos os dias. Apesar dessa variação no comportamento individual de viagens dia a dia, também 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 ainda não era tão comum entre as pesquisas de transporte.

A posse de dados de múltiplos dias, que permitem avaliar a existência da variabilidade intrapessoal, aumenta, contudo, 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.

Em locais como favelas e assentamentos informais, a complexidade do ambiente urbano em que estão inseridos pode apresentar considerável influência nas atividades realizadas pela população de baixa renda que ali vive. Conforme apontam Ewing e Cervero (2010), diversos

estudos fornecem evidências econômicas e comportamentais da influência do ambiente construído no padrão de viagens em centros urbanos.

Além disso, as condições geográficas, sociais e econômicas da população de baixa renda residente em favelas, podem refletir desigualdades de acesso a oportunidades no ambiente urbano. A não participação de conjuntos de indivíduos em atividades-chave da sociedade em que vive, como o consumo de bens e serviços, a produção econômica ou social, a participação em processos de tomada de decisão e interações sociais, está relacionada à exclusão social destes grupos.

No entanto, por envolver questões políticas, sociais e ambientais, a exclusão social não é consequência apenas da estratificação econômica, podendo ocorrer também entre grupos que não são economicamente distintos uns dos outros de forma tão significativa (Barry, 2002). Por exemplo, entre homens e mulheres (Kwan, 1999), usuários de automóveis e pedestres ou usuários de transporte público (Farber e Páez, 2011), pessoas com deficiência e pessoas saudáveis (Casas, 2007), crianças e adolescentes (Casas et al., 2009).

Como demonstra a revisão de literatura realizada por Koch (2013), os estudos de transporte em favelas tendem a ser localizados na África e Ásia, com pouco foco na América do Sul ou no Brasil. Puderam ser identificados estudos com população de baixa renda que tratam da exclusão social e sua relação com o transporte, mas não com população residente em favelas. Mais recentemente, Koch (2013) analisou o padrão de viagens em três favelas do Rio de Janeiro, a partir de uma coleta de dados com 2000 residentes, a fim de guiar políticas futuras de investimentos em transportes nas favelas, para o desenvolvimento das comunidades.

1.1. Objetivo

Considerando o contexto de transporte em favelas, bem como de comportamento de diferentes grupos sob uma mesma condição econômica, esse estudo visa investigar 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. São definidos a seguir os objetivos específicos do trabalho:

1. Avaliar o processo de coleta e processamento de dados sobre comportamentos individuais por múltiplos dias com base no uso de dados de GPS coletados com *smartphones*.
2. Analisar os padrões de atividades de diferentes indivíduos de uma mesma classe social – população de baixa renda residente em uma favela.

3. Identificar grupos de pessoas que apresentam padrões semelhantes entre si e distintos dos demais.
4. Verificar as características sociodemográficas e econômicas dos indivíduos classificados em um mesmo grupo.
5. Compreender se os indivíduos que apresentam padrões de atividades semelhantes, quando considerados os múltiplos dias de análise, assemelham-se também considerando sua ocupação, renda, posse de automóvel, gênero, idade e outras características – o perfil que apresentam tem associação com o seu comportamento ou não há relação explícita nesse sentido?

1.2. Etapas do Método

Para alcançar os objetivos propostos, serão utilizados dados de uma semana de atividades realizadas por habitantes de Paraisópolis, uma favela na cidade de São Paulo. Os dados foram obtidos a partir de um aplicativo instalado nos *smartphones* de 105 residentes. Além disso, foi realizada uma pesquisa domiciliar por meio de um questionário impresso, na qual os participantes respondiam perguntas sobre suas condições socioeconômicas e também informações gerais sobre seus hábitos de transporte.

O método empregado neste estudo 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 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 socioeconômicas das pessoas que os formavam. O esquema da Figura 1 ilustra este processo.

Figura 1. Etapas do método. Fonte: Elaboração própria.

1.3. Estrutura do Documento

Após esta introdução, o documento está organizado em mais cinco capítulos, sendo o conteúdo de cada um deles resumido a seguir.

Primeiramente, é definido o contexto mais amplo do estudo com base em uma revisão da literatura. Em seguida, é apresentada a área de estudo e detalhada a coleta de dados realizada. O quarto capítulo descreve a abordagem metodológica adotada para processamento dos dados. O quinto capítulo relata o método usado para identificação de padrões de atividades em múltiplos dias e os principais resultados. Finalmente, o sexto capítulo apresenta as conclusões finais do estudo.

2. Revisão Bibliográfica

A revisão bibliográfica visa o embasamento do leitor sobre os principais temas e as lacunas observadas nessa linha de pesquisa. Foram visitados pela autora e aqui apresentados os seguintes assuntos: favelas, população de baixa renda e exclusão social; métodos de coleta e processamento de dados de viagens e atividades; e estudos com dados de múltiplos dias, incluindo métricas usadas para identificação de padrões de atividades e viagens e resultados verificados para diferentes grupos de indivíduos.

Inicialmente, por se tratar de uma pesquisa realizada em Paraisópolis, o leitor será contextualizado sobre a relação entre as favelas, população de baixa renda e o transporte, com foco na dificuldade de acesso a oportunidades a qual seus moradores por vezes estão submetidos no ambiente urbano.

2.1. Favelas e População de Baixa Renda

De acordo com o Censo 2010¹, 84% da população brasileira vive em áreas urbanas. O rápido processo de urbanização do país foi caracterizado por uma distribuição desigual dos bens, com pessoas de baixa renda vivendo em terrenos urbanos inapropriados, em áreas de risco e/ou localizadas na periferia das cidades com acesso limitado a oportunidades e serviços, incluindo escolas, centros de saúde e transporte público. A maioria dos domicílios em assentamentos de baixa renda está concentrada em 20 Regiões Metropolitanas, especialmente em São Paulo, no Rio de Janeiro, em Belém, em Salvador e no Recife (Maia et al., 2016).

Estes assentamentos são caracterizados normalmente pela ocupação ilegal e informal de alta densidade em terrenos públicos ou privados, com infraestruturas básicas inadequadas. São áreas que costumam apresentar uma grande mistura de usos do solo, com pequenos comércios locais. Sua localização geográfica em relação aos centros comerciais da cidade pode variar.

Koch et al. (2013) afirmam que ainda são necessários estudos detalhados para compreender as restrições específicas de viagens e as necessidades de acessibilidade das comunidades de baixa renda, desconectadas há tempos dos processos políticos, econômicos e sociais

¹ IBGE (Instituto Brasileiro de Geografia e Estatística). Disponível em: <https://censo2010.ibge.gov.br/>.

formais nas cidades brasileiras. Por sua vez, Ciccantell e Bunker (1998) e Martell (2016) apontam para a importância de se ter em mente as características dos sistemas de subsistência nos quais muitos residentes urbanos pobres – jovens e idosos, homens e mulheres – podem operar em cidades em desenvolvimento.

Lucas e Porter (2016) atentam ainda para o fato de muitas das soluções de mobilidade adotadas pelas cidades em desenvolvimento serem “copiadas” de países desenvolvidos, não levando em consideração o contexto local. As autoras esclarecem que cada cidade em desenvolvimento possui seu próprio sistema socioespacial complexo e requer uma compreensão contextualizada dos impactos socioeconômicos que diferentes soluções de mobilidade terão em diferentes setores da população, a fim de assegurar que seus habitantes sejam adequadamente providos de acesso sustentável a bens, serviços e atividades, hoje e no futuro.

Em contraste com o Norte Global, onde os sistemas de transporte já estão melhor estabelecidos, as cidades em desenvolvimento têm potencial considerável para reformular a trajetória futura da mobilidade das pessoas, e o potencial de subsistência associado, através da introdução de novos sistemas de transporte melhor alocados e regulamentados, que considerem os interesses da população local, incluindo os mais pobres.

Maia et al. (2016) pesquisaram diferentes grupos de moradores de baixa renda em duas comunidades contrastantes na região metropolitana de Recife, no Brasil: uma na favela do Coque, localizada no centro da cidade, e outra na do Alto Santa Teresina, na periferia. Os autores identificaram que os moradores usam a caminhada como parte ou integralmente para todas as viagens e muitos usam transporte informal. Uma questão-chave examinada no estudo é até que ponto os padrões restritos de mobilidade e atividades dos cidadãos influenciam sua qualidade de vida.

No estudo de caso com moradores da comunidade de Soacha, na periferia urbana de Bogotá na Colômbia, Hernandez e Titheridge (2016) relatam que muitas das pessoas que vivem em áreas com baixo acesso a serviços de transporte público desenvolvem estratégias alternativas de mobilidade para manter o acesso a oportunidades na cidade e em outras áreas periféricas. Esses serviços de transporte informais tendem a oferecer uma possibilidade flexível, geralmente barata, para os residentes de baixa renda chegarem a seus destinos de emprego, educação e compras quando os serviços públicos de transporte não estão disponíveis em seus locais de origem.

Koch et al. (2013) realizaram estudo mais aprofundado, que descreveu os padrões de viagens e atividades das pessoas em três favelas da cidade do Rio de Janeiro. A partir de uma

pesquisa com 2000 residentes, foram analisados o índice de mobilidade, a posse de automóvel, o uso de transporte ativo, motivos e tempos de viagem e as percepções de segurança viária. Os autores concluíram que “os investimentos em transporte são um componente central dos programas de urbanização, mas a base de conhecimento sobre o uso do transporte nas favelas do Rio é totalmente inadequada e os impactos de esforços anteriores para melhorar o transporte permanecem pouco compreendidos ou analisados”.

Maia et al. (2016) apontam também a necessidade de mudança do foco de políticas exclusivamente voltadas para a mobilidade para políticas que considerem também a acessibilidade e perspectivas de bem-estar social. Em países como Reino Unido, Holanda e Austrália estão sendo adotadas políticas em reconhecimento de que, embora a mobilidade seja um importante contribuinte para a inclusão social da população de baixa renda, o fator chave para isso é o acesso às atividades em vez do acesso ao transporte em si.

No Brasil, no entanto, há poucos estudos que reconheçam a interface entre transporte e acessibilidade às atividades, e a desvantagem social relacionada a isso. A existência de transporte comunitário, que atenda às comunidades de baixa renda de muitas cidades brasileiras, pode comprovar a inadequação do transporte coletivo regular às necessidades dessa parcela da população. Como exemplos podem ser citados moto-táxis, vans e micro-ônibus, altamente valorizados pelas comunidades urbanas mais pobres, mas que, em geral, permanecem desregulamentados pelo setor público, sendo considerados transporte informal ou ilegal (Maia e Orrico Filho, 2007).

Ao passo que se observa considerável número de estudos sobre programas de urbanização de favelas em comunidades de baixa renda nas cidades brasileiras, há poucos estudos sobre suas necessidades de transporte. A provisão de transporte em programas de melhoria de favelas geralmente se limita à pavimentação das vias e melhorias nas escadarias, não considerando o acesso dos indivíduos a oportunidades socioeconômicas externas às comunidades (Maia et al., 2016). Considera-se, então, que pesquisas que investiguem as atividades que populações urbanas de baixa renda realizam podem vir a ampliar o entendimento de suas necessidades.

Contudo, coletar dados de atividades e viagens de população de baixa renda, que vive em favelas, é um desafio. Diversas categorias na população têm sido identificadas como “grupos de difícil acesso” quanto à coleta de dados de transporte, dependendo de seus atributos socioeconômicos, culturais, geográficos e físicos. Grupos de baixa renda, bem como comunidades segregadas, são frequentemente assim denominadas (Behren et al., 2009).

Residentes de favelas, em específico, podem ser consideradas “grupos de difícil acesso” por alguns motivos, segundo Ampt e Hickman (2015): dificuldade de enumerar seus habitantes, já que os dados do Censo podem ser incompletos ou faltantes para essa parcela da população; dificuldade de escolher uma amostra, dada a lacuna de endereços formais e números de telefone; dificuldade de atingir a população alvo da pesquisa, dados os horários de trabalho normalmente atípicos; necessidade de respeitar o crime local organizado, quando existente. No Brasil, por exemplo, o Censo demográfico começou a registrar sistematicamente informações sobre favelas apenas em 2010.

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. Por esses motivos, certos incentivos para redução de taxas de não resposta podem ser necessários (Behrens et al., 2009).

2.2. Métodos de Coleta e Processamento de Dados de Viagens e Atividades

Pesquisas que investiguem as atividades de indivíduos de baixa renda trazem consigo uma questão operacional ainda não bem resolvida no campo de estudos de transporte, que é a coleta de informações para esse extrato da população. Além disso, o fato de moradores das favelas serem consideradas “grupos de difícil acesso”, também dificulta a obtenção de dados de viagens e atividades diárias de seus moradores.

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. Além disso, o recente desenvolvimento da tecnologia e seu barateamento tem possibilitado a inclusão tecnológica dos extratos mais pobres da população. O avanço nessa área pode ser um aliado na obtenção de dados de comportamento dos indivíduos, até mesmo para grupos historicamente considerados como de difícil acesso.

2.2.1. Novos Métodos de Coleta de Dados

Os métodos tradicionais usados na coleta de dados de viagens para previsões de demanda, como as pesquisas domiciliares de origem e destino, possuem incontestável valor. Além de obter dados sobre os padrões de mobilidade individuais e do domicílio, também incluem informações relevantes sobre os modos e motivos das viagens realizadas. No entanto, segundo Anda et al. (2017), duas principais limitações podem ser identificadas: em primeiro lugar, representam apenas uma pequena amostra da população (normalmente em torno de 1%); e em segundo lugar, geralmente são atualizados apenas a cada cinco ou dez anos.

Além disso, no caso de pesquisas com questionários, um problema comum é que geralmente elas retratam o comportamento regular das viagens, sendo que muitas vezes o número de perguntas é limitado devido aos custos e ao tempo necessário para preencher o questionário. Com isso, a identificação de comportamentos individuais detalhados fazendo uso de questionários tradicionais é bastante difícil, não sendo suficiente para estudos aprofundados (Shen e Stopher, 2014).

Novos métodos mais eficientes de coleta de dados têm atraído o interesse de pesquisadores nos últimos anos. O uso de dados de dispositivos GPS, telefones celulares e bilhetagem eletrônica estão se tornando cada vez mais populares. Sua principal vantagem é a coleta automática de uma grande quantidade de dados de trajetórias espaço-temporais sem a necessidade de pesquisas caras e demoradas (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.

Maia et al. (2016), por exemplo, constataram que na comunidade do Coque, em Recife, localizada a 2 km da área central da cidade, apesar dos habitantes descreverem a área como bem servida por transporte público, eles relataram que não o usam de fato, porque podem realizar a maioria de suas atividades diárias a pé, evitando o custo associado.

Ainda que haja desafios relacionados ao uso desses tipos de tecnologia, acredita-se na necessidade de explorar novas alternativas de ferramentas para coleta de dados, especialmente em áreas pobres, que são historicamente mal representadas em modelos de planejamento devido à falta ou baixa qualidade de dados.

2.2.2. Uso de Receptores de GPS para Coleta de Dados

Tradicionalmente, as pesquisas sobre padrão de viagens costumam ser realizadas por meio de entrevistas domiciliares, nas quais os entrevistados são solicitados a reportar suas atividades e viagens via papel, internet ou entrevistas por telefone. Além de caros, esses métodos consomem tempo dos participantes e são desgastantes para todas as partes (Stopher e Metcalf, 1996).

Recentemente, outras formas de coleta de dados têm sido desenvolvidas, como as pesquisas com uso de *smartphones* com receptor de GPS embutido, que registram as trajetórias diárias dos indivíduos rastreados. Observa-se também uma tendência de combinar a coleta de dados

com *smartphones* às pesquisas tradicionais, nas quais os participantes respondem a um questionário. Entre as vantagens desse método estão a diminuição do número de viagens não reportadas (Bricka e Bhat, 2006) e maior precisão do tempo e local dos deslocamentos capturados pelos dispositivos GPS (Schönfelder *et al.*, 2002). Também há métodos que se baseiam nos registros do GPS para realizar, em seguida, uma pesquisa de confirmação com os participantes.

Pesquisas que fazem uso do GPS para coletar dados de atividades e viagens necessitam da solicitação ativa aos participantes, pois requerem a instalação de aplicativos que captem os sinais recebidos pelo GPS, no caso de uso dos *smartphones* para registro dos dados. Devido a isso, essas pesquisas são limitadas a uma amostra de indivíduos, o que faz com que os dados de tal natureza sejam comumente caracterizados como “*small data*”, em contraposição aos dados classificados como “*big data*”, como os registros de chamadas de celulares (*Call data records* - CDR) coletados por empresas de telefonia e dados de bilhetagem eletrônica dos sistemas de ônibus urbano.

Há também algumas limitações técnicas do uso de receptores de GPS para a coleta de dados, sendo que o principal problema 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 – efeito urbano de *canyon* (Stopher *et al.*, 2008).

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 segmentos da população. Até mesmo em uma comunidade de baixa renda como Paraisópolis, *smartphones* são comuns e a internet 3G é amplamente disponível.

De acordo com uma pesquisa da Kantar Worldpanel (2016), a penetração dos *smartphones* está cada vez maior no Brasil: 57% da população brasileira é proprietária de ao menos um aparelho deste tipo. Os celulares, sejam os *smartphones* ou aqueles que têm funções mais básicas, estão inseridos em praticamente toda a população brasileira: 94% disseram ter um telefone móvel. Mesmo entre as pessoas das classes mais pobres D e E, 91% possuem pelo menos um aparelho, segundo a pesquisa realizada.

Os receptores GPS podem ser encontrados em diversos formatos. Os dois tipos mais comuns de aparelhos usados para coleta de dados de GPS em transportes são os dispositivos dedicados somente ao GPS e dispositivos integrados a *smartphones*.

Dispositivos dedicados ao GPS

Os dispositivos dedicados são aparelhos usados exclusivamente para registro de dados de GPS. São mais antigos que os dispositivos integrados a telefones celulares, sendo que após o surgimento destes, seu uso foi bastante reduzido, devido principalmente à maior praticidade de se carregar os telefones celulares, que são menores e mais leves que os dispositivos dedicados.

Apesar de hoje o preço de um aparelho desses ser muito menor do que no final do século 20, ainda assim pode ser uma despesa significativa, especialmente em pesquisas com muitos entrevistados. Além disso, tal dispositivo portátil requer esforço e disciplina dos entrevistados, que devem levá-lo consigo e recarregar ou trocar baterias, algo que é feito naturalmente com os *smartphones*, visto que receptores de GPS consomem uma quantidade considerável de energia (Vlassenroot et al., 2015).

Os primeiros estudos com uso de dispositivos dedicados ao GPS surgiram em meados da década de 1990 e incluíram principalmente métodos de localização de veículos e análise do comportamento relacionado às viagens de indivíduos (Gadzinski, 2018). Entre suas principais vantagens estão a alta precisão dos dados de localização e a alta frequência de gravação dos dados, além da possibilidade de serem usados para avaliar a velocidade e a direção de um indivíduo ou veículo equipado (Duncan et al., 2009).

Dispositivos GPS integrados a smartphones

Nos últimos anos, os *smartphones* começaram a ser vistos como ferramentas promissoras para a obtenção de dados para análises de comportamento de viagens. Os *smartphones* normalmente podem coletar dados de diferentes fontes e sensores: antenas de celular, sinal de GPS, sinal Wi-Fi, acelerômetro, Bluetooth, tornando-se possível, portanto, obter grandes quantidades de dados precisos de localização sem o uso de dispositivos dedicados (Gadzinski, 2018). Além disso, podem ser instalados aplicativos para simplificar o processo de coleta de dados ou para ampliar o alcance da pesquisa, por exemplo, por meio de um questionário específico ou confirmação ativa dos dados coletados.

Todos os *smartphones* são geralmente equipados com receptores de GPS e também alguns telefones celulares convencionais podem inclui-lo. Para calcular a posição, a partir dos dados do sistema GPS, o receptor usa informações de pelo menos quatro satélites em um método chamado “trilateração”. Em áreas abertas, a precisão pode alcançar até 5 metros, mas sofre degradação em áreas com predomínio de edifícios altos ou árvores e dentro de edifícios, onde os sinais de GPS são prejudicados. O GPS assistido (A-GPS) pode aumentar a precisão da localização em áreas urbanas, combinando sinais de localização de GPS com dados de localização do celular, alcançando precisão inferior a 10 metros. Similarmente, outras formas

de sistemas de localização hibridizados podem incluir também o uso de sinais de rede Wi-Fi (Anda et al., 2017).

Há também algumas desvantagens tecnológicas do uso de *smartphones* com GPS: a precisão da localização pode ser um pouco pior do que no caso de dispositivos dedicados, e o seu uso pode ser limitado devido à bateria relativamente curta e ao alto custo de transferência de dados para *data centers*. No entanto, parece que essas limitações estão se tornando cada vez menos importantes com o intenso desenvolvimento tecnológico observado (Shen e Stopher, 2014).

Gadzinski (2018) desenvolveu um aplicativo de baixo custo e realizou um estudo piloto com mais de 100 estudantes em Poznań, na Polônia. O autor concluiu que há um enorme e crescente potencial relacionado aos telefones celulares, mas ainda existem algumas barreiras importantes, como problemas de amostragem, questões tecnológicas e limitações analíticas.

Para o autor, o uso de dados obtidos com *smartphones* pode ampliar as possibilidades analíticas nos estudos de transporte, sendo sua grande vantagem o fato de que eles podem resolver muitos dos problemas enfrentados pelas pesquisas tradicionais, como a subdeclaração de viagens curtas, relatos de locais e horários imprecisos e informações sobre um dia “típico” em vez do dia real, além da grande flexibilidade na concepção do questionário.

No entanto, o autor faz observações a respeito desse tipo de pesquisa: possíveis problemas de amostragem, devido à distribuição desigual de *smartphones* na sociedade; necessidade de limpeza dos dados brutos antes do processamento devido a interrupções; necessidade de algoritmos avançados e sofisticados para processamento dos dados. Por fim, afirma que um uso mais amplo de *smartphones* em estudos de comportamento em viagens parece ser uma perspectiva ainda distante.

2.2.3. Identificação de Atividades e Viagens

Anda et al. (2017) apontam uma desvantagem a ser considerada ao utilizar esses novos tipos de dados: a sua natureza bruta, sendo necessário um esforço analítico adicional para identificação de atividades, viagens, modos e motivos. 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.

Anda et al. (2017) citam autores que utilizaram técnicas de clusterização de tempo e distância para inferir atividades dos indivíduos a partir de dados de CDR (*Call Data Records*, obtidos a partir das antenas de telefonia celular). Primeiramente, eles agruparam pontos que eram espacialmente próximos, medindo a distância entre dois pontos consecutivos e comparando-a a um limiar de distância (por exemplo, 300 m). Em seguida, os clusters obtidos foram considerados atividades potenciais se o intervalo de tempo entre a primeira e a última observação no cluster fosse maior que um limiar de tempo (por exemplo, 10 min). A localização geográfica da atividade foi configurada para ser o centroide de todos os pontos do cluster.

Uma das formas diretas de inferir informações contextuais, como o tipo da atividade ou o motivo da viagem, é por regras de frequência e horário. Segundo Anda et al. (2017), há estudos que usaram a frequência e horário da atividade para identificar trabalho, casa e outros locais, a partir de dados de CDR. Eles definiram o local de residência do usuário como o ponto de permanência mais frequentemente observado entre 20h e 7h durante dias da semana e fins de semana. O local de trabalho foi definido como o ponto de permanência, diferente de casa, que os usuários mais visitaram entre 7 e 20 horas nos dias úteis. Outra variação usada é a definição do local de trabalho como aquele para o qual o usuário percorre a máxima distância da residência, sendo todos os demais locais de atividade designados como “outros”.

Widhalm et al. (2015) usaram dados de CDR para agrupar as atividades em categorias significativas com base no horário de início, duração, sequência de locais de permanência e informação de uso do solo. Ortega-Tong (2013) usou dados de bilhetagem eletrônica para classificar as viagens de usuários de transporte público em Londres; Jiang et al. (2012) clusterizaram padrões diários de atividades usando dados de uma pesquisa domiciliar realizada em Chicago.

Sila-Nowicka *et al.* (2016) propuseram uma metodologia para identificar e analisar o comportamento dinâmico (viagens) e estático (atividades) das trajetórias obtidas a partir de dados de GPS. A partir de um conjunto de dados de 205 voluntários de três cidades escocesas, continuamente monitorados em seus deslocamentos diários por uma semana,

foram inferidos os modos de viagem e identificados e classificados os locais de atividades realizadas.

Anda et al. (2017) destacam a importância de se incluir variáveis como os locais visitados anteriormente, o horário de início e duração das atividades no processo de caracterização de atividades, a partir do comportamento diário dos indivíduos. Os autores evidenciam que embora essa abordagem seja limitada para dados de CDR e bilhetagem, tem sido adotada com uso de dados de GPS.

2.3. Estudos com Dados de Múltiplos Dias

O uso de receptores de GPS facilita a coleta de dados de múltiplos dias, permitindo estudar com profundidade a variabilidade intrapessoal dos indivíduos em relação aos seus padrões diários de viagens e atividades. Apesar da pouca disponibilidade de dados de múltiplos dias, a literatura registra alguns estudos que coletaram “*small data*” por períodos mais longos, variando de 7 dias até um mês ou mais. A seguir, são apresentados alguns exemplos de estudos que usaram esse tipo de conjunto de dados.

Hanson e Huff (1982, 1986) e Huff e Hanson (1986) usaram dados coletados em Uppsala, Suécia, em 1971, por 35 dias consecutivos (*Uppsala Household Travel Survey*). A partir de diários detalhados das viagens e atividades realizadas por uma amostra de indivíduos, foram conduzidas análises para investigar a variabilidade intrapessoal existente no período de observação.

Pas (1988) selecionou dados coletados pela pesquisa *Reading Activity Diary Survey* (Shapcott, 1978), realizada na Inglaterra entre Janeiro e Março de 1973, compondo uma amostra de 112 pessoas que trabalhavam. Os dados foram coletados com uma entrevista pessoal e um diário de viagens. O diário requeria ao respondente reportar cada atividade (tanto dentro como fora de casa), em um período de cinco dias úteis, tendo sido analisadas as viagens realizadas.

Outro conjunto de dados de múltiplos dias foi coletado na Holanda. Conhecido como *Dutch National Mobility Panel*, compreendeu uma amostra nacional de domicílios que foi pesquisada em intervalos de seis meses, com início em abril de 1984. Em cada uma das fases da pesquisa os participantes completavam um diário de viagens de sete dias e reportavam alterações em suas condições sociodemográficas em relação à interação anterior. Pesquisas que usaram esse conjunto de dados incluem Golob e Meurs (1987), Kitamura (1986, 1987, 1998), Kitamura e Bovy (1987) e Kitamura e van der Hoorn (1987).

Mas, talvez o conjunto de dados de múltiplos dias mais relevante é o *Mobidrive*. Axhausen *et al.* (2002) coletaram dados de seis semanas das cidades alemãs de Karlsruhe e Halle. Uma pesquisa similar, aperfeiçoada com a experiência adquirida na Alemanha, foi conduzida em 2003 em Thurgau, na Suíça. Esses conjuntos de dados foram amplamente usados em diversos estudos posteriores.

2.3.1. Métricas Usadas para Identificação de Padrões de Atividades e Viagens em Múltiplos Dias

Dado que as necessidades, desejos e restrições individuais não são constantes de um dia para o outro, o padrão de atividades e viagens de uma pessoa não é necessariamente repetido a cada dia. Há algumas atividades, como comer e dormir, que se repetem todo dia, mas outras atividades como compras, negócios pessoais e atividades sociais, não são realizadas necessariamente da mesma forma todos os dias.

Se de um lado há ritmos e rotinas que caracterizam certa repetição de atividades e viagens diariamente, de outro há uma diversidade de compromissos, desejos e oportunidades, que caracterizam o padrão de viagens como um processo dinâmico que envolve variabilidade das viagens diárias. Hanson e Huff (1981) concluíram que os padrões de atividades e viagens dos indivíduos são caracterizados tanto por repetição quanto por variabilidade.

Por sua vez, Pas e Koppelman (1987) exploraram os fatores que afetam a variabilidade intrapessoal em viagens urbanas. Para eles, a variação no comportamento diário de atividades e viagens dos indivíduos deve ser resultado da variação de suas necessidades e desejos a serem satisfeitos, além de ser afetada pelos recursos disponíveis para viajar e restrições de tempo e espaço, que limitam a liberdade individual de variar seu comportamento dia a dia. Esse processo de caracterização dos padrões de atividades e viagens é resumido na Figura 2.

Figura 2. Processo de caracterização dos padrões de atividades e viagens dos indivíduos. Fonte: Elaboração própria.

Estudos mostram que não há um único dia da semana que possa ser identificado como representativo para a maioria das pessoas, em relação ao seu comportamento de viagens (Hanson e Huff, 1988; Huff e Hanson, 1990; Yamamoto e Kitamura, 1999; Kang e Scott, 2010).

No entanto, a realização de atividades principais como estudo e trabalho implica em algum grau de regularidade temporal (Huff e Hanson, 1990). O nível de repetição varia para diferentes grupos sociodemográficos e de comportamento de viagens (Pas, 1987; Pas e Koppelman, 1986; Pas e Sundar, 1995; Dharmowijoyo *et al.*, 2015).

Compreender a estabilidade entre diferentes combinações de atividades e viagens pode auxiliar a criação de melhores métodos de estimativas para as decisões diárias dos indivíduos. Além disso, pode colaborar com a análise da influência dos atributos das viagens, das características socioeconômicas e do ambiente construído no desenvolvimento de padrões mais ou menos habituais ou repetitivos (Susilo e Axhausen, 2014).

Mesmo com disponibilidade de dados para diversos dias, a análise da repetição e variabilidade dos padrões de atividades e viagens ainda é um desafio. O principal motivo para isso é esses padrões serem multifacetados, ou seja, há diversas formas de caracterizar um padrão de atividade e viagem. Por exemplo, pode-se focar no número de viagens ou no encadeamento de viagens para representar o padrão de um indivíduo (Susilo e Kitamura, 2005). Pode-se também utilizar abordagens mais elaboradas, que considerem diversas métricas das viagens realizadas por um indivíduo para representar seu padrão de viagens.

Anda et al. (2017) consideram longos períodos de observação como uma oportunidade para aplicar modernas técnicas de mineração dos dados a fim de inferir informações adicionais sobre o comportamento dos indivíduos.

A seguir são apresentadas, em maior detalhe, algumas métricas encontradas na literatura para identificação de padrões de atividades e viagens em múltiplos dias. Foram organizadas em ordem crescente de complexidade, em relação ao número e tipo de atributos considerados e métodos de análise utilizados.

1. Variância da Taxa de Geração de Viagens

Pas e Koppelman (1987) usaram dados coletados pela *Reading Activity Diary Survey* na Inglaterra em 1973, que continha o registro de viagens de 112 pessoas por um período de cinco dias úteis. Para cada pessoa foi calculado o número de viagens diárias para cada um dos cinco dias. A variabilidade intrapessoal foi medida em termos das taxas de geração de viagens diárias, definida pela seguinte expressão:

$$V_m = \frac{1}{N} \cdot \sum_{i=1}^N \left[\frac{1}{N_D - 1} \cdot \sum_{j=1}^{N_D} (t_{ij} - t_{mi})^2 \right] \quad (1)$$

Sendo:

V_m = Variabilidade média intrapessoal de taxas de viagem para o segmento de população em análise;

t_{ij} = Número de viagens realizadas pelo indivíduo i no dia j ;

t_{mi} = Número médio diário de viagens realizadas pelo indivíduo i no período;

N_D = Número de dias no período de análise;

N = Número de indivíduos no segmento de população em análise.

Pode-se observar que essa medida de variabilidade equivale à média da variância do número de viagens: um alto valor resultante para a variabilidade média intrapessoal indica um segmento de população cujos membros apresentam maior variação em suas taxas de geração de viagens de um dia para o outro.

2. Variabilidade do Espaço de Atividades

Fazendo uso dos dados do Mobidrive, Susilo e Kitamura (2005) examinaram as características do espaço de atividades dos indivíduos, representando-o pelo “segundo momento de locais de atividades”, composto pela combinação de duas parcelas, I_C e I_H :

$$I_C = \sum_{i=1}^n \{(X_i - X_C)^2 + (Y_i - Y_C)^2\} \quad (2)$$

$$I_H = L^2 = (X_H - X_C)^2 + (Y_H - Y_C)^2 \quad (3)$$

$$X_C = \frac{1}{n} \sum_{i=1}^n X_i \quad (4)$$

$$Y_C = \frac{1}{n} \sum_{i=1}^n Y_i \quad (5)$$

Sendo:

I_C = Reflete a dispersão das atividades de um indivíduo ao redor do centroide (C) do conjunto de atividades de um dia, calculado em termos de distância euclidiana;

I_H = Descreve a distância entre a residência e o centroide das atividades do dia em análise;

(X_i, Y_i) = Coordenadas da atividade i ;

(X_C, Y_C) = Coordenadas do centroide dos locais de atividades;

(X_H, Y_H) = Coordenadas do local da residência;

n = Número de atividades do indivíduo no dia em análise;

L = Distância entre a residência e o centroide de atividades.

A Figura 3 exemplifica o conceito de “segundo momento de locais de atividades”:

Figura 3. Exemplo de cálculo do espaço de atividades pelo segundo momento dos locais de atividades.
Fonte: Susilo e Kitamura, 2005.

Foi analisada a variabilidade dos segundos momentos de locais de atividades dos indivíduos entre os dias. Os resultados indicaram que o compromisso de realizar atividades externas à casa, como por exemplo, atividades obrigatórias nos dias de semana e atividades de lazer nos fins de semana, influenciam a extensão do espaço de atividades.

3. Herfindahl-Hirschman Index (HHI)

Susilo e Axhausen (2014) examinaram o grau de repetição das escolhas dos indivíduos em suas combinações diárias de locais de atividades e viagens. Para tanto, foi aplicado o índice Herfindahl-Hirschman nos conjuntos de dados diários de 6 semanas do Mobidrive (Axhausen *et al.*, 2002) e de Thurgau (Löchl *et al.*, 2005; Axhausen *et al.*, 2007).

Normalmente usado na Economia, o Herfindahl-Hirschman Index, também conhecido como HHI, é uma medida de concentração das empresas em relação ao setor a que pertencem e um indicador de competição entre elas. Pode variar de 0 a 1, sendo 0 a representação de um grande número de pequenas empresas enquanto 1 indica a existência de um único grande produtor monopolista (Herfindahl, 1950).

No caso do estudo, o HHI foi usado para medir a repetitividade de combinações idênticas de escolhas individuais de local, atividade e viagem nos múltiplos dias do período observado. Foram usados como atributos a localização de atividades, o motivo das atividades, o modo das viagens e o horário de partida, combinados dois a dois. Quanto mais alto o valor do HHI, maiores os índices de repetição de combinações idênticas no período de seis semanas.

Valores mais baixos para o HHI indicavam que os indivíduos tinham maior tendência de mudar suas combinações de local, atividade e viagem mais frequentemente. Em outras palavras, valores altos do índice são associados com o hábito ou comportamento mais resistente à mudança, enquanto valores baixos são associados a comportamentos menos repetitivos e maior busca por variação.

Portanto, esse índice indica a estabilidade/nível de repetição das escolhas dos indivíduos em relação aos atributos selecionados, no que diz respeito a um evento recorrente (por exemplo, repetição do uso do carro quando visita um determinado local, que pode ser, por exemplo, o trabalho).

A principal vantagem da aplicação do HHI é a sua simplicidade na análise da interdependência das escolhas dos indivíduos em relação a um conjunto de atributos. Seu ponto fraco, no entanto, é que não considera os impactos do encadeamento de viagens que podem influenciar uma ou mais escolhas individuais.

4. Caracterização de Padrões Semanais a Partir de Padrões Diários

Pas (1988) descreveu uma metodologia para análise de padrões de atividades e viagens em múltiplos dias na qual os padrões de múltiplos dias são a composição de alguns tipos de padrões diários. Essa abordagem facilita a identificação de tipos gerais de padrões semanais de viagens e atividades bem como a investigação das relações entre o comportamento semanal e as características sociodemográficas dos indivíduos participantes.

O método foi aplicado para uma amostra de dados de 112 pessoas que trabalham - parte da *Reading Activity Diary Survey*, já apresentada anteriormente. Os resultados mostraram que grupos gerais de comportamento semanal puderam ser identificados e que esses grupos são relacionados a características sociodemográficas particulares dos indivíduos e domicílios. Além disso, os resultados também evidenciaram que, para pessoas empregadas, os padrões diários de viagens e atividades independem do dia da semana, ou seja, os diferentes padrões diários identificados não estão relacionados a dias da semana específicos.

Por fim, os resultados obtidos não refutaram a hipótese de que o padrão semanal depende das características sociodemográficas dos indivíduos, enquanto os padrões diários independem de suas características, sendo afetados na verdade por outras características situacionais geralmente não observadas, como por exemplo: doença de um membro da família, o clima do dia ou problemas com o automóvel.

Os conjuntos de atividades diários foram representados como pontos distintos em um espaço euclidiano, sendo que a proximidade entre eles indicava similaridade entre os padrões

correspondentes. Em seguida foi realizada uma análise de clusters para o agrupamento de padrões similares e interpretação desses grupos.

Cada padrão de atividades diário foi descrito em dois níveis: no primeiro, os atributos considerados foram o tipo de atividade realizada, o horário e a distância do local à casa do indivíduo; para o segundo nível, foi considerado o número total de atividades realizadas naquele dia.

A análise de clusters empregada agrupou os padrões de atividades e viagens diários em um pequeno número de grupos relativamente homogêneos em relação à sua proximidade no “espaço de padrões”. Foram selecionados conjuntos de padrões diários representativos para cada cluster, sendo que para isso análises de subclusters foram feitas para cada cluster.

Para descrição dos principais padrões obtidos, foram selecionados cinco grupos de padrão diário de viagens e atividades, apresentados na tabela a seguir. Para interpretação das descrições dos padrões, o termo “tour” pode ser definido como: *conjuntos de viagens com origem e destino no mesmo local, podendo envolver paradas durante o percurso*.

Tabela 1. Descrição dos tipos de padrões diários de viagens e atividades

Grupo de padrão diário de viagens e atividades	Descrição
D1	Padrões nulos – isto é, nenhuma atividade realizada fora de casa em um dado dia.
D2	Predominantemente um único tour entre a casa e a atividade de subsistência, ocorrendo durante os períodos de pico. Algumas poucas paradas rotineiras ou de lazer.
D3	Duas ou três atividades fora de casa realizadas em um ou dois tours a partir de casa. A primeira atividade é de subsistência, geralmente durante o período de pico da manhã. As demais paradas geralmente são de tarefas rotineiras.
D4	Três ou quatro atividades fora de casa, realizadas geralmente em dois tours longe de casa. Paradas no caminho antes ou após a atividade de subsistência. Geralmente um tour à noite depois de retornar da atividade de subsistência para casa.
D5	Múltiplas atividades fora de casa geralmente em dois tours partindo de casa. Primeira atividade geralmente de subsistência e feita durante o período pico da manhã. Além dessa, são realizadas tanto atividades rotineiras quanto de lazer.

Fonte: Pas, 1988.

Apesar da distribuição dos padrões diários não ter resultado estatisticamente diferente para dias da semana específicos, os resultados apontaram que alguns tipos de padrão diário são mais ou menos prováveis de acontecerem em determinados dias da semana. Por exemplo, o

tipo D2 é menos provável de acontecer em uma quinta-feira do que a média da semana, enquanto o tipo D4 é mais provável de acontecer nesse dia.

A frequência de distribuição dos padrões diários ao longo da semana permitiu, por sua vez, a caracterização de cinco tipos de padrões semanais, em função dos tipos de padrões diários apresentados na Tabela 1. As descrições dos tipos de padrões semanais obtidos são apresentadas na Tabela 2. A mesma definição de “tour” pode ser aplicada.

Tabela 2. Descrição dos tipos de padrões semanais de viagens e atividades

Grupo de padrão semanal de viagens e atividades	Descrição
S1	Principalmente padrões nulos (D1), com algum tour ocasional de uma parada durante os períodos de pico (D2).
S2	Principalmente um único tour para uma atividade, durante os períodos de pico (D2). Ocasionalmente, duas ou três atividades fora de casa realizadas em um ou dois tours a partir de casa (D3).
S3	Principalmente duas ou três atividades fora de casa realizadas em um ou dois tours a partir de casa (D3). Atividade de subsistência geralmente feita durante o período pico da manhã, além de outras atividades rotineiras. Ocasionalmente, padrões de múltiplas paradas (D5).
S4	Padrões diários com um número médio de paradas (D3, D4, D5). São mais prováveis três ou quatro atividades fora de casa, com paradas no caminho antes ou após a atividade de subsistência, bem como um tour à noite depois de retornar da atividade de subsistência para casa.
S5	Principalmente padrões diários de múltiplas paradas (D5). Ocasionalmente, padrões de duas ou três atividades fora de casa realizadas em um ou dois tours a partir de casa (D3).

Fonte: Pas, 1988.

2.3.2. Resultados Obtidos para Diferentes Grupos de Indivíduos

A análise do comportamento de atividades e viagens relacionado às características sociodemográficas e econômicas dos indivíduos foi avaliada pelos diversos autores de duas formas: primeiramente define-se os grupos de indivíduos com características semelhantes para então analisar o comportamento de atividades e viagens; ou são analisadas as atividades e viagens realizadas por diversos indivíduos, agrupando-se aqueles que apresentam comportamento semelhante primeiro, para então avaliar suas características e ver se podem ser encontradas similaridades.

Nesta seção são apresentados os principais resultados obtidos a partir das análises de dados de atividades e viagens de múltiplos dias, considerando diferentes grupos de indivíduos. Na maior parte dos casos estudados, os autores definiram previamente os grupos segundo suas

características sociodemográficas e econômicas para então analisar seu comportamento. Essa abordagem é baseada na hipótese de que as possibilidades individuais de se engajar em eventos e processos são dependentes de um conjunto de circunstâncias ligadas ao indivíduo bem como ao seu ambiente. Consequentemente, o nível de repetição ou variação individual dos locais de atividade e viagens deve ser diferente para diferentes grupos.

Os resultados do estudo de Susilo e Kitamura (2005) apontam para o fato de que nos dias de semana, para trabalhadores e estudantes, a abrangência dos locais de atividade (I_C) e a distância de casa para o centroide dos locais de atividade (I_H) são relativamente estáveis de um dia para outro. Por outro lado, nos fins de semana, quando as atividades tendem a ser mais dispersas, os locais de atividades são mais variados e os espaços de atividades tendem a variar.

Susilo e Axhausen (2014) utilizaram regressões para analisar o índice pessoal de repetição (HHI) dos padrões de viagens em relação aos aspectos sociodemográficos dos entrevistados, para checar se a variabilidade podia ser associada a atributos particulares. Residentes urbanos apresentaram menor variância na combinação de atividades e locais do que residentes rurais. Entretanto, entre os residentes urbanos, a variância nas combinações cresce dos subúrbios urbanos para as regiões mais centrais. Além disso, residentes das áreas rurais tendem a apresentar maior constância na escolha do modo de viagem do que residentes urbanos, provavelmente devido à sua dependência de carros e grandes distâncias para acesso às atividades.

Os resultados obtidos mostram que a repetitividade dos locais e atividades e do modo de transporte é altamente influenciada pelos compromissos dos indivíduos fora de casa, pelas condições internas do domicílio e pela acessibilidade aos locais de atividades. Além disso, o grau de repetição do padrão diário de atividades e viagens é menos correlacionado com o modo de viagem, mas mais com os compromissos e obrigações dos indivíduos. Por exemplo, atividades obrigatórias como trabalho e estudo apresentam padrões mais repetitivos que outras. Como esperado, viagens a lazer e negócios pessoais apresentam maior flexibilidade em seus padrões do que outras viagens. Já a repetição das escolhas dos modos de transporte é mais relacionada às condições ou acessibilidade dos locais de atividade, mas não diretamente relacionada às atividades em si.

Pas e Koppelman (1987) mostraram que indivíduos com menos restrições econômicas, profissionais e familiares apresentam níveis mais altos de variabilidade intrapessoal em suas frequências de viagem. Além disso, indivíduos que possuem necessidades pessoais e domésticas que não requerem viagens diárias para atendê-las apresentam níveis mais altos

de variabilidade intrapessoal em suas viagens, por possuírem maior liberdade para outros tipos de deslocamentos, em relação às pessoas que possuem compromissos que requerem sua participação em atividades diárias e viagens relacionadas. Algumas premissas específicas foram formuladas e demonstradas, são elas:

- Pessoas empregadas apresentam níveis consideravelmente mais baixos de variabilidade intrapessoal do que pessoas não empregadas;
- A variabilidade intrapessoal dos padrões de viagem está diretamente relacionada à ocupação do chefe de família e ao nível de educação do indivíduo;
- Pessoas casadas apresentam níveis de variabilidade intrapessoal inferiores ao de pessoas solteiras;
- Mulheres apresentam níveis mais elevados de variabilidade intrapessoal em seu comportamento diário de viagens do que homens;
- A presença de crianças em casa não apresenta efeitos significativos na variabilidade intrapessoal do comportamento de viagens de homens casados, mas indica grande variação do número de viagens realizadas por mulheres casadas;
- Indivíduos que possuem automóvel apresentam níveis mais altos de variabilidade intrapessoal em seu comportamento de viagens;
- Indivíduos de renda mais alta apresentam níveis mais altos de variabilidade intrapessoal em seu comportamento diário de viagens, devido principalmente a atividades sociais e de lazer.

Pas (1988) identificou um total de 560 padrões diários de viagens e atividades, que correspondem a cinco dias úteis de uma amostra de 112 indivíduos da pesquisa realizada em Reading, na Inglaterra. Os resultados obtidos a partir da análise das características sociodemográficas das pessoas e sua relação com cada um dos cinco padrões de comportamento semanal definidos (ver Tabela 2), mostraram que a composição dos padrões se diferencia entre os seguintes segmentos da população: mulher/homem, baixa renda/alta renda, chefe de família/outras membros da família.

Mais especificamente, homens que trabalham se mostraram ser mais prováveis que a média de pessoas que trabalham a apresentar um padrão de atividades e viagens semanal tipo S2. Esse padrão consiste principalmente em uma viagem para uma atividade, durante os períodos de pico, e ocasionalmente, duas ou três atividades fora de casa realizadas em uma ou duas viagens a partir de casa. Já as mulheres que trabalham são significativamente menos prováveis que a média de pessoas que trabalham a apresentar um padrão de viagens e atividades do mesmo tipo S2.

Quanto a pessoas de baixa renda que trabalham, essas apresentam maior tendência do que a média de pessoas que trabalham a apresentar padrões semanais de viagens e atividades mais “simples”. Em particular, pessoas de baixa renda são significativamente mais prováveis a apresentar um padrão semanal que consiste principalmente em padrões nulos, com uma viagem ocasional de uma parada durante os períodos de pico.

A análise de dados coletados por múltiplos dias contribui para um diagnóstico aprofundado dos padrões de atividades. Usualmente as pesquisas de transporte são realizadas para apenas um dia, portanto a análise do comportamento individual ao longo de múltiplos dias caracteriza um diferencial desse estudo.

Além disso, a falta de dados disponíveis e bibliografia existente acerca da demanda por viagens e das reais necessidades das favelas quanto ao transporte, representam um entrave à integração dos seus residentes à cidade formal e à proposição de melhorias para circulação segura dos seus habitantes. Percebe-se com a isso a necessidade de estudos relacionados ao transporte e ao ambiente urbano nas favelas, justificando também, dentro deste contexto, o estudo proposto.

3. Coleta de Dados

Esse capítulo apresenta o contexto da área de estudo e relata a coleta de dados em Paraisópolis, bem como os principais dados obtidos.

A pesquisa de campo foi realizada como parte de um projeto desenvolvido pelo Banco Mundial, denominado “*Paraisópolis Mobility Diagnosis: Providing More Equitable and Sustainable Transport Options for Vulnerable Populations in the Community of Paraisópolis, São Paulo, Brazil*”, em parceria com o Laboratório de Geoprocessamento (LabGEO) do Departamento de Engenharia de Transportes da Escola Politécnica da Universidade de São Paulo. A pesquisa foi aprovada pelo CEP e pela Conep² em um processo submetido pela Plataforma Brasil, que regula a questão ética de realização de pesquisas envolvendo humanos no país.

3.1. Área de Estudo

Em São Paulo, durante as décadas de 1970 e 1980, as favelas cresceram por toda parte, mas aquelas localizadas em áreas com maior valor imobiliário foram removidas para locais de terra mais barata, ambientalmente mais frágeis, concentrando-se predominantemente na periferia da cidade (Lacerda, 2007; Lima e Somekh, 2013).

Paraisópolis, no entanto, permaneceu localizada no meio do ambiente urbano formal em uma área relativamente bem-dotada³ de serviços públicos, adjacente a bairros ricos como o Morumbi. 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, embora outras estimativas apresentem valores superiores, e 17.700 domicílios⁴), além de apresentar densidade construída 12 vezes maior que a média da cidade (Carvalho,

² O sistema CEP-CONEP foi instituído em 1996 para proceder a análise ética de projetos de pesquisa envolvendo seres humanos no Brasil. Este processo é baseado em uma série de resoluções e normativas deliberados pelo Conselho Nacional de Saúde (CNS), órgão vinculado ao Ministério da Saúde. A maioria dos processos relacionados à análise ética ocorre em ambiente eletrônico por meio da ferramenta eletrônica chamada Plataforma Brasil.

Disponível em: <http://plataformabrasil.saude.gov.br/>

³ Heliópolis, a maior favela de São Paulo por população, também se manteve em uma área bem servida da cidade.

⁴ Censo 2010 do IBGE (Instituto Brasileiro de Geografia e Estatística).

Disponível em: <https://censo2010.ibge.gov.br/>.

2009). A Figura 4 apresenta a localização de Paraisópolis e a Figura 5, uma vista aérea, na qual pode-se observar o contexto urbano em que está inserida, sendo notável a diferença do padrão de construção das residências de Paraisópolis e dos bairros próximos.

Figura 4. Localização de Paraisópolis. Fonte: Elaboração própria, com dados do Metrô de São Paulo⁵ e CEM – Centro de Estudos da Metrópole⁶.

⁵ Base georreferenciada das Zonas OD 2007 disponível em: <https://transparencia.metrosp.com.br/dataset/pesquisa-origem-e-destino/>

⁶ Base georreferenciada da rede viária disponível em: <http://web.fflch.usp.br/centrodametropole/>

Figura 5. Vista aérea de Paraisópolis. Fonte: Google Earth, 2016.

Paraisópolis foi desenvolvida inicialmente em terra privada, o que dificultou processos de intervenção pública. A comunidade cresceu enquanto delimitava os quarteirões e principais ruas em seu território, restrito à presença de terrenos íngremes sujeitos a riscos de deslizamentos e enchentes. Ao mesmo tempo, a alta densidade de moradia informal, existente nas estreitas ruas internas e vielas, complicaram futuros projetos de transporte e urbanização.

Em 2005, a Secretaria Municipal de Habitação (SEHAB) lançou o Plano de Urbanização de Paraisópolis. A primeira fase do plano foi completada em 2009 e focou principalmente em desafios legislativos a respeito da terra privada enquanto adquiria espaço para reassentar 11% dos domicílios localizados em áreas de risco. No entanto, a comunidade permaneceu com baixos níveis de renda e acesso limitado a serviços básicos.

A segunda e terceira fases do Plano de Urbanização de Paraisópolis ampliaram os serviços básicos em 2011. Com essas ações, em 2012 Paraisópolis passou de 17% para 75% de cobertura de esgoto em três anos (Santos, 2012). Também em 2012, Paraisópolis tinha 7 centros médicos, 25 escolas, 14 creches e 11 igrejas (França e Costa, 2012).

Os planos de urbanização são favoráveis às favelas ao promoverem acesso a serviços básicos, como água encanada, eletricidade e serviços de resíduos sólidos, mas geralmente não consideram aspectos de mobilidade.

Paraisópolis vivenciou uma significativa melhora em suas condições econômicas e sociais ao longo dos últimos dez anos, seguida pelo aumento no preço da terra (Meyer et al., 2016). O maior acesso ao crédito resultou em um aumento da posse de automóvel, com consequente

aumento do congestionamento interno. Em 2007, 10% dos residentes possuíam carro (Metrô SP); apesar de não haver dado mais recente, é esperado que esse valor tenha aumentado.

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 Figura 6 ilustra locais onde ocorreram acidentes de trânsito em Paraisópolis, nos anos de 2010 e 2017.

Figura 6. Acidentes de trânsito 2010 e 2017, em Paraisópolis e região. Fonte: Elaboração própria, com dados da Companhia de Engenharia de Tráfego (CET-SP).

Existem sete linhas oficiais de ônibus que passam por Paraisópolis (Figura 7), sendo que cinco delas operam com vans ou micro-ônibus, a fim de permear suas estreitas vias, embora não haja pontos de parada dentro de Paraisópolis, o que força as pessoas a caminharem até os pontos localizados no entorno.

Figura 7. Linhas de ônibus que passam por Paraisópolis e pontos de ônibus localizados no entorno.
Fonte: Elaboração própria com dados do GeoSampa⁷.

Segundo Arbex et al. (2016), menos de 20% dos empregos formais disponíveis no município são acessíveis a 1 hora de deslocamento de Paraisópolis a partir de transporte público. Os tempos de viagem dos ônibus que atendem a área são altamente imprevisíveis devido ao congestionamento, com tempos de viagem frequentemente 2,5 vezes mais longos do que o tempo de fluxo livre. Isso tem implicações na organização das atividades diárias, já que é necessário contar com uma margem de tempo de segurança para chegar aos destinos no horário desejado.

Atualmente, os sistemas de transporte sobre trilhos mais próximos da favela são a Linha 5-Lilás do Metrô (Companhia do Metropolitano de São Paulo) e a Linha 9-Esmeralda da CPTM (Companhia Paulista de Trens Metropolitanos) (Figura 8). Um aspecto positivo a ser ressaltado é que duas estações de monotrilho da Linha 17-Ouro, atualmente em construção,

⁷ Bases georreferenciadas disponíveis em: <http://geosampa.prefeitura.sp.gov.br/>

estão planejadas para a região de Paraisópolis. Essas estações conectarão a comunidade ao sistema de metrô, o que deverá aliviar algumas das atuais restrições ao acesso dos habitantes à rede de transporte coletivo da cidade.

Figura 8. Linhas de ônibus, metrô e trem na região de Paraisópolis. Fonte: Elaboração própria com dados do GeoSampa.

3.2. Desenho da Pesquisa

A coleta de dados em Paraisópolis foi realizada em duas fases. Na primeira fase (piloto), foram coletados dados de uma pequena amostra não representativa (21 indivíduos). A segunda fase prévia à 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 (apresentado no Anexo 1), sobre seus dados 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.

Aplicativo *FollowMee*

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 permite 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.

Amostra Piloto (Fase 1)

A primeira fase da pesquisa, realizada com uma pequena amostra de moradores, foi realizada em outubro de 2015. A equipe de campo foi formada por colaboradores do Laboratório de Geoprocessamento (LabGEO) da Escola Politécnica da Universidade de São Paulo (POLI-USP), do Banco Mundial e outros voluntários.

A amostra foi composta por 21 residentes, 52% de sexo feminino e 48% de sexo masculino. A entrevista em questionário coletou o perfil do residente (composição da família, dados socioeconômicos, aspectos de saúde) e hábitos de transporte.

Essa etapa foi importante para testar o aplicativo selecionado, bem como o processo de compra de créditos e do pacote de dados 3G, a fim de garantir o acesso à internet. O aplicativo foi instalado com sucesso em 17 *smartphones* (80% dos participantes). Os pesquisadores forneceram instruções para desinstalar o aplicativo quando a coleta de dados fosse finalizada. A Figura 9 ilustra alguns dos participantes desta fase sendo entrevistados pelos pesquisadores.

Figura 9. Aplicação da primeira fase da pesquisa em Paraisópolis (outubro de 2015). Foto: Irene Quintáns.

Pesquisa Oficial (Fase 2)

A segunda fase da pesquisa foi realizada entre 24 de maio e 5 de agosto de 2016 por uma empresa com experiência na área – a Cobrape⁸, que apresentou boas estratégias para se aproximar da comunidade e estabelecer contato com a população, com cuidado especial em relação aos líderes locais.

O treinamento dos pesquisadores de campo foi realizado pelo LabGEO em 16 de maio de 2016 (Figura 10). Durante o treinamento, foram apresentados o escopo do projeto, os objetivos da pesquisa, a metodologia adotada, o processo de instalação do aplicativo e de aquisição de crédito e pacote de dados 3G para os *smartphones* dos participantes.

Figura 10. Treinamento de pesquisadores de campo no LabGeo. Foto: Cobrape.

3.2.1. Planejamento Amostral

Segundo Ortúzar e Willumsen (2011), o dimensionamento do tamanho da amostra em estudos de transporte pode seguir dois caminhos: (a) ser baseado em uma resposta quantitativa, sendo necessário conhecer a variância esperada para esse valor ou (b) ser baseado em uma resposta binomial (de resposta sim ou não).

No caso de variáveis categóricas, a amostra pode ser calculada como:

⁸ Companhia Brasileira de Projetos e Empreendimentos (<http://www.cobrape.com.br/>).

$$n = \frac{Z^2 \cdot p \cdot q}{E^2} \quad (6)$$

Sendo:

n = Número de indivíduos na amostra;

Z = Valor crítico da normal padronizada correspondente ao grau de confiança desejado. Para um grau de confiança de 95%, é igual a 1,96;

p = Proporção populacional de indivíduos que pertence à categoria de interesse (no desconhecimento da proporção, adota-se 50%, resultando na maior amostra possível);

q = Proporção populacional de indivíduos que não pertence à categoria de interesse ($q = 1 - p$).

E = Erro aceitável relativo;

Para o caso de Paraisópolis, onde a população do universo pode ser considerada finita, aplica-se um fator de correção à fórmula anterior, resultando na seguinte fórmula corrigida:

$$n = \frac{N \cdot p \cdot q \cdot Z^2}{p \cdot q \cdot Z^2 + (N - 1) \cdot E^2} \quad (7)$$

Sendo N = Tamanho da população.

Na Tabela 3 são apresentados diferentes tamanhos de amostra de indivíduos, de acordo com os valores de erro amostral admissíveis, para a população aproximada de Paraisópolis, estimada em 70.000 pessoas.

Tabela 3. Amostra de indivíduos calculada por erro amostral

Erro amostral (E)	Amostra (n)
3%	1.051
4%	595
5%	381
7%	195
10%	96

Admitindo-se um erro amostral máximo de 5%, são necessários **381 indivíduos** para representar a população estudada. 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 de maneira aleatória, 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 (Figura 11). Em cada subárea, foram selecionados no mínimo 20 e no máximo 40 domicílios.

Figura 11. Subáreas de amostragem dos domicílios. Fonte: Elaboração própria.

Os 381 domicílios sorteados aleatoriamente formaram uma lista prioritária de domicílios a serem abordados, chamada "Lista A". Nos domicílios em que não foi encontrado nenhum residente na primeira visita, foi deixado um aviso informando a data e hora de retorno do pesquisador. Se nenhum residente daquele domicílio foi encontrado em três visitas e mais uma quarta visita no fim de semana, ou houve recusa de participação, outro domicílio de uma "Lista B" (contendo 1.170 domicílios aleatoriamente sorteados) seria então selecionado.

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.

Sabe-se que há maior homogeneidade entre as unidades amostrais dentro de um mesmo estrato de população, por exemplo: sexo, idade, condição socioeconômica. Na tentativa de

diversificar a amostra e garantir alguns mínimos de entrevistas realizadas para cada estrato de população em Paraisópolis, foram estabelecidas quotas gerais que a pesquisa deveria atingir, considerando todos os participantes. Assim, a pesquisa deveria apresentar no mínimo as seguintes porcentagens da amostra de 381 residentes para cada categoria definida:

- 40% são do sexo feminino e 40% são do sexo masculino;
- 50% têm um emprego (formal ou informal);
- 20% trabalham fora da comunidade;
- 20% trabalham dentro da comunidade;
- 20% desloca-se pelo menos 3 vezes por semana;
- 20% desloca-se pelo menos 3 vezes por semana por modos motorizados.

A fim de verificar continuamente as quotas estabelecidas e, se necessário, ajustar a amostra da pesquisa, a empresa Cobrape desenvolveu um *website* de monitoramento. Esta ferramenta permitiu a consulta direta do progresso da coleta por atributos, como subárea da pesquisa, percentuais de gênero, trabalho e tipo de transporte utilizado. A Figura 12 apresenta o formulário de acompanhamento de campo, enquanto a Figura 13 ilustra as informações do *website* de monitoramento sobre o progresso da coleta de dados.

Figura 12. Formulário de acompanhamento de campo. Foto: Irene Quintáns.

Figura 13. Website de acompanhamento da pesquisa, com dados e gráficos em tempo real sobre o progresso da coleta. Fonte: Cobrape.

3.2.2. Desafios da Pesquisa

Esse item apresenta barreiras sociais e técnicas, bem como ações mitigadoras adotadas para implementação da pesquisa e coleta dos dados de atividades e dos indivíduos, além de avaliar as dificuldades de atingir a amostra selecionada.

Algumas dificuldades não previstas se sucederam durante o trabalho de campo. As cartografias dos domicílios obtidas junto à SEHAB não estavam atualizadas, pois o último mapeamento tinha sido realizado em 2006. Devido às múltiplas mudanças ocorridas nesse período em Paraisópolis, os dados disponíveis eram imprecisos ou não estavam suficientemente bons para identificação dos domicílios e seus acessos de forma fácil e rápida. Com isso, o conhecimento de campo da empresa contratada foi fundamental para localizar os domicílios selecionados. A utilização de mapas georreferenciados ajudou a identificar os domicílios visitados nas subáreas definidas.

A Figura 14 apresenta um exemplo de mapa georreferenciado para acompanhamento dos domicílios sorteados, neste caso para a subárea 01. Os domicílios eram diferenciados por cor, segundo o status da entrevista: 1ª, 2ª ou 3ª visita realizada, ainda não visitado, concluído (com/sem aplicativo), substituído.

Figura 14. Mapa georreferenciado para identificação dos domicílios visitados. Fonte: Cobrape.

A desconfiança das pessoas com relação ao projeto foi uma das razões para recusas de participação na pesquisa; ainda assim os pesquisadores se esforçaram em explicitar os benefícios de participar.

Os dados de localização (GPS) foram coletados a uma baixa frequência (5 minutos), considerando a necessidade de evitar gasto de bateria dos *smartphones*, dado que não se sabia as especificações dos aparelhos que uma população de baixa renda possui. Conforme apontado na revisão bibliográfica, outros diversos aplicativos foram desenvolvidos, inclusive gratuitos, desde a realização da pesquisa, em 2016.

A Tabela 4 apresenta um resumo das principais barreiras encontradas para a coleta de dados e respectivas ações propostas para superá-las.

Tabela 4. Principais barreiras para a coleta de dados e ações propostas (Pizzol et al., 2017)

Estágio da pesquisa	Riscos/barreiras	Ações
Seleção da tecnologia	<ul style="list-style-type: none"> Alto consumo de bateria, perda de sinal do GPS; Frequência de registro do GPS versus precisão no estágio de processamento de dados; Necessidade de interagir com o usuário para obtenção dos dados. 	<ul style="list-style-type: none"> Seleção de intervalo de coleta que não comprometeu a bateria (5 minutos); Seleção de aplicativo que permitia a extração remota dos dados, salvando-os no servidor.
Processo de amostragem	<ul style="list-style-type: none"> Base de residências desatualizada (2006); Unidades de habitação irregulares e não registradas; Mapas e endereços não confiáveis. 	<ul style="list-style-type: none"> Amostragem de domicílios espacialmente estratificada, com cobertura espacial mínima em áreas de difícil acesso; Equipe da pesquisa de campo com experiência na área.

Estágio da pesquisa	Riscos/barreiras	Ações
Contato com o residente do domicílio	<ul style="list-style-type: none"> • Acesso restrito a áreas e períodos limitados durante o dia e nos finais de semana (devido ao crime local organizado); • Endereços não localizados; • Ninguém em casa. 	<ul style="list-style-type: none"> • Reunião com as partes interessadas; • Parceria com grupos confiáveis; • Apresentações de resultados para os líderes locais em diferentes etapas da pesquisa; • Foram realizadas até três visitas para encontrar um residente no domicílio.
Identificação de pessoas qualificadas para a pesquisa	<ul style="list-style-type: none"> • Adultos (maior que 16 anos); • Posse de <i>smartphone</i> com acesso à internet; • Aceitação em participar. 	<ul style="list-style-type: none"> • Disponibilização de créditos e pacote de dados para acesso à internet por duas semanas.
Instalação do aplicativo e inserção de créditos para acesso à internet	<ul style="list-style-type: none"> • Questões de privacidade; • Especificações técnicas do <i>smartphone</i>. 	<ul style="list-style-type: none"> • Formulário de consentimento de participação e acordo de confidencialidade; • Manual de instalação do aplicativo; • Treinamento da equipe de campo para a instalação do aplicativo; • Equipe do escritório apoiando a equipe de campo sobre diferentes especificações dos <i>smartphones</i> e situações adversas.
Coleta e processamento dos dados	<ul style="list-style-type: none"> • Períodos menores que 14 dias; • Dias descontínuos; • Sinal de GPS ruim em algumas áreas de Paraisópolis. 	<ul style="list-style-type: none"> • Período mais longo (2 semanas) para obter dados de uma semana completa (7 dias); • Interpolação temporal dos dados a cada 2 minutos.

3.2.3. Taxas de Resposta

A seguir são apresentados os números finais da coleta de dados. 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.

Segundo indicado na Tabela 5, 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).

Tabela 5. Indivíduos contatados na pesquisa

Status Final da Coleta de Dados	Indivíduos	%
Concluídos com coleta de dados de GPS	272	34,1%
Concluídos, mas desinstalaram o aplicativo em seguida	88	11,0%
Concluídos sem instalação do aplicativo	21	2,6%
Recusaram participar	271	33,9%
Não possuíam <i>smartphone</i>	147	18,4%
Total	799	100%

3.3. Dados de GPS dos *Smartphones*

O aplicativo *FollowMee* foi configurado para rastrear a localização do GPS do *smartphone* de cada indivíduo em intervalos de 5 minutos. As informações atualizadas eram enviadas para um servidor sempre que houvesse uma conexão Wi-Fi disponível. Usando a interface do website <http://www.followmee.com/>, após o período de coleta de dados, foi realizado o *download* do banco de dados do servidor para cada indivíduo. Conforme apresentado na Tabela 6, os registros de localização de cada *smartphone* eram organizados em ordem cronológica (campo *Date*). Os demais atributos são descritos em seguida.

Tabela 6. Exemplo de registros de localização para um indivíduo, obtida a partir do website do aplicativo *FollowMee*

Id_ GPS	Date	Lat	Long	Speed (km/h)	Altitude (m)	Accuracy (m)	Type	Distance (km)
0	25/05/2016 00:05	-23.61878	-46.73117	3	819	5	GPS	
1	25/05/2016 00:23	-23.61874	-46.73131	0	805	200	GPS	0.02
2	25/05/2016 00:29	-23.61872	-46.73106	11	829	10	GPS	0.03
3	25/05/2016 00:35	-23.61879	-46.73094	0	947	30	GPS	0.02
4	25/05/2016 00:41	-23.61857	-46.73118	2	768	30	GPS	0.04
5	25/05/2016 01:03	-23.61875	-46.73115	0	793	5	GPS	0.03

Id_ GPS	Date	Lat	Long	Speed (km/h)	Altitude (m)	Accuracy (m)	Type	Distance (km)
6	25/05/2016 01:30	-23.61861	-46.73132	3	793	30	GPS	0.03

- **Id_GPS:** Código de identificação do registro de localização do GPS.
- **Date:** Dia e horário do registro de localização.
- **Lat:** Latitude da localização, no sistema de coordenadas geográficas WGS84.
- **Long:** Longitude da localização, no sistema de coordenadas geográficas WGS84.
- **Speed:** Velocidade instantânea registrada pelo GPS, em km/h.
- **Altitude:** Altitude registrada pelo GPS, em metros.
- **Accuracy:** Precisão do GPS, em metros.
- **Type:** Tipo do dado.
- **Distance:** Distância medida pelo GPS em relação ao registro anterior, em metros.

No caso desta pesquisa só foi coletado um tipo de dado – GPS, não tendo sido feito uso de redes de celular, redes de *Wi-Fi*, acelerômetro ou *Bluetooth*.

O aplicativo estava configurado para registrar a localização continuamente, mesmo quando o indivíduo estava parado no mesmo lugar. Portanto, não havia rastreamento somente em momentos em que o *smartphone* estava desligado ou o sinal do GPS estava fraco. Nestes casos, a localização era registrada em intervalos maiores que 5 minutos, causando falhas na sequência de registros. No exemplo da Tabela 6, pode-se observar 18 minutos de intervalo entre os registros 0 e 1; 22 minutos entre os registros 4 e 5; e 27 minutos entre os registros 5 e 6.

Dos 272 participantes que coletaram dados com o aplicativo, 175 (ou 64,3%) tiveram dados coletados por mais de uma semana. Em média, foram coletados dados por 8 dias por indivíduo, variando de 0 a 15 dias, conforme ilustram a Tabela 7 e a Figura 15.

Tabela 7. Número de dias de coleta

Número de dias de coleta	Frequência absoluta de indivíduos (nº)	Frequência relativa de indivíduos (%)	Frequência relativa acumulada (%)
0	8	2,9%	2,9%
1	10	3,7%	6,6%
2	13	4,8%	11,4%
3	10	3,7%	15,1%
4	17	6,3%	21,4%

Número de dias de coleta	Frequência absoluta de indivíduos (nº)	Frequência relativa de indivíduos (%)	Frequência relativa acumulada (%)
5	12	4,4%	25,8%
6	27	9,9%	35,7%
7	12	4,4%	40,1%
8	18	6,6%	46,7%
9	9	3,3%	50,0%
10	11	4,0%	54,0%
11	9	3,3%	57,3%
12	18	6,6%	63,9%
13	84	30,9%	94,8%
14	13	4,8%	99,6%
15	1	0,4%	100,0%
Total	272	100%	-

Figura 15. Polígono de frequências de dias de coleta por indivíduo. Fonte: Elaboração própria.

4. Processamento dos Dados

A abordagem adotada para análise do comportamento dos indivíduos neste estudo se baseou nas atividades que realizaram em um período de sete dias consecutivos, escolhido por permitir representar o padrão de atividades de uma semana completa.

Após a coleta dos dados de localização, apresentada no capítulo anterior, foi necessário realizar uma limpeza dos dados brutos, seguida por um processamento intensivo dos dados, para que fosse possível usá-los como entrada no processo de agrupamento das atividades, e em seguida, dos indivíduos. Neste capítulo foram desenvolvidas as etapas do método destacadas na Figura 16.

Figura 16. Etapas do método desenvolvidas no Capítulo 4. Fonte: Elaboração própria.

4.1. Definições

Primeiramente, são definidos alguns conceitos adotados neste trabalho, que serão mencionados por diversas vezes neste e nos próximos capítulos:

- Pontos de GPS: Pontos de localização do GPS, obtidos a partir do aplicativo instalado no *smartphone* de cada indivíduo.
- POI's (*Points of Interest*): Locais de interesse, onde os indivíduos realizam suas atividades cotidianas, como o local de sua residência, locais de trabalho, locais de estudo, locais de compras e serviços.

- Atividades: Cada atividade corresponde a uma visita a um determinado POI, que começa em algum horário do dia e ocorre durante um tempo específico. Diversas atividades podem acontecer no mesmo POI, em diferentes horários do mesmo dia ou em diferentes dias da semana. Essa definição corresponde ao que geralmente é denominado *Stay* na literatura.
- Categorias de atividades: Tipos de atividades definidos a partir de seus atributos, como o horário de início, a duração, o local em que ocorrem, entre outros, visto que não há informação sobre o motivo das atividades.
- Padrões de atividades: Caracterização do conjunto de atividades que uma pessoa realiza, a partir de determinados critérios, como por exemplo, o número total de atividades, a repetição ou variação de diferentes tipos de atividades.
- Categorias de indivíduos: Tipos de pessoas que apresentam padrões de atividades semelhantes, podendo ser socioeconomicamente homogêneos ou não.
- Caracterização socioeconômica: Atributos de cunho social, demográfico ou econômico que uma pessoa apresenta, como gênero, idade, renda, composição da família, condição de trabalho, grau de escolaridade, posse de automóvel, entre outros.

4.2. Limpeza dos Dados e Seleção dos Indivíduos para Análise

O processo de limpeza dos dados se iniciou com o conjunto dos 272 indivíduos que não desinstalaram o aplicativo logo após concluir a entrevista, e chegaram a ter seus dados de localização registrados por algum período, conforme descrito no item 3.3.

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 (ver item 3.3). 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.

Também ocorreram casos em que os dados sofreram grandes saltos de localização (ou seja, *outliers*), com registro de pontos que ocorreram repentinamente a quilômetros de distância em um curto período de tempo. Assim, alguns tratamentos foram necessários para minimizar a presença de tais ocorrências.

Em relação aos *outliers*, foi adotado um critério de distância de 10 km em relação ao registro anterior, correspondente à situação da pessoa estar em um automóvel a 120 km/h, com os dados sendo coletados a cada 5 minutos. Portanto, todos os dados de GPS cujo valor do campo *Distance* (correspondente à distância em relação ao ponto anterior) era maior que 10

km foram excluídos, o que correspondeu a 0,05% do total de registros coletados ao longo de todo o período.

Os altos intervalos de tempo observados entre dados de GPS foram tratados segundo uma abordagem conservadora: indivíduos que tiveram 6% ou mais dos intervalos entre registros de GPS maiores que 30 minutos foram excluídos. A Figura 17 apresenta a distribuição de intervalos entre todos os registros de GPS, excluídos os *outliers*, podendo-se observar que cerca de 80% deles foram coletados em intervalos de até 30 minutos.

Figura 17. Distribuição de intervalos entre os registros de GPS. Fonte: Elaboração própria.

A Figura 18 ilustra a relação entre o número de registros por indivíduo e a respectiva porcentagem de intervalos maiores que 30 minutos. Pode-se notar que grandes porcentagens de altos intervalos estão mais associadas a indivíduos que coletaram menor volume de dados.

Figura 18. Relação entre o número de registros por indivíduo e a porcentagem de intervalos maiores que 30 minutos. Fonte: Elaboração própria.

A Figura 19, por sua vez, apresenta a porcentagem de intervalos maiores que 30 minutos por indivíduo com indicação do limite de 6%, critério adotado para a exclusão dos indivíduos.

Figura 19. Porcentagem de intervalos maiores que 30 minutos por indivíduo. Fonte: Elaboração própria.

Com isso, partindo dos 272 participantes, a amostra foi reduzida para 189 indivíduos.

Para a análise dos padrões de atividades de uma semana completa, 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. Mais detalhes desse procedimento podem ser encontrados no trabalho de Bianco (2018).

Por fim, durante a fase de identificação e caracterização das atividades, que será apresentada em detalhe no item 4.3.3, foram identificadas 4 pessoas cujo celular enviou dados sempre da mesma localização e, por ser uma situação atípica, tiveram que ser eliminadas, já que a falta de registro de algum movimento qualquer pode estar associada a problemas na coleta.

Portanto, a amostra final contém 105 indivíduos, cujos dados de GPS e informações coletadas pela entrevista apresentam as características necessárias para identificação de padrões de atividades e análise do perfil socioeconômico associado. A Figura 20 resume o processo de redução progressiva da amostra.

Figura 20. Processo de seleção dos indivíduos para análise. Fonte: Elaboração própria.

4.3. 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. A seguir esse processo é apresentado em detalhe, incluindo a análise de sensibilidade e inspeção visual, realizadas para validação do método e resultados obtidos.

4.3.1. Density-Based Clustering Algorithm (DBSCAN)

O método usado para identificação dos POI's foi implementado inicialmente pelo pesquisador Renato Arbex, do Departamento de Engenharia de Transportes da Escola Politécnica da Universidade de São Paulo.

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 (*Density-Based Spatial Clustering of Application with Noise*) 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. Ester et al. (1996) salientam que qualquer função de distância pode ser usada para *Eps*, sendo que uma função apropriada deverá ser escolhida de acordo com a aplicação.

O algoritmo inicia aleatoriamente a partir de um ponto e pesquisa por pontos vizinhos que estejam dentro da distância máxima estabelecida para formação de um cluster. Esse conjunto de pontos se expandirá caso os pontos vizinhos estiverem suficientemente próximos e compuserem o mínimo de pontos necessário para a formação de um cluster. Caso contrário, os pontos vizinhos serão designados como ruído. Este processo é repetido até que nenhum ponto adjacente possa ser encontrado. Se o número de pontos em qualquer cluster for menor

que o valor definido para o parâmetro *MinPts*, ele será descartado. A Figura 21 ilustra clusters identificados pelo DBSCAN, diferenciados pela cor dos pontos.

Figura 21. Ilustração de clusters identificados pelo DBSCAN. Fonte: Ester et al., 1996.

O processo foi realizado em dois estágios, sendo a identificação de atividades o primeiro estágio, seguido pela identificação de POI's:

Identificando atividades: No primeiro estágio, uma distância equivalente foi considerada, combinando distância e tempo entre cada par de pontos do GPS. Assim, visitas diferentes ao mesmo local (por exemplo, em diferentes períodos do dia ou dias diferentes) foram identificadas como atividades distintas, caracterizando-as como clusters espaço-temporais de pontos do GPS. A localização geográfica da atividade foi configurada para ser o centroide de todos os pontos do cluster.

A Figura 22 ilustra a composição da distância equivalente usada para identificação das atividades, a partir das métricas de distância euclidiana e tempo convertido em distância.

Figura 22. Exemplo de ilustração da distância equivalente usada para identificação das atividades. Fonte: Adaptação de esquema proposto por Renato Arbex.

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.

As diferentes atividades foram entradas 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.

Enquanto o número mínimo de pontos para formação de clusters no primeiro estágio foi variado (10 ou 15, conforme apresentado no item 4.3.2); na segunda etapa este parâmetro foi fixado como 1, uma vez que mesmos locais que foram visitados apenas uma vez ao longo da semana também devem ser considerados um POI do indivíduo.

4.3.2. Análise de Sensibilidade e Inspeção Visual

Uma análise de sensibilidade foi conduzida, visando selecionar a configuração ótima de parâmetros para o processamento do DBSCAN, seguido por um teste de inspeção visual para validação dos resultados.

Três parâmetros podem ser definidos, segundo a abordagem adotada nesse estudo: i) a distância máxima entre dois pontos consecutivos (*Eps 1*); ii) o intervalo de tempo máximo entre dois pontos consecutivos (*Eps 2*) e iii) o número mínimo de pontos (*MinPts*) necessário para formar um cluster. Dois valores alternativos foram testados para cada um dos três parâmetros: i) 50 e 100 metros para a distância máxima; ii) 30 ou 60 minutos para o intervalo máximo de tempo e iii) 10 ou 15 pontos mínimos.

Pontos de GPS com precisão excedendo o parâmetro de distância máxima foram excluídos antes de processar o algoritmo. Considerando a distância de 100 metros, foram cerca de 13% do total de pontos de GPS, e 23% para a distância de 50 metros.

A estatística *silhouette* foi então usada para avaliar a eficácia do processo de clusterização. Valores mais altos de *silhouette* indicam melhor discriminação entre clusters. A Tabela 8 apresenta os resultados da análise de sensibilidade, considerando os 272 indivíduos que coletaram dados de GPS por algum período. Os valores indicam que a melhor combinação dos parâmetros para aplicação do DBSCAN foi: *Eps 1* = 100 metros, *Eps 2* = 30 minutos e *MinPts* = 15. Essa configuração tem implicações no número, localização e duração das atividades, que corresponderam a 6.108 atividades e 1.974 POI's.

Tabela 8. Análise de sensibilidade dos parâmetros para o algoritmo DBSCAN

Eps 1 (metros)	Eps 2 (minutos)	Min Pts	Número de Atividades	Silhouette Calculado
50	30	10	8.879	0,83
100	30	10	7.107	0,88
50	60	10	7.823	0,81
100	60	10	6.512	0,88
50	30	15	7.735	0,85
100	30	15	6.108	0,90
50	60	15	6.565	0,83
100	60	15	5.514	0,89

Uma análise visual usando imagens do *Google Maps* em um ambiente SIG (*Sistema de Informação Geográfica*) foi realizada para avaliar a qualidade do processo de identificação dos POI's. Seis indivíduos foram selecionados aleatoriamente; para cada um deles, os POI's resultantes para cada combinação de parâmetros do DBSCAN foram comparados com os dados de GPS brutos.

A Figura 23 ilustra os resultados obtidos para um dos indivíduos. O exemplo apresenta a variação de apenas dois parâmetros, tempo e distância, já que o número mínimo de pontos (*MinPts*) igual a 15 apresentou valores de *silhouette* sistematicamente superiores, segundo os resultados da Tabela 8. Os parâmetros de distância (50 e 100 metros) são variados nas colunas e os de tempo (30 e 60 minutos) nas linhas.

No quadrante superior direito são exibidos os POI's, decorrentes das atividades identificadas para *Eps 1* = 100 metros e *Eps 2* = 30 minutos, combinação que correspondeu ao maior valor de *silhouette*, representados por círculos vazios pretos. As seguintes cores foram usadas para representar os POI's obtidos para as demais variações propostas: laranja para *Eps 1* = 50 metros e *Eps 2* = 30 minutos; amarelo para *Eps 1* = 50 metros e *Eps 2* = 60 minutos e verde para *Eps 1* = 100 metros e *Eps 2* = 60 minutos. Os dados brutos do GPS são representados pelos pontos azuis.

É um verdadeiro desafio verificar a precisão dos POI's, dada a alta densidade em Paraisópolis, onde residências, lojas e serviços estão localizados bastante próximos uns dos outros. Para tanto, a análise visual se deu nas seguintes etapas: i) verificar se a localização dos POI's coincide com uma alta concentração de dados brutos do GPS (a fim de investigar também a influência do enriquecimento dos dados através da interpolação realizada); ii) analisar a proximidade dos POI's (supondo que seja menos provável que POI's próximos correspondam

a diferentes locais de atividade); e iii) verificar se os mesmos POI's foram identificados com as diferentes configurações de parâmetros e ao longo dos dias, quando aplicável.

O *Eps 2* parece afetar menos a identificação de atividades e, conseqüentemente, de POIs, do que o *Eps 1*, já que 30 e 60 minutos resultaram em número e localização de POI's semelhantes. Então, o limite de 30 minutos foi adotado, pois há atividades curtas que podem ser realizadas nesse intervalo de tempo.

Quanto ao parâmetro de distância, caso seja definido um limite muito baixo, o número de atividades e POI's decorrentes pode ser superestimado. De fato, foram identificados mais POI's para 50 metros do que para 100 metros. Mas os resultados para os clusters de 50 metros não pareciam representar tão bem os POI's visto que: (i) os clusters não coincidiram com uma alta densidade de dados originais de GPS (o que pode ser observado no POI número 2, apresentado na figura) e (ii) alguns clusters foram gerados muito próximos entre si (três clusters diferentes para o que provavelmente é um mesmo POI número 1, gerado pelas configurações com distância de 100 metros).

Figura 23. Inspeção visual da distribuição espacial dos Pontos de Interesse (POI's) para um indivíduo selecionado aleatoriamente. Fonte: Elaboração própria, com imagens do Google Maps.

Com isso, a partir da inspeção visual, a combinação de parâmetros de 100 metros, 30 minutos e mínimo de 15 pontos, que apresentou o maior *silhouette*, foi a escolhida

4.3.3. Identificação de Atividades e POI's

Quando do processamento do DBSCAN, os pontos que não foram atribuídos a nenhum dos clusters formados no primeiro estágio, foram tratados como ruído e interpretados como viagens entre diferentes locais de atividade (POI's).

Com base nos resultados, foi possível associar a cada registro de GPS, uma Atividade e um POI. Para cada indivíduo, as atividades foram numeradas em ordem crescente, de acordo com a sequência cronológica dos registros de localização do GPS. Com relação a atividades diferentes, deve-se observar que podem se referir a um mesmo POI ou a POI's distintos, com isso a numeração dos POI's por vezes se repete. Por sua vez, todos os ruídos foram designados como atividades "-1". A Tabela 9 apresenta exemplos de associação dos registros de localização do indivíduo de IdPessoa=3 a atividades e POI's. Os atributos são descritos em seguida.

Tabela 9. Exemplo de identificação de Atividades e POI's após o processamento do algoritmo DBSCAN

Id_GPS	Id Pessoa	Dados Artificiais	Date Original	Date Interpolado	Lat	Long	Atividade	POI
839	3	0	04/06/2016 10:45	04/06/2016 10:45	-23.62054	-46.73238	-1	-
840	3	1	04/06/2016 10:45	04/06/2016 10:47	-23.62054	-46.73238	-1	-
841	3	0	04/06/2016 10:50	04/06/2016 10:50	-23.62449	-46.73397	-1	-
842	3	1	04/06/2016 10:50	04/06/2016 10:52	-23.62449	-46.73397	-1	-
843	3	0	04/06/2016 10:55	04/06/2016 10:55	-23.62819	-46.73225	39	19
844	3	1	04/06/2016 10:55	04/06/2016 10:58	-23.62819	-46.73225	39	19
845	3	0	04/06/2016 11:00	04/06/2016 11:00	-23.62946	-46.7314	39	19
846	3	1	04/06/2016 11:00	04/06/2016 11:03	-23.62946	-46.7314	39	19
847	3	0	04/06/2016 11:06	04/06/2016 11:06	-23.62892	-46.73186	39	19
848	3	1	04/06/2016 11:06	04/06/2016 11:08	-23.62892	-46.73186	39	19
849	3	0	04/06/2016 11:11	04/06/2016 11:11	-23.62916	-46.73143	39	19
850	3	1	04/06/2016 11:11	04/06/2016 11:14	-23.62916	-46.73143	39	19
851	3	0	04/06/2016 11:16	04/06/2016 11:16	-23.62922	-46.73143	39	19
852	3	1	04/06/2016 11:16	04/06/2016 11:18	-23.62922	-46.73143	39	19
853	3	1	04/06/2016 11:16	04/06/2016 11:20	-23.62922	-46.73143	39	19
854	3	1	04/06/2016 11:16	04/06/2016 11:22	-23.62922	-46.73143	39	19

Id_GPS	Id Pessoa	Dados Artificiais	Date Original	Date Interpolado	Lat	Long	Atividade	POI
...	...	...	...	...	...	...	...	...
888	3	0	04/06/2016 12:33	04/06/2016 12:33	-23.61969	-46.73062	40	0
889	3	1	04/06/2016 12:33	04/06/2016 12:35	-23.61969	-46.73062	40	0
890	3	1	04/06/2016 12:33	04/06/2016 12:37	-23.61969	-46.73062	40	0
891	3	1	04/06/2016 12:33	04/06/2016 12:39	-23.61969	-46.73062	40	0
892	3	1	04/06/2016 12:33	04/06/2016 12:41	-23.61969	-46.73062	40	0

- **Id_GPS:** Código de identificação sequencial dos pontos de GPS originais e interpolados.
- **Id_Pessoa:** Código de identificação do indivíduo.
- **Dados_Artificiais:** Indicação se é um dado interpolado ou original, sendo igual a 1 para dados interpolados e igual a 0 para dados originais.
- **Date_Original:** Dia e horário do registro de localização original.
- **Date_Interpolado:** Dia e horário do registro de localização interpolado
- **Lat:** Latitude da localização.
- **Long:** Longitude da localização.
- **Atividade:** Identificação da Atividade do indivíduo a qual o registro do GPS está associado.
- **POI:** Identificação do POI no qual a atividade ocorreu.

No exemplo da Tabela 9, pode-se observar que o campo *Date_Interpolado* é igual ao campo *Date_Original* quando o dado é original, ou seja, *Dados_Artificiais* = 0. Para os dados interpolados (*Dados_Artificiais* = 1), observa-se a diferença de 2 ou 3 minutos entre os horários consecutivos do campo *Date_Interpolado*, devido à aproximação dos segundos. Os valores de *Date_Original* para os dados interpolados são iguais ao do último dado original, tal como as coordenadas *Lat* e *Long*, repetidas para todos os dados interpolados até a ocorrência de um novo dado original.

Os registros indicam que o indivíduo 3 tinha realizado 38 atividades em 18 POI's. Esse indivíduo estava em deslocamento até 10:50, do dia 04/06/2016. Às 10:55 ele inicia a Atividade 39 no POI 19, que dura, no mínimo, até às 11:16. Às 12:33 ele inicia a Atividade 40 no POI 0, o primeiro a ser identificado pelo algoritmo, o que evidencia a recorrência de atividades nesse POI.

As estatísticas gerais para a amostra de 105 pessoas são apresentadas na Tabela 10. O número de atividades se refere a todas as atividades exceto viagens (-1), já que a abordagem adotada neste estudo se baseia apenas nas atividades realizadas. Foram identificados mais 4 indivíduos que realizaram apenas 1 atividade cada, desconsideradas para a caracterização das atividades.

Tabela 10. Número de registros de GPS, Atividades e POI's para a amostra de 105 indivíduos

Estatística	Valores
Amostra	105
Número de registros de GPS originais	39.746
Número médio de registros de GPS originais por indivíduo por dia	54,0
Número de registros de GPS originais + interpolados	473.816
Número de Atividades	1.600
Número de POI's	668

O próximo passo consistiu em converter o banco de dados de registros de GPS em um banco de dados de atividades, a fim de calcular as variáveis a serem usadas no processo de agrupamento. Cada uma das 1.600 atividades identificadas se tornou uma observação no banco de dados de atividades.

O banco foi ordenado por indivíduo; para cada indivíduo as atividades foram organizadas em ordem temporal crescente. Foi necessário excluir a primeira e a última atividades de cada indivíduo para prosseguimento das análises, pelos seguintes motivos: i) a primeira atividade já estava em andamento no momento de início da coleta de dados, não sendo possível identificar o seu horário de início real; e ii) a coleta de dados foi finalizada durante a realização da última atividade, não sendo possível identificar o seu horário de fim real. Conforme apresentado na Tabela 11, pode-se observar que 90% das primeiras atividades dos indivíduos aconteceram no domicílio, bem como 86% das últimas atividades.

Tabela 11. Distribuição das primeiras e últimas atividades entre residências e demais POI's

POI	Primeira Atividade	%	Última Atividade	%
Residência	94	90%	90	86%
Outro	11	10%	15	14%
Total	105	100%	105	100%

Além disso, foi observado um caso de duas atividades consecutivas, identificadas como diferentes pelo algoritmo, mas que ocorreram no mesmo POI. Sendo assim, foram agrupadas em uma única atividade. Com esses dois procedimentos, as 1.600 atividades foram reduzidas para 1.389 [$1600 - (105 \times 2) - 1 = 1389$].

A Tabela 12 apresenta as estatísticas finais para a amostra de 105 indivíduos, referentes às atividades e POI's considerados para a etapa de clusterização.

Tabela 12. Número de Atividades e POI's finais para a amostra de 105 indivíduos

Estatísticas finais	Valores
Número de Atividades	1.389
Número de POI's	649
Número médio de atividades por indivíduo por dia	1,9
Número médio de POI's por indivíduo no período	6,2

4.4. 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, etc.). Neste estudo, uma abordagem multidimensional é usada, incluindo todos os atributos disponíveis e relevantes para a caracterização de uma atividade. Diferentemente da tradicional classificação de atividades por motivo, aqui diversas variáveis foram usadas para compor o perfil de uma atividade, sendo nenhuma delas diretamente relacionada ao propósito de realização das atividades, embora todas possuam aspectos que contribuem para essa inferência.

Os atributos considerados podem ser divididos em quatro categorias: Campos de Identificação, Variáveis Espaciais, Variáveis Temporais e Variáveis de Repetição e Sequência. Eles são apresentados na Tabela 13 e descritos a seguir:

Tabela 13. Atributos das atividades

Campos de Identificação	Descrição
<i>Id_Pessoa_Atividade</i>	Código de identificação do indivíduo e ordem da atividade na sequência de atividades do indivíduo (chave de identificação).
<i>POI</i>	Identificação do POI no qual a atividade ocorreu.
<i>Lat_POI</i>	Latitude do POI no qual a atividade ocorreu (sistema de coordenadas geográficas WGS84).
<i>Lon_POI</i>	Longitude do POI no qual a atividade ocorreu (sistema de coordenadas geográficas WGS84).
<i>POI_Resid</i>	Identificação do POI da residência do indivíduo.

Variáveis Espaciais	Descrição
<i>Local</i>	Indicação se a atividade ocorreu na residência ou em outro local, dentro ou fora de Paraisópolis.
<i>Local_Antes</i>	Indicação se a atividade anterior ocorreu na residência, dentro ou fora de Paraisópolis.
<i>Local_Depois</i>	Indicação se a atividade posterior ocorreu na residência, dentro ou fora de Paraisópolis.
<i>Dist_Resid</i>	Distância linear do POI da atividade à residência do indivíduo, em metros.
<i>Dist_POI_Antes</i>	Distância linear do POI da atividade ao POI da atividade anterior, em metros.
<i>Dist_POI_Depois</i>	Distância linear do POI da atividade ao POI da atividade posterior, em metros.
<i>Uso_Solo</i>	Uso predominante do solo na quadra do POI da atividade.
Variáveis Temporais	Descrição
<i>Date_Início</i>	Dia do início da atividade.
<i>Date_Fim</i>	Dia do fim da atividade.
<i>Dia_Útil_FDS</i>	Indicação se a atividade ocorreu em um dia útil ou no fim de semana ("fds").
<i>Hora_Início</i>	Horário de início da atividade, admitido como o instante do primeiro registro de GPS associado àquela atividade, no formato HHMM.
<i>Hora_Fim</i>	Horário de fim da atividade, admitido como o instante do último registro de GPS associado àquela atividade, no formato HHMM.
<i>Alocação_Hora_Início</i>	Alocação do horário de início da atividade, no formato HHMM.
<i>Alocação_Hora_Fim</i>	Alocação do horário de fim da atividade, no formato HHMM.
<i>Dur_Curta_Aloc</i>	Duração curta da atividade, calculada pelo método da alocação, no formato HHMM.
<i>Dur_Curta_Arred</i>	Duração curta da atividade, calculada pelo método do arredondamento, no formato HHMM.
<i>Dur_Longa_Arred</i>	Duração longa da atividade, calculada pelo método do arredondamento, no formato HHMM.
<i>Dur_Curta_Arred_Antes</i>	Duração curta da atividade anterior, calculada pelo método do arredondamento, no formato HHMM.
<i>Dur_Longa_Arred_Antes</i>	Duração longa da atividade anterior, calculada pelo método do arredondamento, no formato HHMM.
<i>Dur_Curta_Arred_Depois</i>	Duração curta da atividade posterior, calculada pelo método do arredondamento, no formato HHMM.
<i>Dur_Longa_Arred_Depois</i>	Duração longa da atividade posterior, calculada pelo método do arredondamento, no formato HHMM.
<i>Intervalo_Ativ_Arred</i>	Intervalo de tempo entre o fim da atividade e o início da próxima, calculada pelo método do arredondamento, no formato HHMM.

Variáveis de Repetição e Sequência	Descrição
<i>Freq_POI_Vezes</i>	Número de vezes que o POI da atividade foi visitado pelo indivíduo no período.
<i>Freq_POI_Dias</i>	Número de dias em que o POI da atividade foi visitado pelo indivíduo no período.
<i>Ordem_Atividade</i>	Ordem da atividade do indivíduo no dia.

A seguir são explicitadas as lógicas e fórmulas envolvidas no processamento de cada variável.

POI de Residência

- Variáveis: *POI_Resid*

A identificação do POI referente à residência de cada indivíduo foi realizada em duas etapas. Inicialmente, partiu-se da premissa de que a maior parte das atividades realizadas por cada indivíduo corresponde às atividades realizadas em casa. A variabilidade de locais de atividade externos à residência é maior, já que apesar de trabalharmos, estudarmos, realizarmos compras fora de casa, normalmente voltamos para a casa no fim do dia, além de outros possíveis retornos à residência entre atividades ao longo do dia. Assim, para cada indivíduo, o POI formado pelo maior número de atividades foi considerado a sua residência.

Em seguida, os endereços de moradia declarados no questionário da pesquisa de campo, georreferenciados para verificação visual dos dados, foram comparados individualmente com a localização do POI residencial estimado a partir do procedimento anterior. Com isso, foram corrigidos 4 casos em que os POI's inicialmente atribuídos à residência se referiam, na verdade, a prováveis locais de trabalho ou estudo, que também apresentavam alta recorrência de atividades.

Distância entre POI's

- Variáveis: *Dist_Resid*, *Dist_POI_Antes* e *Dist_POI_Depois*

A variável de distância entre POI's foi calculada com base na latitude e longitude identificadas, calculada como a distância euclidiana.

Local da Atividade

- Variáveis: *Local*, *Local_Antes* e *Local_Depois*

A identificação se o local da atividade ocorreu na residência, dentro ou fora de Paraisópolis foi realizada considerando um *buffer* de 100 metros ao redor dessa área, delimitada a partir dos setores censitários do IBGE. Vale ressaltar que o *buffer* foi utilizado para representar a comunidade e seu entorno próximo, já que os setores do IBGE incluem apenas as quadras de Paraisópolis, excluindo, por exemplo, importantes avenidas ao redor. Além dos comércios e serviços que existem dentro da comunidade, nos quais os moradores costumam realizar uma parcela expressiva de suas atividades, há também estabelecimentos no entorno próximo que estão na mesma escala de proximidade dos anteriores. Por esse motivo, os POI's localizados nessa região foram incluídos na categoria "dentro da comunidade".

Uso do Solo

- Variáveis: *Uso_Solo*

A variável de uso do solo pode auxiliar a inferência sobre os motivos das atividades. A associação do POI ao uso de solo correspondente pode diferenciar, por exemplo, imóveis residenciais, onde normalmente são realizadas atividades domésticas, de imóveis de comércios e serviços, onde são realizadas compras ou atividades de trabalho, ou ainda de escolas, onde são realizadas atividades de estudo (ou trabalho, no caso de professoras e assistentes).

Os dados de uso do solo foram obtidos a partir do portal GeoSampa da prefeitura de São Paulo⁹, podendo ser definidos como:

"Tipo de uso do solo predominante (maior ou igual a 60%) por quadra fiscal. Com a finalidade de análise do espaço urbano, a então Secretaria Municipal de Planejamento (Sempla) estabeleceu uma metodologia de agregação resultante do cruzamento entre os valores USO e PADRÃO atribuídos pelo TPCL (Cadastro Territorial Predial de Conservação e Limpeza), para cada imóvel cadastrado, gerando as 16 tipologias de uso relacionadas na tabela Descrição de Tipologias de Uso do Solo".

As referidas tipologias são apresentadas na Tabela 14, a seguir. Em Paraisópolis, as residências aparecem, em sua maioria, como "Terrenos vagos", acentuando o caráter de assentamento informal da favela, sem registros oficiais sobre os domicílios.

⁹ Disponível em: <http://geosampa.prefeitura.sp.gov.br>. Acessado em fevereiro de 2018.

Tabela 14. Descrição de Tipologias de Uso do Solo.

Valor	Uso do solo predominante por quadra
0	Sem informação
1	Residencial horizontal baixo padrão
2	Residencial horizontal médio/alto padrão
3	Residencial vertical baixo padrão
4	Residencial vertical médio/alto padrão
5	Comércio e serviços
6	Indústria e armazéns
7	Residencial e Comércio/serviços
8	Residencial e Indústria/armazéns
9	Comércio/serviços e Indústria/armazéns
10	Garagens
11	Equipamentos públicos
12	Escolas
13	Terrenos vagos
14	Outros
15	Sem predominâncias

Fonte: GeoSampa, seção “Acessar Metadados”.

Alocação do Horário

- Variáveis: *Hora_Início*, *Hora_Fim*, *Alocação_Hora_Início* e *Alocação_Hora_Fim*

Em função de intervalos de coleta maiores que os programados, não havia precisão de quando os registros de GPS indicavam o início ou término de uma atividade. Com isso, foi necessário definir as variáveis *Hora_Início* e *Hora_Fim*, a partir das quais puderam ser definidas todas as demais variáveis temporais. Essas variáveis são importantes considerando que o período de realização e a duração das atividades podem estar associadas à natureza da atividade.

Devido à incerteza com relação aos horários de início e fim das atividades, optou-se por realizar a alocação do início e fim das atividades em faixas horárias, com a finalidade de padronizar e facilitar a interpretação dos resultados. A alocação foi realizada segundo os critérios:

- HH_(i):00 a HH_(i):29 → Alocação na faixa horária HH_(i); e
- HH_(i):30 a HH_(i):59 → Alocação na faixa horária HH_(i+1).

No exemplo da Tabela 16, pode-se observar que os horários de início 1523 (15h23min), 1624 (16h24min) e 1723 (17h23min) foram alocados nas faixas horárias 1500, 1600 e 1700, respectivamente, enquanto os horários 1634 (16h34min) e 1646 (16h46min) foram alocados na faixa horária 1700.

Duração da Atividade

- Variáveis: *Dur_Curta_Aloc*, *Dur_Curta_Arred*, *Dur_Longa_Arred*, *Dur_Curta_Arred_Antes*, *Dur_Longa_Arred_Antes*, *Dur_Curta_Arred_Depois* e *Dur_Longa_Arred_Depois*

Durante o processo de identificação das atividades com o DBSCAN, foi observado que na maior parte das vezes as atividades eram intercaladas por ruído, ou seja, viagens (*Atividade* = -1). Contudo, havia também casos em que atividades distintas se sucediam sem a ocorrência de ruídos entre elas. Há duas hipóteses acerca de tais ocorrências: as atividades ocorreram em dois POI's bastante próximos, sendo que o intervalo de coleta de 5 minutos foi incapaz de captar o caminho entre eles; ou houve perda de sinal no caminho entre os POI's, caracterizado por grandes intervalos de tempo entre dois registros do GPS consecutivos, o que pode ocorrer em automóveis, passagens subterrâneas como o metrô ou até mesmo em áreas densamente construídas (ver 2.2.2). Em função da incerteza sobre os horários de início e fim das atividades, a duração das atividades foi calculada de duas formas, aqui denominadas de *Duração curta* e *Duração longa*. Os critérios usados para o cálculo foram os seguintes:

Duração curta

- i) O início da atividade ocorre no instante do primeiro registro associado àquela atividade; e
- ii) O fim da atividade ocorre no instante do último registro associado àquela atividade.

Duração longa

- i) O início da atividade ocorre no instante do último registro anterior à atividade; e
- ii) O fim da atividade ocorre no instante do primeiro registro posterior à atividade.

Os conceitos de *Duração curta* e *Duração longa* são ilustrados na Figura 24, na forma de 4 exemplos diferentes observados nos dados:

Exemplo 1: Padrão regular, com dados coletados a alta frequência e atividades intercaladas por ruído; pouca diferença entre a duração curta e a duração longa.

Exemplo 2: Padrão regular, com dados coletados a alta frequência, mas são identificadas atividades consecutivas sem ruído entre elas, sendo provavelmente o caso de atividades realizadas em POI's próximos; pouca diferença entre a duração curta e a duração longa.

Exemplo 3: Padrão irregular, a falta de sinal gera “buracos” no banco de dados, entre duas atividades; a duração curta difere bastante da duração longa, que cobre o intervalo de tempo sem coleta de dados.

Exemplo 4: Padrão irregular, a falta de sinal gera “buracos” no banco de dados durante o decorrer de uma atividade; pouca diferença entre a duração curta e a duração longa, pois ambas cobrem o intervalo de tempo sem coleta de dados. Apesar disso, a situação é menos problemática que a anterior, pois pode-se assumir que o indivíduo permaneceu no mesmo local nesse período, embora, dependendo do intervalo, seria possível a intercalação de uma atividade em outro local.

A fim de avaliar a sua influência, ambas as durações calculadas foram usadas no processo de clusterização de atividades. Se as duas variáveis apresentarem comportamento semelhante na formação dos clusters, entende-se que não há diferenças importantes.

Figura 24. Ilustração dos conceitos de *Duração Curta* e *Duração Longa* das atividades. Fonte: Elaboração própria.

Com o objetivo de avaliar o impacto das formas de representação do horário no cálculo da duração das atividades, esta foi calculada por dois métodos: Método da Alocação, no qual a duração é calculada pela diferença dos horários de início e fim alocados em faixas horárias; e Método do Arredondamento, no qual a duração é calculada pela diferença entre o primeiro e o último registro de GPS associado à atividade, e arredondada para a faixa horária superior.

O Método da Alocação foi aplicado apenas para a *Duração curta*, visando evitar a multiplicidade de variáveis que se referem a um mesmo atributo; o Método do Arredondamento foi aplicado para a *Duração curta* e *Duração longa*. As fórmulas são apresentadas a seguir:

- i) Método da Alocação:

$$Duração_Curta_Aloc = Alocação_Hora_Fim - Alocação_Hora_Início$$

- ii) Método do Arredondamento:

$$Duração_Curta_Arred = Hora_Fim - Hora_Início$$

$$Duração_Longa_Arred = \text{Instante do primeiro registro posterior à atividade} - \text{Instante do último registro anterior à atividade}$$

(A partir de HH_(i):01, arredondadas para a faixa horária superior HH_(i+1))

Segundo o Método da Alocação, é possível uma duração resultar igual a 0, no caso de atividades curtas, se os horários de início e fim foram alocados na mesma faixa horária, como aconteceu com a atividade 3 da pessoa 10, apresentada no exemplo da Tabela 16. As durações curtas calculadas pelos dois métodos resultaram iguais ou diferiram em no máximo 1 hora, como é o caso das atividades 2, 3, 7, 8, 9, 11 e 13 do mesmo exemplo, sendo a duração alocada sempre menor que a duração arredondada.

Quanto às durações das atividades que ocorreram antes e depois, pode-se observar que não há registro das variáveis *Dur_Curta_Arred_Antes* e *Dur_Longa_Arred_Antes* para a primeira atividade da sequência (2), enquanto não há registro das variáveis *Dur_Curta_Arred_Depois* e *Dur_Longa_Arred_Depois* para a última atividade da sequência (13). Em função da exclusão das primeiras e últimas atividades, não foi possível realizar o cálculo de durações que dependessem dessas atividades.

Intervalo entre atividades

- Variáveis: *Intervalo_Ativ_Arred*

O intervalo entre atividades pode retratar falhas de sinal ocorridas nesse intervalo, como representado no exemplo 3 da Figura 24, ou as viagens realizadas entre atividades, aqui

representadas por ruídos (exemplos 1 e 4 da Figura 24). Essa variável foi calculada pelo Método do Arredondamento, considerando o intervalo entre o final da atividade analisada e o início da atividade posterior, segundo a fórmula:

$$\text{Intervalo_Ativ_Arred} = \text{Hora_Início} (\text{Atividade}_{(i+1)}) - \text{Hora_Fim} (\text{Atividade}_{(i)})$$

Frequência do POI

- Variáveis: *Freq_POI Vezes* e *Freq_POI Dias*

A frequência de visitas a um POI pode expressar a repetição ou a variação espacial do comportamento de atividades do indivíduo. Neste estudo, foram calculadas duas variáveis de repetição: o número de vezes que o POI em que ocorreu a atividade foi visitado na semana, e o número de dias nos quais isso aconteceu.

As tabelas a seguir (Tabela 15, Tabela 16 e Tabela 17) apresentam as variáveis descritas acima para o indivíduo de *IdPessoa*=10, exemplificando o banco de dados de atividades usado para a próxima etapa, o agrupamento das atividades.

Tabela 15. Exemplo dos Campos de Identificação e Variáveis Espaciais para as atividades de um indivíduo

Campos de Identificação					Variáveis Espaciais						
<i>Id_Pessoa Atividade</i>	<i>POI</i>	<i>Lat_POI</i>	<i>Lon_POI</i>	<i>POI Resid</i>	<i>Local</i>	<i>Dist Resid</i>	<i>Local Antes</i>	<i>Local Depois</i>	<i>Dist_POI Antes</i>	<i>Dist_POI Depois</i>	<i>Uso Solo</i>
10_2	1	-23.59812	-46.73050	0	fora	2082	residência	dentro	2082	2124	4
10_3	3	-23.61535	-46.72135	0	dentro	146	fora	residência	2124	146	7
10_4	0	-23.61442	-46.72034	0	residência	0	dentro	fora	146	2082	13
10_5	1	-23.59812	-46.73050	0	fora	2082	residência	residência	2082	2082	4
10_6	0	-23.61442	-46.72034	0	residência	0	fora	fora	2082	3477	13
10_7	9	-23.64339	-46.73346	0	fora	3477	residência	residência	3477	3477	1
10_8	0	-23.61442	-46.72034	0	residência	0	fora	fora	3477	2082	13
10_9	1	-23.59812	-46.73050	0	fora	2082	residência	residência	2082	2082	4
10_10	0	-23.61442	-46.72034	0	residência	0	fora	fora	2082	2082	13
10_11	1	-23.59812	-46.73050	0	fora	2082	residência	residência	2082	2082	4
10_12	0	-23.61442	-46.72034	0	residência	0	fora	fora	2082	2082	13
10_13	1	-23.59812	-46.73050	0	fora	2082	residência	residência	2082	2082	4

Tabela 16. Exemplo das Variáveis Temporais para as atividades de um indivíduo (continua)

Identificação	Variáveis Temporais									
<i>Id_Pessoa Atividade</i>	<i>Date_Início</i>	<i>Date_Fim</i>	<i>Dia_Útil FDS</i>	<i>Hora Início</i>	<i>Alocação Hora_Início</i>	<i>Hora Fim</i>	<i>Alocação Hora_Fim</i>	<i>Duração Curta_Aloc</i>	<i>Duração Curta_Arred</i>	<i>Duração Longa_Arred</i>
10_2	09/06/2016	09/06/2016	útil	0754	0800	1523	1500	0700	0800	0800
10_3	09/06/2016	09/06/2016	útil	1634	1700	1716	1700	0000	0100	0100
10_4	09/06/2016	10/06/2016	útil	1723	1700	0647	0700	1400	1400	1500
10_5	10/06/2016	10/06/2016	útil	0803	0800	1530	1600	0800	0800	0800
10_6	10/06/2016	10/06/2016	útil	1646	1700	1944	2000	0300	0300	0400
10_7	10/06/2016	10/06/2016	útil	2035	2100	2215	2200	0100	0200	1400
10_8	11/06/2016	13/06/2016	fds	0938	1000	0124	0100	3900	4000	5800
10_9	13/06/2016	13/06/2016	útil	0801	0800	1422	1400	0600	0700	0700
10_10	13/06/2016	14/06/2016	útil	1523	1500	0358	0400	1300	1300	1600
10_11	14/06/2016	14/06/2016	útil	0754	0800	1522	1500	0700	0800	0800
10_12	14/06/2016	15/06/2016	útil	1624	1600	0332	0400	1200	1200	1500
10_13	15/06/2016	15/06/2016	útil	0755	0800	1518	1500	0700	0800	0800

Tabela 17. Exemplo das Variáveis Temporais (continuação) e Variáveis de Repetição e Sequência para as atividades de um indivíduo

Identificação	Variáveis Temporais					Variáveis de Repetição e Sequência		
<i>Id_Pessoa</i> <i>Atividade</i>	<i>Dur_Curta</i> <i>Arred Antes</i>	<i>Dur_Longa</i> <i>Arred Antes</i>	<i>Dur_Curta</i> <i>Arred Depois</i>	<i>Dur_Longa</i> <i>Arred Depois</i>	<i>Intervalo</i> <i>Ativ_Arred</i>	<i>Freq_POI</i> <i>Veze</i>	<i>Freq_POI</i> <i>Dias</i>	<i>Ordem</i> <i>Atividade</i>
10_2	-	-	0100	0100	0200	5	5	1
10_3	0800	0800	1400	1500	0100	1	1	2
10_4	0100	0100	0800	0800	0200	5	5	3
10_5	1400	1500	0300	0400	0200	5	5	1
10_6	0800	0800	0200	1400	0100	5	5	2
10_7	0300	0400	4000	5800	1200	1	1	3
10_8	0200	1400	0700	0700	0700	5	5	1
10_9	4000	5800	1300	1600	0200	5	5	1
10_10	0700	0700	0800	0800	0400	5	5	2
10_11	1300	1600	1200	1500	0200	5	5	1
10_12	0800	0800	0800	0800	0500	5	5	2
10_13	1200	1500	-	-	0300	5	5	1

5. Padrões de Atividades em Múltiplos Dias

Este capítulo apresenta a metodologia usada para a classificação de atividades e indivíduos e discute os resultados de sua aplicação, desenvolvendo as etapas do método destacadas na Figura 25. Inicialmente o item 5.1 apresenta os métodos de agrupamento considerados neste trabalho; o item 5.2 descreve o processo de classificação das atividades, enquanto o item 5.3 descreve o processo de classificação de padrões de atividades dos indivíduos. O capítulo termina com a análise das características socioeconômicas dos grupos de indivíduos, realizada no item 5.4.

Figura 25. Etapas do método desenvolvidas no Capítulo 5. Fonte: Elaboração própria.

5.1. 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.

As técnicas de clusterização podem ser classificadas em *hierárquicas* ou *não hierárquicas*. Os esquemas hierárquicos permitem a identificação do ordenamento e da alocação das

observações, sendo possível avaliar a quantidade de clusters formados. Nos esquemas não hierárquicos parte-se de um número predefinido de clusters, a partir do qual as observações são alocadas, com posterior avaliação da representatividade de cada variável para a formação deles. Portanto, o resultado de um método pode servir de entrada para a realização do outro, tornando a análise cíclica, conforme aponta o esquema da Figura 26.

Figura 26. Lógica para elaboração da análise de agrupamentos. Fonte: Favero e Fávero, 2017.

A alocação das observações em grupos é feita a partir de *medidas de dissimilaridade* entre as variáveis originais (medidas de distâncias), no caso de variáveis métricas, ou de *medidas de similaridade* (medidas de semelhanças), no caso de variáveis binárias. A formação de clusters é bastante sensível à presença de *outliers*, sendo que observações que apresentam valores das variáveis muito distantes ou distintos das demais provavelmente irão formar clusters pequenos.

Medidas de distância (*dissimilaridade*)

Existem diversas medidas de distâncias que podem ser usadas em análises de agrupamentos, sendo a distância linear euclidiana a mais usada. Para um banco de dados com n observações, no qual cada observação i ($i = 1, 2, \dots, n$) apresenta valores correspondentes a cada uma das j ($j = 1, \dots, k$) variáveis métricas X_{ji} , a definição geral da distância euclidiana entre duas observações quaisquer p e q será:

$$d_{pq} = \sqrt{(X_{1p} - X_{1q})^2 + (X_{2p} - X_{2q})^2 + \dots + (X_{kp} - X_{kq})^2} = \sqrt{\sum_{j=1}^k (X_{jp} - X_{jq})^2} \quad (8)$$

Neste estudo, além da distância euclidiana, foi utilizada também a distância quadrática euclidiana, recomendada para os casos em que as variáveis apresentam pequena dispersão

de seus valores, fazendo com que o uso da distância euclidiana ao quadrado facilite a interpretação dos resultados e a alocação das observações nos grupos. Sua expressão é dada por:

$$d2_{pq} = (X_{1p} - X_{1q})^2 + (X_{2p} - X_{2q})^2 + \dots + (X_{kp} - X_{kq})^2 = \sum_{j=1}^k (X_{jp} - X_{jq})^2 \quad (9)$$

Medidas de semelhança (*similaridade*)

A medida de semelhança usada neste estudo foi o logaritmo da verossimilhança ou *log-likelihood*, aplicada no método *TwoStep Clustering*, apresentado a seguir.

A distância de log da verossimilhança pode ser usada tanto para variáveis contínuas quanto categóricas. É uma distância baseada em probabilidade; a distância entre dois clusters está relacionada à diminuição do log da verossimilhança quando eles são combinados em um cluster. No cálculo do log da verossimilhança, são consideradas distribuições normais para as variáveis contínuas e distribuições multinomiais para as variáveis categóricas. Também é admitido que as variáveis e observações são independentes umas das outras (IBM, 2011).

Os métodos de agrupamentos ou esquemas de aglomeração utilizados neste estudo são descritos a seguir. Neste estudo, todas as análises de agrupamento foram realizadas por meio do *IBM SPSS Statistics Software*®.

5.1.1. Método Hierárquico

De acordo com a forma que é iniciado o processo, os esquemas de aglomeração hierárquicos podem ser aglomerativos ou divisivos. 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, são mais comumente utilizados os métodos de encadeamento único (*nearest neighbor* ou *single linkage*), completo (*furthest neighbor* ou *complete linkage*) ou médio (*between groups* ou *average linkage*), conforme apresentado na Figura 27.

Figura 27. Esquemas de aglomeração em análise de agrupamentos. Fonte: Favero e Fávero (2017).

O método de encadeamento completo privilegia as maiores distâncias entre observações, para que sejam formados novos clusters a cada estágio de agrupamento, pela incorporação de observações ou grupos. Recomenda-se utilizá-lo para os casos em que não exista considerável afastamento entre as observações, sendo necessário identificar heterogeneidade entre elas. No caso deste estudo, apesar de terem sido testados os outros métodos, a composição dos grupos puderam ser melhor interpretadas com o método de encadeamento completo.

A partir dos estágios de aglomeração e das distâncias entre os agrupamentos formados, é possível construir um gráfico em formato de árvore, denominado *dendrograma*, que resume o processo de aglomeração e explicita a alocação de cada observação em cada cluster. A partir do dendrograma, é possível analisar a magnitude dos saltos de distância para que se defina o número de clusters mais adequado. Por exemplo, um salto com magnitude elevada, em comparação aos demais, pode indicar que determinada observação ou cluster consideravelmente distinto foi incorporado a clusters já formados e melhor estabelecidos, não sendo necessário um próximo estágio de aglomeração.

A Figura 28 apresenta um exemplo de dendrograma; no eixo x estão as observações e no eixo y a medida de distância. Pode-se observar consideráveis saltos de distância em duas etapas, representados pelas ligações de cor cinza: a primeira indica a formação de 4 clusters (azul, verde, vermelho e roxo), enquanto a segunda indica o agrupamento dos 4 clusters em 2 clusters (azul com verde e vermelho com roxo).

Figura 28. Exemplo de dendrograma para o método de encadeamento completo, com distância euclidiana. Fonte: *Minitab Support*¹⁰.

Se as observações apresentarem valores muito próximos para as variáveis, os saltos de distância do dendrograma resultante podem ficar pequenos, dificultando a leitura do gráfico. Nesse caso, recomenda-se o uso da distância quadrática euclidiana, a fim de explicitar os saltos de forma mais nítida, facilitando a identificação dos clusters no dendrograma.

5.1.2. *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).

¹⁰ Disponível em: <https://support.minitab.com/en-us/minitab/18/help-and-how-to/modeling-statistics/multivariate/how-to/cluster-observations/interpret-the-results/all-statistics-and-graphs/dendrogram/>. Acessado em abril de 2018.

A Figura 29 apresenta um exemplo de ilustração do procedimento *k-means*, no qual as observações foram agrupadas em três clusters, representados pelas cores azul, verde e vermelho, sendo os centros dos clusters indicados pelo círculo com x em preto.

Figura 29. Exemplo de ilustração do procedimento *k-means*. Fonte: Maciej Pacula¹¹.

Após a conclusão do procedimento *k-means*, é necessário verificar se os valores das variáveis métricas diferem entre os grupos definidos, ou seja, se a variabilidade entre os clusters é significativamente superior à variabilidade interna de cada cluster. Essa análise pode ser conduzida a partir de uma análise de variância de um fator (ANOVA), sendo as hipóteses nula e alternativa definidas como:

H_0 : a variável em análise apresenta a mesma média em todos os grupos formados; e

H_1 : a variável em análise apresenta média diferente em pelo menos um dos grupos em relação aos demais.

Com base no resultado dos ANOVAS, as variáveis podem ser ordenadas segundo o valor da estatística *F*, que permite identificar as variáveis cujas médias mais diferem entre os grupos.

¹¹ Disponível em: <http://blog.mpacula.com/2011/04/27/k-means-clustering-example-python/>. Acessado em abril de 2018.

5.1.3. *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 (IBM, 2011), o que não é possível com os métodos apresentados anteriormente.

O método também permite realizar a análise de agrupamentos considerando apenas variáveis categóricas ou contínuas independentemente, o que o torna bastante flexível. Ademais, o número de clusters pode ser inserido pelo usuário *a priori* ou definido automaticamente pelo algoritmo.

O *TwoStep Clustering*, conforme indica o próprio nome, é processado em duas etapas (IBM, 2011):

1. 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 clusters dos dados originais que serão usados no lugar dos dados originais, na etapa de agrupamento hierárquico.
2. O segundo passo consiste na aplicação do método hierárquico padrão, descrito no item 5.1.1, com os *pré-clusters* formados na primeira etapa. Como o número de subgrupos é muito menor do que o número de registros originais, os métodos tradicionais de agrupamento podem ser usados de forma eficiente.

Nas configurações do algoritmo, é possível definir o número de *pré-clusters*. Em geral, quanto maior o número de *pré-clusters*, mais preciso o resultado final da segunda etapa.

5.2. Classificação de Atividades

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 serão compostos padrões de atividades individuais.

O método de agrupamentos usado foi o *TwoStep Clustering*, pela possibilidade de se trabalhar com variáveis qualitativas ou categóricas e também pela facilidade de visualização e análise dos resultados. O número de clusters ótimo foi analisado semanticamente a partir dos resultados do algoritmo e também com base no valor da estatística *silhouette*, apresentada no item 4.3.1.

O item 5.2.1 apresenta as variáveis consideradas para a formação dos clusters, o item 5.2.2 descreve a condução do processo de agrupamento das atividades e, finalmente, o item 5.2.3 apresenta e analisa os grupos de atividades resultantes.

5.2.1. Variáveis Consideradas

Neste item são apresentadas as variáveis descritivas das atividades, definidas no item 4.4, usadas como entrada para o processo de clusterização.

O algoritmo do *TwoStep Clustering* permite especificar o nível de medição das variáveis como escala (dados numéricos em um intervalo ou escala de razão), ordinal ou nominal. Os dados nominais e ordinais podem ser uma sequência de caracteres ou numéricos, conforme definições a seguir, segundo o *Guia do Usuário do Sistema Principal do IBM SPSS Statistics 24* (IBM, 2016):

- **Nominal:** Uma variável pode ser tratada como nominal quando seus valores representam categorias sem ranqueamento intrínseco, como dados que expressam nomeações, características.
- **Ordinal:** Uma variável pode ser tratada como ordinal quando seus valores representam categorias com algum ranqueamento intrínseco, como dados que expressam ordem, níveis de classificação ou dados de contagem.
- **Escala:** Uma variável pode ser tratada como escala (ou contínua) quando os seus valores representam categorias ordenadas com uma métrica significativa, de forma que seja possível comparar as medições de distância entre os valores, como número de anos ou renda, por exemplo.

A Tabela 18 apresenta as variáveis usadas, com o nível de medição adotado para cada uma delas. Todas as variáveis numéricas foram categorizadas ou consideradas como dados de contagem, e incluídas como ordinais.

Tabela 18. Variáveis usadas para o agrupamento das Atividades e respectivo nível de medição adotado

Tipo da variável	Variável	Nível de medição
Variáveis Espaciais	<i>Local</i>	Nominal
	<i>Local_Antes</i>	Nominal
	<i>Local_Depois</i>	Nominal
	<i>Dist_Resid</i>	Ordinal
	<i>Dist_POI_Antes</i>	Ordinal
	<i>Dist_POI_Depois</i>	Ordinal
	<i>Uso_Solo</i>	Nominal
Variáveis Temporais	<i>Dia_Útil_FDS</i>	Nominal

Tipo da variável	Variável	Nível de medição
	<i>Alocação_Hora_Início</i>	Ordinal
	<i>Dur_Curta_Aloc</i>	Ordinal
	<i>Dur_Curta_Arred</i>	Ordinal
	<i>Dur_Longa_Arred</i>	Ordinal
	<i>Dur_Curta_Arred_Antes</i>	Ordinal
	<i>Dur_Longa_Arred_Antes</i>	Ordinal
	<i>Dur_Curta_Arred_Depois</i>	Ordinal
	<i>Dur_Longa_Arred_Depois</i>	Ordinal
	<i>Intervalo_Ativ_Arred</i>	Ordinal
Variáveis de Repetição e Sequência	<i>Freq_POI_Vezes</i>	Ordinal
	<i>Freq_POI_Dias</i>	Ordinal
	<i>Ordem_Atividade</i>	Ordinal

Para as variáveis espaciais *Dist_Resid*, *Dist_POI_Antes* e *Dist_POI_Depois*, foram criadas categorias para classificar os dados contínuos (Tabela 19). Para as demais variáveis ordinais cujo formato original já era categórico, foram estabelecidas categorias diferentes das originais, a fim de se reduzir o número de classes e facilitar a análise dos resultados (Tabela 20 e Tabela 21). O mesmo processo foi realizado para a variável nominal de uso do solo, sendo que as 16 categorias originais foram agregadas em 9 categorias (Tabela 22). Destacam-se como exceções as variáveis de repetição e frequência *Freq_POI_Dias* e *Ordem_Atividade*, para as quais foram usadas as categorias originais (7 e 9 categorias, respectivamente).

Tabela 19. Categorias adotadas para as Variáveis Espaciais Ordinais

<i>Dist_Resid,</i>
<i>Dist_POI_Antes,</i>
<i>Dist_POI_Depois</i>
0km
Até 0,3km
0,3km - 0,5km
0,5km - 1km
1km - 3km
3km - 5km
5km - 10km
Acima 10km

Tabela 20. Categorias adotadas para a Variável de Repetição Ordinal

<i>Freq_POI_Vezes</i>
1
2 e 3
4 e 5
6 e 7
8 a 10
Acima 10

Tabela 21. Categorias adotadas para as Variáveis Temporais Ordinais

<i>Alocação_Hora_Início</i>	<i>Dur_Curta_Aloc</i>	<i>Durações Arredondadas</i>	<i>Intervalo_Ativ_Arred</i>
06h30 - 10h29	1h	0 - 1h	Até 1h
10h30 - 12h29	2h	1h01- 2h00	1h - 3h
12h30 - 15h29	3h	2h01- 3h00	3h - 6h
15h30 - 17h29	4h - 6h	3h01 - 6h00	6h - 10h
17h30 - 19h29	7h - 9h	6h01 - 9h00	Acima 10h
19h30 - 23h29	10h - 12h	9h01 - 12h00	
23h30 - 06h29	13h - 15h	12h01 - 15h00	
	16h - 24h	15h01 - 24h00	
	Acima 24h	Acima 24h	

Tabela 22. Categorias adotadas para a Variável Espacial Nominal *Uso_Solo*

Categoria original	Categoria adotada
Sem informação	Sem informação
Residencial horizontal baixo padrão	Residencial BP
Residencial horizontal médio/alto padrão	Residencial MAP
Residencial vertical baixo padrão	Residencial BP
Residencial vertical médio/alto padrão	Residencial MAP
Comércio e serviços	Com/serv/ind
Indústria e armazéns	Com/serv/ind
Residencial e Comércio/serviços	Misto
Residencial e Indústria/armazéns	Misto
Comércio/serviços e Indústria/armazéns	Com/serv/ind
Garagens	Outros
Equipamentos públicos	Eq. públicos
Escolas	Escolas
Terrenos vagos	Terrenos vagos
Outros	Outros
Sem predominâncias	Outros

5.2.2. Processo de Agrupamento das Atividades

A identificação das variáveis relevantes para a formação dos clusters, bem como a escolha do número de clusters para a classificação das atividades foi realizada em um processo de tentativas e análises sucessivas.

Em um primeiro momento, todas as variáveis apresentadas acima foram consideradas e o algoritmo foi configurado para definir o número de clusters de forma automática. O resultado foram apenas 2 clusters, sendo que a partir daí foram realizados diversos testes, variando o número de clusters e variáveis de entrada. A cada etapa os resultados foram analisados a partir dos valores do silhouette, da razão entre o tamanho do maior e do menor clusters formados e de uma interpretação da própria composição dos clusters.

Em diversos trabalhos, principalmente aqueles que utilizam grandes bases de dados, o silhouette é usado como indicador principal para avaliação do melhor número de clusters. Ou seja, varia-se o número de clusters entre uma gama considerável de valores e aquele que apresentar o maior silhouette é considerado o valor ótimo. Contudo, a grande variabilidade do perfil das atividades e principalmente a relativa pequena extensão do banco de dados (apenas 1.389 entradas) deste estudo, fez com que o silhouette resultante do processo de clusterização das atividades não fosse tão expressivo como medida de coesão e separação dos grupos de atividade.

Por esse motivo, fez-se importante a análise de outros aspectos, como a razão entre o número de observações no maior e menor clusters, e principalmente, a análise detalhada dos próprios grupos formados e as atividades que os compõe. Esse processo foi necessário para verificar se faziam sentido em relação ao que se esperava de uma classificação das atividades, com base no conhecimento a respeito do banco de dados e comportamento dos indivíduos em relação às suas atividades.

Os testes conduzidos no processo de agrupamento das atividades podem ser divididos em duas etapas: T1 a T5, nos quais foi realizado o aumento progressivo do número de clusters (Tabela 23); e T6 a T9, nos quais foi realizada a definição das variáveis, o refinamento do número de clusters e a análise detalhada da composição de atividades dos clusters (Tabela 24).

A partir dos 2 clusters sugeridos pelo algoritmo do *TwoStep Clustering*, aumentou-se gradualmente o número de clusters até 7 (T1 a T5). Entre eles, 6 clusters pareceu a melhor opção em termos de coerência dos grupos formados e menor razão entre maior e menor cluster.

Em seguida, foi observado que a presença das variáveis de duração das atividades anteriores e posteriores da atividade analisada (*Dur_Curta_Arred_Antes*, *Dur_Longa_Arred_Antes*, *Dur_Curta_Arred_Depois* e *Dur_Longa_Arred_Depois*) e de intervalo entre atividades (*Intervalo_Ativ_Arred*) eliminava algumas observações, devido ao fato de que o algoritmo não trata *missing values*, causando a redução da amostra. Visto que essa redução representava

uma porcentagem relevante da amostra de atividades (2 atividades por indivíduo: primeiras e últimas), essas variáveis foram retiradas, para evitar a redução da amostra.

Procedeu-se com a análise de uma variável importante na classificação das atividades, a sua duração. A duração da atividade foi calculada de três formas diferentes, conforme explicado no item 4.4, as quais foram incluídas como entradas do algoritmo, uma a uma. No entanto, a inclusão de todas as três variáveis de duração, em conjunto com as demais variáveis, resultou em uma melhor composição dos grupos. Com isso, optou-se por manter as três durações como variáveis de entrada.

Essa configuração de variáveis foi aplicada para a formação de 6 clusters, que tinha apresentado o melhor resultado *a priori* (T6). Prosseguiu-se assim novamente com a variação do número de clusters, aumentando dessa vez para 8, 9 e 10 (T7, T8 e T9, respectivamente). Observou-se nesse momento a importância de se verificar mais detalhadamente as atividades que compunham cada cluster, para decisão da melhor alternativa.

A análise das atividades que compunham os clusters foi sistematizada da seguinte forma: se a maioria das atividades foram realizadas em casa, o cluster era designado como “C”; se dentro de Paraisópolis, mas em outro lugar que não a residência, como “D”; e se fora de Paraisópolis, como “F”, evidenciando a importância da variável *Local* na formação dos clusters.

No cenário de 6 clusters, foram identificados 2 clusters com maioria das atividades realizadas em casa (2C). Por sua vez, nos cenários com 8, 9 e 10 clusters, foram identificados 4 clusters com maioria das atividades realizadas em casa (4C). Quanto aos clusters de atividades realizadas em sua maioria dentro e fora de Paraisópolis, nos cenários com 6 e 8 clusters foram identificadas 2D e 2F, no cenário de 9 clusters, 2D e 3F, enquanto no cenário com 10 clusters, foram identificadas 3D e 3F.

Diante desses resultados, a alternativa T6 (6 clusters), ressaltada em vermelho na Tabela 24, foi a considerada para a classificação das atividades. Portanto, a análise do padrão de atividades dos residentes de Paraisópolis será conduzida considerando 6 grupos diferentes de atividades.

Tabela 23. Variáveis e resultados dos testes conduzidos no processo de agrupamento das atividades

	Variável	Automático	T1	T2	T3	T4	T5
Entrada	Número de clusters	2	3	4	5	6	7
Resultados	Razão maior cluster/ menor cluster	1,39	1,86	2,52	1,93	1,65	2,33
	Silhouette	0,2	0,2	0,2	0,1	0,1	0,1
Variáveis Espaciais	<i>Local</i>	X	X	X	X	X	X
	<i>Dist_Resid</i>	X	X	X	X	X	X
	<i>Local_Antes</i>	X	X	X	X	X	X
	<i>Local_Depois</i>	X	X	X	X	X	X
	<i>Dist_POI_Antes</i>	X	X	X	X	X	X
	<i>Dist_POI_Depois</i>	X	X	X	X	X	X
	<i>Uso_Solo</i>	X	X	X	X	X	X
Variáveis Temporais	<i>Dia_Útil_FDS</i>	X	X	X	X	X	X
	<i>Alocação_Hora_Início</i>	X	X	X	X	X	X
	<i>Dur_Curta_Aloc</i>	X	X	X	X	X	X
	<i>Dur_Curta_Arred</i>	X	X	X	X	X	X
	<i>Dur_Longa_Arred</i>	X	X	X	X	X	X
	<i>Dur_Curta_Arred_Antes</i>	X	X	X	X	X	X
	<i>Dur_Longa_Arred_Antes</i>	X	X	X	X	X	X
	<i>Dur_Curta_Arred_Depois</i>	X	X	X	X	X	X
	<i>Dur_Longa_Arred_Depois</i>	X	X	X	X	X	X
	<i>Intervalo_Ativ_Arred</i>	X	X	X	X	X	X
Variáveis de Repetição e Sequência	<i>Freq_POI_Vezes</i>	X	X	X	X	X	X
	<i>Freq_POI_Dias</i>	X	X	X	X	X	X
	<i>Ordem_Atividade</i>	X	X	X	X	X	X

Tabela 24. Variáveis e resultados dos testes conduzidos na etapa final de agrupamento das atividades

	Variável	T6	T7	T8	T9
Entrada	Número de clusters	6	8	9	10
Resultados	Razão maior cluster/ menor cluster	2,45	4,03	3,10	3,09
	Silhouette	0,1	0,1	0,1	0,1
	Composição dos clusters	2C-2D-2F	4C-2D-2F	4C-2D-3F	4C-3D-3F
Variáveis Espaciais	<i>Local</i>	X	X	X	X
	<i>Dist_Resid</i>	X	X	X	X
	<i>Local_Antes</i>	X	X	X	X
	<i>Local_Depois</i>	X	X	X	X
	<i>Dist_POI_Antes</i>	X	X	X	X
	<i>Dist_POI_Depois</i>	X	X	X	X
	<i>Uso_Solo</i>	X	X	X	X
Variáveis Temporais	<i>Dia_Útil_FDS</i>	X	X	X	X
	<i>Alocação_Hora_Início</i>	X	X	X	X
	<i>Dur_Curta_Aloc</i>	X	X	X	X
	<i>Dur_Curta_Arred</i>	X	X	X	X
	<i>Dur_Longa_Arred</i>	X	X	X	X
	<i>Dur_Curta_Arred_Antes</i>				
	<i>Dur_Longa_Arred_Antes</i>				
	<i>Dur_Curta_Arred_Depois</i>				
	<i>Dur_Longa_Arred_Depois</i>				
	<i>Intervalo_Ativ_Arred</i>				
Variáveis de Repetição e Sequência	<i>Freq_POI_Vezes</i>	X	X	X	X
	<i>Freq_POI_Dias</i>	X	X	X	X
	<i>Ordem_Atividade</i>	X	X	X	X

5.2.3. Análise dos Grupos de Atividades

Após a classificação das atividades, são analisados, nesse item, os grupos formados e sua composição, em relação às características que suas atividades apresentam, ou seja, as variáveis descritivas usadas no processo de clusterização.

Neste item serão analisados os resultados da classificação dos 6 grupos de atividades, pois, conforme será apresentado no item 5.3., foi a usada para o agrupamento dos indivíduos, a fim de reduzir a complexidade de interpretação dos resultados. A Tabela 25 e Figura 30 apresentam a alocação das atividades entre os clusters resultantes.

Tabela 25. Alocação das atividades por cluster

Cluster	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%

Figura 30. Porcentagem de atividades por cluster. Fonte: Elaboração própria.

Na Figura 31 e Figura 32 é apresentada uma comparação entre os clusters, uma das saídas do *software SPSS*. Os círculos indicam as categorias dominantes para cada variável, sendo o tamanho do círculo um indicativo da porcentagem de observações naquela categoria. Apesar de ser representada a categoria mais frequente para cada variável, pode haver outras

categorias próximas quanto ao número de observações, como por exemplo ocorreu nos casos de *Dur_Curta_Aloc* e *Dur_Curta_Arred*

A Figura 31 apresenta os clusters cuja maioria das atividades foi realizada na residência (Clusters 1 e 2) segundo a variável *Local*, enquanto a Figura 32 apresenta os clusters cuja maioria das atividades foi realizada dentro de Paraisópolis mas em outro lugar que não a residência (Clusters 4 e 5) e os clusters cuja maioria das atividades foi realizada fora de Paraisópolis (Clusters 3 e 6).

Figura 31. Clusters de atividades realizadas em sua maioria na residência.

Figura 32. Clusters de atividades realizadas em sua maioria dentro e fora de Paraisópolis.

Da Figura 33 a Figura 46 são apresentadas em barras vermelhas, para cada variável, a distribuição entre as categorias por cluster de atividade. As barras vermelhas mais fracas ao fundo apresentam a distribuição das variáveis para todos os clusters. Com base nesse conjunto de saídas do SPSS, os grupos foram analisados conforme descrito a seguir.

Clusters 1 e 2 - Atividades na residência

Em relação aos clusters de atividades realizadas na residência, pode-se observar que se diferenciam 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. Analisando a Figura 41, Figura 42 e Figura 43, nota-se que o Cluster 2 apresenta distribuição bastante variada segundo os três métodos de cálculo da duração, o que indica que este é um cluster de atividades de durações diversas, entre 1 e 12 horas. Já o Cluster 1 possui apenas atividades bastante longas, com duração a partir de 12 horas, que em alguns casos chegam a ultrapassar 24 horas, indicando situações em que os indivíduos ficaram em casa por períodos longos (ou ao menos deixaram o celular em casa).

Quanto ao horário de início da atividade (Figura 40), 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.

Os Clusters 1 e 2 diferem levemente pela frequência de visitas à residência, sendo que as atividades do Cluster 2, no geral, ocorrem em POI's que recebem mais visitas na semana do que os POI's onde são realizadas as atividades do Cluster 1 (Figura 44 e Figura 45). Por outro lado, as localizações dos POI's onde foram realizadas as atividades anteriores e posteriores às atividades dos dois clusters estão quase igualmente distribuídas entre dentro e fora de Paraisópolis, não apresentado diferença expressiva (Figura 34 e Figura 35).

Uma análise interessante trata do uso do solo dos POI's, referentes às atividades desses clusters, caracterizadas como atividades realizadas na residência: a grande maioria dos registros foram identificados como terrenos vagos (Figura 38). Há alguns poucos domicílios identificados como residencial de baixo padrão, e outros como médio ou alto padrão, que por sua vez podem se referir a residências de famílias com renda mais alta, existentes no entorno de Paraisópolis, responsáveis pela caracterização do uso do solo predominante da quadra.

A maioria das atividades, em ambos os clusters, foram identificadas como a segunda atividade realizada pelos indivíduos no dia (Figura 46), podendo indicar o retorno para casa depois da atividade principal, como trabalho ou estudo, que seria a primeira atividade do dia, ou ainda a volta para o almoço. Foram identificadas também atividades como sendo a terceira ou quarta realizada no dia, o que pode indicar outras atividades que ocorrem ao longo do dia antes do retorno para casa.

Os clusters não se distinguem dos demais quanto ao dia em que as atividades foram realizadas, sendo que a distribuição das atividades dos Clusters 1 e 2 entre dias úteis ou fim de semana é similar à distribuição para todos os clusters (Figura 39).

Clusters 4 e 5 - Atividades dentro de Paraisópolis

Em relação aos clusters de atividades que ocorrem majoritariamente dentro de 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 (Figura 40 a Figura 45). As atividades curtas apresentam duração de até duas horas, enquanto as atividades médias a longas duram de 2 a 6 horas, principalmente.

A maioria dos POI's onde foram realizadas as atividades anteriores às atividades do Cluster 4 correspondem à residência, o que pode representar atividades realizadas pela manhã em Paraisópolis no início do dia ou após o almoço (Figura 34). Para o Cluster 5, no entanto, também foram identificados POI's dentro de Paraisópolis onde foram realizadas as atividades anteriores, o que indica a realização de atividades distintas realizadas em sequência dentro da favela, e fora de Paraisópolis, o que pode representar os indivíduos chegando na favela e atendendo a compromissos na redondeza antes de ir para casa.

Analogamente, essa característica é observada para os POI's onde são realizadas as atividades posteriores (Figura 35); para o Cluster 4 a maioria corresponde à residência, enquanto para o Cluster 5, além de POI's de residência, foram identificados também POI's dentro e fora de Paraisópolis.

Tal como para as atividades realizadas na residência, as atividades realizadas dentro de Paraisópolis também são caracterizadas por uso do solo identificado como terrenos vagos (Figura 38), mas com proporção maior de residencial médio ou alto padrão, possivelmente em decorrência do *buffer* adotado para distinção do limite de Paraisópolis, conforme explicado em 4.4, que caracteriza as atividades como sendo realizadas dentro ou no entorno de Paraisópolis, onde há residências de média e alta renda. Isto pode indicar atividades de

trabalhos realizadas pelos residentes de Paraisópolis, servindo os proprietários dos imóveis da região, como faxineiras, porteiros, jardineiros, cozinheiras, babás.

Clusters 3 e 6 - Atividades fora de Paraisópolis

Tal como os clusters descritos acima, 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. Pode-se observar que são os clusters que apresentam maiores valores de distância entre os POI's onde foram realizadas as atividades e a residência dos indivíduos, com distribuição variada a partir de 1 km, ultrapassando os 10 km (Figura 36).

Comparados com os clusters com atividades principais dentro de Paraisópolis, o Cluster 3 apresenta atividades mais longas que aquelas observadas no Cluster 4, o que pode indicar atividades principais realizadas fora de Paraisópolis, como trabalho ou estudo, com duração entre 6 e 9 horas em sua maioria (Figura 41, Figura 42 e Figura 43). Já as atividades curtas do Cluster 6 assemelham-se às atividades do Cluster 5, com a maioria delas com duração de até 1 hora, mas havendo também atividades com duração de até 3 horas, o que pode representar compras, refeições, tarefas, ou pequenos serviços realizados esporadicamente 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 mais irregular, de menor previsibilidade (Figura 44 e Figura 45).

O horário das atividades desses clusters reforçam essa conclusão, visto que a maioria das atividades do Cluster 3 se iniciaram no início da manhã entre 6h30 e 10h30, enquanto as atividades do Cluster 6 apresentam horário de início bastante diverso, com maioria das atividades se iniciando das 12h30 às 15h30, apesar de ocorrerem ao longo de todo o dia (Figura 40).

Ademais, a maior parte das atividades do Cluster 3 é a primeira atividade realizada no dia (Figura 46), o que reforça a possibilidade de ser uma atividade principal como estudo ou trabalho. Para o Cluster 6, apesar da prevalência de primeiras atividades, há um número considerável de segundas, terceiras e quartas atividades realizadas no dia.

A maioria dos POI's onde foram realizadas as atividades anteriores às atividades do Cluster 3, correspondem à residência (Figura 34), o que pode representar atividades realizadas fora de Paraisópolis logo após os indivíduos deixarem a residência, reforçando a hipótese de

trabalho ou estudo. Para o Cluster 6, as atividades anteriores se dividem principalmente entre residência, podendo indicar atividades curtas realizadas fora de Paraisópolis logo depois dos indivíduos deixarem a residência, mas antes de iniciarem a atividade principal, e fora de Paraisópolis, o que pode representar atividades curtas eventuais, como compras, tarefas ou atividades de lazer, realizadas a partir do local da atividade principal.

Esse padrão se repete para os POI's onde foram realizadas as atividades posteriores (Figura 35); para o Cluster 3 a maioria corresponde à residência, indicando padrão mais regular, enquanto para o Cluster 6, há também registros de atividades realizadas fora da favela.

A partir da análise do uso do solo, é possível observar maioria de comércio e serviços para as atividades do Cluster 6, mas ainda assim alta porcentagem de residencial de médio ou alto padrão e algum uso misto (Figura 38). Para o Cluster 3, por outro lado, a maioria do uso do solo é residencial de médio ou alto padrão, com alguma porcentagem 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 imóveis de alto padrão, como mencionado anteriormente, como faxineiras, jardineiros, pedreiros, etc., ao mesmo tempo que a presença de comércio e serviços pode indicar atividades de trabalho em lojas, escritórios ou outros estabelecimentos comerciais. Por outro lado, a predominância de comércio e serviços para as atividades mais curtas do Cluster 6 nos faz supor a realização de compras, refeições, outras tarefas, ocasionalmente realizadas sem um padrão de visitas regular.

A partir das análises realizadas, pode-se concluir que os grupos de atividades identificados podem ser caracterizados principalmente quanto aos atributos de localização, duração, hora de início e frequência. A nomenclatura adotada, contudo, privilegiou duas dimensões (localização e duração), embora as demais também sejam relevantes, a fim de facilitar a interpretação dessas categorias para o seu uso na classificação dos indivíduos, que será apresentada no próximo item.

Assim, os clusters de atividade, até então identificados por números, serão denominados segundo a localização (residência, dentro ou fora de Paraisópolis) e duração das atividades (curta, curta-média, média-longa ou longa) predominantes, conforme indicado na Tabela 26.

Tabela 26. Identificação dos grupos de atividades

Cluster de atividade	Identificação
Cluster 1	Residência – Longa

Cluster de atividade	Identificação
Cluster 2	Residência – Curta-média
Cluster 3	Fora – Média-longa
Cluster 4	Dentro – Média-longa
Cluster 5	Dentro – Curta
Cluster 6	Fora – Curta

Figura 33. Distribuição da variável *Local* para cada cluster.

Figura 34. Distribuição da variável *Local_Antes* para cada cluster.

Figura 35. Distribuição da variável *Local_Depois* para cada cluster.

Figura 36. Distribuição da variável *Dist_Resid* para cada cluster.

Figura 37. Distribuição da variável *Dist_POI_Depois* para cada cluster.

Cluster de Atividades 1

Cluster de Atividades 2

Cluster de Atividades 3

Cluster de Atividades 4

Cluster de Atividades 5

Cluster de Atividades 6

Figura 38. Distribuição da variável *Uso_Solo* para cada cluster.

Figura 39. Distribuição da variável *Dia_Útil_FDS* para cada cluster.

Figura 40. Distribuição da variável *Alocação_Hora_Início* para cada cluster.

Figura 41. Distribuição da variável *Dur_Curta_Aloc* para cada cluster.

Figura 42. Distribuição da variável *Dur_Curta_Arred* para cada cluster.

Figura 43. Distribuição da variável *Dur_Longa_Arred* para cada cluster.

Figura 44. Distribuição da variável *Freq_POI_Vezes* para cada cluster.

Figura 45. Distribuição da variável *Freq_POI_Dias* para cada cluster.

Figura 46. Distribuição da variável *Ordem_Atividade* para cada cluster.

5.3. Classificação de Indivíduos

Esta seção descreve a metodologia usada para identificar diferentes grupos de indivíduos com padrões de atividades semelhantes entre si. Para tanto, foram utilizadas variáveis descritivas obtidas a partir das 6 categorias de atividades identificadas na seção anterior. Os itens a seguir descrevem as variáveis usadas para a formação dos clusters, o processo de classificação e validação dos resultados, terminando com a análise das características de cada grupo formado.

5.3.1. Variáveis Consideradas

Esse item explica o processo de obtenção das variáveis necessárias para definir uma segmentação dos indivíduos, a partir das categorias de atividades identificadas.

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.

A Tabela 27 apresenta as variáveis consideradas no processo de clusterização dos indivíduos, com identificação do critério usado para a caracterização do padrão de atividades e descrição da variável. Todas as variáveis foram testadas no processo de classificação, mas somente algumas foram mantidas na solução adotada para a formação dos grupos de indivíduos, conforme descrito no item seguinte.

Tabela 27. Variáveis usadas para a clusterização dos Indivíduos

Variável por indivíduo	Critério de caracterização do padrão de atividade	Descrição da variável
<i>NumAtividades</i>	Intensidade	Número total de atividades realizadas no período.
<i>NumClusters</i>	Variação	Número de clusters de atividades em que foi identificada pelo menos uma atividade do indivíduo.
<i>MaxRepClusters</i>	Repetição	Máxima repetição de atividades em um mesmo cluster.
<i>MaxRepClusters_porAtiv</i>	Repetição	$MaxRepClusters / NumAtividades$
<i>6Clusters</i>	Intensidade / Variação / Repetição	Número de atividades classificadas em cada um dos seis clusters.

Variável por indivíduo	Critério de caracterização do padrão de atividade	Descrição da variável
%6Clusters	Variação / Repetição	Porcentagem de atividades por cluster de atividade.
NumPOIs	Intensidade / Variação	Número total de POIs visitados no período.
NumAtividades_porPOI	Repetição	Número de atividades por POI = $\text{NumAtividades} / \text{NumPOIs}$

Pode-se observar que todas as variáveis consideradas são escalares, já que se referem a um número inteiro de atividades, uma razão ou uma porcentagem. As variáveis *6Clusters* e *%6Clusters* na realidade correspondem a 6 variáveis cada, visto que se referem às 6 categorias de atividade identificadas, conforme apresentado no item 5.2.3.

5.3.2. Processo de Classificação

Devido ao fato das variáveis consideradas serem escalares, os três métodos de agrupamento apresentados no item 5.1 puderam ser utilizados, visando a comparação dos resultados e escolha da configuração mais adequada para os grupos de indivíduos. Inicialmente, foi aplicado o método Hierárquico, cuja análise do dendrograma resultante possibilitou estimar a quantidade de clusters, a qual pode ser então utilizada como entrada para aplicação e avaliação dos métodos *K-means* e *TwoStep*.

Método Hierárquico

O método Hierárquico foi utilizado para a análise do número de clusters, já que nesse método não há a necessidade de predefinir um número de clusters como nos demais. Neste caso, não foram realizados testes com o conjunto completo de variáveis, mas sim foram consideradas aquelas que, em função de julgamento *a priori* e de resultados de testes preliminares, foram consideradas como mais determinantes para a caracterização do padrão de atividades individual: *NumAtividades*, *NumClusters*, *MaxRepClusters* e *%6Clusters*.

O dendrograma indicou a formação de 6 ou 7 clusters; o número de 6 clusters foi o escolhido para a análise dos resultados, uma vez que o sétimo cluster seria composto por apenas um indivíduo. Os resultados são apresentados e analisados no Apêndice 1.

Método K-means

O algoritmo K-means é um método de agrupamento simples, bastante utilizado em processos de classificação. Foram consideradas as mesmas variáveis que aquelas para o método Hierárquico: *NumAtividades*, *NumClusters*, *MaxRepClusters* e *%6Clusters*; o número de 6

clusters de indivíduos foi predefinido para entrada no método. Os resultados são apresentados e analisados no Apêndice 1.

TwoStep Clustering

O método *TwoStep* foi usado para realizar testes com diferentes conjuntos de variáveis, como foi feito para a classificação de atividades. A Tabela 28 apresenta as variáveis consideradas em cada um deles. Os resultados dos métodos Hierárquico e K-means foram úteis para a análise dos grupos formados a partir do método *TwoStep*.

Em um primeiro momento, foram incluídas todas as variáveis das categorias de atividades. Também foram predefinidos 6 clusters de indivíduos como entrada (T1), a partir dos resultados obtidos para os métodos Hierárquico e K-means. Contudo, com relação à distribuição das variáveis, não se observaram diferenças expressivas entre os clusters.

Em seguida, foram testadas diversas combinações de variáveis (T2 a T7). Entre elas, a combinação que envolvia as variáveis *NumAtividades*, *NumClusters*, *MaxRepClusters* e *%6Clusters* (T3) demonstrou os melhores resultados, sendo possível identificar padrões de atividades bem definidos. Com isso, foram realizados testes com 7 e 8 clusters (T8 e T9, respectivamente) considerando as mesmas variáveis, para os quais pode-se observar silhouette ligeiramente maior.

Foi possível observar diversos comportamentos comuns aos diferentes testes realizados. No entanto, quando considerados 7 grupos de indivíduos, havia a segmentação de um grupo de pessoas que apresentava um padrão característico de uma atividade regular principal, como trabalho ou estudo, dentro de Paraisópolis, o qual não foi identificado quando considerados 6 grupos de indivíduos. Por outro lado, quando selecionados 8 grupos, um deles não podia ser claramente caracterizado.

Assim, os 7 grupos de indivíduos, formados a partir da configuração de variáveis do T8, destacado em vermelho na Tabela 28, foram selecionados para representação dos padrões de atividades.

Tabela 28. Variáveis e resultados dos testes conduzidos no processo de clusterização dos indivíduos – Método *TwoStep*

	Variável	T1	T2	T3	T4	T5	T6	T7	T8	T9
Entrada	Número de clusters	6	6	6	6	6	6	6	7	8
Resultados	Razão maior cluster/ menor cluster	2,50	2,70	2,70	3,40	5,83	5,20	5,67	2,50	2,67
	Silhouette	0,3	0,3	0,4	0,4	0,4	0,3	0,3	0,4	0,4
Variáveis das categorias de atividades	<i>NumAtividades</i>	X	X	X	X	X	X	X	X	X
	<i>NumClusters</i>	X	X	X	X		X	X	X	X
	<i>MaxRepClusters</i>	X	X	X				X	X	X
	<i>MaxRepClusters_porAtiv</i>	X	X		X		X			
	<i>6Clusters</i>	X					X	X		
	<i>%6Clusters</i>	X	X	X	X	X			X	X
Outras variáveis	<i>NumPOIs</i>						X	X		
	<i>NumAtividades_porPOI</i>						X	X		

A Tabela 29, a seguir, apresenta a análise de variância (ANOVA), para teste da hipótese de que a média de cada variável é igual entre os grupos, segundo os resultados do método *TwoStep*. Pode-se observar que todos os *valores-P* são consideravelmente menores que 0,05 (*valor-P* = 0,000 para todas as variáveis). Portanto, é possível rejeitar a hipótese nula de que a média de cada uma das 9 variáveis é estatisticamente igual para os 7 clusters de indivíduos, justificando o agrupamento dos indivíduos, ao nível de significância de 5%.

Tabela 29. Análise de variância (ANOVA) dos clusters de indivíduos – Método *TwoStep*

Variáveis		Soma dos Quadrados	gl	Erro Quadrático Médio	F	Valor-P
NumAtividades	Entre clusters	3553,764	6	592,294	25,495	0,000
	Intra cluster	2276,750	98	23,232		
	Total	5830,514	104			
NumClusters	Entre clusters	126,666	6	21,111	30,795	0,000
	Intra cluster	67,182	98	0,686		
	Total	193,848	104			
MaxRepClusters	Entre clusters	437,193	6	72,865	15,785	0,000
	Intra cluster	452,369	98	4,616		
	Total	889,562	104			
%Cluster 1 (Residência – Longa)	Entre clusters	0,524	6	0,087	9,015	0,000
	Intra cluster	0,950	98	0,010		
	Total	1,474	104			
%Cluster 2 (Residência – Curta-média)	Entre clusters	1,011	6	0,169	27,967	0,000
	Intra cluster	0,591	98	0,006		
	Total	1,602	104			
%Cluster 3 (Fora – Média- longa)	Entre clusters	2,139	6	0,357	61,157	0,000
	Intra cluster	0,571	98	0,006		
	Total	2,711	104			
%Cluster 4 (Dentro – Média- longa)	Entre clusters	1,483	6	0,247	73,275	0,000
	Intra cluster	0,331	98	0,003		
	Total	1,814	104			
%Cluster 5 (Dentro – Curta)	Entre clusters	5,156	6	0,859	52,403	0,000
	Intra cluster	1,607	98	0,016		
	Total	6,763	104			
%Cluster 6 (Fora – Curta)	Entre clusters	3,185	6	0,531	36,788	0,000
	Intra cluster	1,414	98	0,014		
	Total	4,599	104			

5.3.3. Análise dos Grupos de Indivíduos

Neste item os 7 clusters de indivíduos são analisados em detalhe, segundo a distribuição das variáveis consideradas no processo de classificação. Cada grupo é caracterizado por um conjunto predominante de comportamentos individuais de atividades, descritos na Tabela 30.

Tabela 30. Identificação dos padrões de atividades dos clusters de indivíduos

Cluster de indivíduos	Padrão de atividade	Descrição
Cluster 1	Casa (C)	Indivíduos que realizam a maior parte de suas atividades em suas residências.
Cluster 2	Casa – Fora (CF)	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.
Cluster 3	Atividade Principal – Dentro (PD)	Indivíduos que apresentam um padrão característico de uma atividade regular principal, como trabalho ou estudo, dentro de Paraisópolis.
Cluster 4	Atividade Principal – Dentro – Variadas (PDV)	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.
Cluster 5	Atividade Principal – Fora (PF)	Indivíduos que apresentam um padrão característico de uma atividade regular principal, como trabalho ou estudo, fora de Paraisópolis.
Cluster 6	Atividade Principal – Fora – Variadas (PFV)	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.
Cluster 7	Atividade Principal – Fora – Ativos (PFA)	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 e bastante ativa, com alto número de atividades observadas no período.

A Tabela 31 e Figura 47 apresentam a alocação dos indivíduos entre os clusters resultantes. O maior cluster (6-PFV) possui 25 indivíduos e os menores (3-PD e 5-PF) 10 indivíduos cada.

Tabela 31. Alocação dos indivíduos por cluster

Cluster	Indivíduos	%
Cluster 1-C	11	10,5%
Cluster 2-CF	15	14,3%
Cluster 3-PD	10	9,5%
Cluster 4-PDV	22	21,0%
Cluster 5-PF	10	9,5%

Cluster	Indivíduos	%
Cluster 6-PFV	25	23,8%
Cluster 7-PFA	12	11,4%
Total	105	100%

Figura 47. Porcentagem de indivíduos por cluster. Fonte: Elaboração própria.

Na Tabela 32 são apresentados os valores das medianas e quartis das variáveis para cada cluster de indivíduo e na Figura 48, em seguida, uma comparação visual entre os clusters. Para cada cluster, uma linha colorida (azul ou vermelha) representa o intervalo entre os quartis de 25% e 75%, enquanto o marcador quadrado representa a mediana. A caixa branca representa o intervalo entre os quartis de 25% e 75% para todos os indivíduos, e o traço preto vertical no interior da caixa, a mediana.

No primeiro quadro são comparados os padrões de atividades baseados principalmente na residência (1-C e 2-CF), no segundo quadro, os padrões baseados em atividades realizadas dentro de Paraisópolis (3-PD e 4-PDV), e no terceiro quadro, os padrões baseados em atividades realizadas fora de Paraisópolis (5-PF, 6-PFV e 7-PFA).

Tabela 32. Medianas e quartis das variáveis para cada cluster de indivíduo

Variáveis	Cluster de indivíduos	1-C	2-CF	3-PD	4-PDV	5-PF	6-PFV	7-PFA	Total
	Nº de Indivíduos	11	15	10	22	10	25	12	105
NumAtividades	Mediana	2,1	9,1	16,1	14,0	11,0	15,0	25,0	13,1
	Q25	1,1	5,1	10,1	12,3	8,0	12,0	23,0	7,0
	Q75	3,0	11,0	23,2	17,0	12,7	16,0	26,0	17,1
NumClusters	Mediana	1,0	3,0	4,5	4,0	4,0	5,0	5,0	4,0
	Q25	1,0	2,5	3,0	4,0	3,0	5,0	5,0	3,0
	Q75	2,0	3,5	5,0	5,0	4,0	6,0	6,0	5,0
MaxRepClusters	Mediana	2,0	4,0	7,0	5,5	5,0	5,0	8,5	5,0
	Q25	1,0	3,0	4,8	4,0	3,5	4,0	8,0	3,0
	Q75	2,0	6,5	9,0	6,0	5,0	6,0	12,2	7,0
%Cluster 1 (Residência – Longa)	Mediana	0%	28%	22%	21%	36%	19%	9%	21%
	Q25	0%	21%	9%	14%	34%	13%	6%	10%
	Q75	33%	33%	33%	24%	38%	21%	13%	29%
%Cluster 2 (Residência – Curta-média)	Mediana	0%	0%	24%	22%	0%	20%	24%	14%
	Q25	0%	0%	12%	14%	0%	14%	18%	0%
	Q75	0%	0%	32%	28%	6%	25%	33%	24%
%Cluster 3 (Fora – Média-longa)	Mediana	0%	0%	0%	0%	40%	27%	14%	0%
	Q25	0%	0%	0%	0%	33%	21%	4%	0%
	Q75	0%	0%	0%	0%	45%	40%	20%	25%
%Cluster 4 (Dentro – Média-longa)	Mediana	0%	0%	40%	13%	0%	0%	2%	0%
	Q25	0%	0%	36%	0%	0%	0%	0%	0%
	Q75	0%	0%	49%	18%	0%	5%	5%	10%

Variáveis	Cluster de indivíduos	1-C	2-CF	3-PD	4-PDV	5-PF	6-PFV	7-PFA	Total
	Nº de Indivíduos	11	15	10	22	10	25	12	105
%Cluster 5 (Dentro – Curta)	Mediana	100%	7%	8%	29%	9%	11%	15%	14%
	Q25	67%	0%	0%	16%	0%	6%	7%	7%
	Q75	100%	12%	15%	42%	14%	20%	23%	31%
%Cluster 6 (Fora – Curta)	Mediana	0%	58%	3%	20%	18%	19%	32%	19%
	Q25	0%	47%	0%	13%	2%	8%	26%	7%
	Q75	0%	67%	8%	29%	24%	24%	42%	30%

Figura 48. Comparação entre os clusters de indivíduos. Fonte: Elaboração própria.

Em complemento, a Figura 55 a Figura 63, apresentadas no Apêndice 2, ilustram a distribuição das variáveis por cluster (vermelho forte) e para o conjunto com todos os indivíduos (vermelho fraco). Com base nesse conjunto de resultados, os grupos de indivíduos puderam ser analisados conforme descrito a seguir.

Clusters 1-C e 2-CF

Os indivíduos desses clusters caracterizam-se por realizar a maior parte de suas atividades em sua própria residência. Pode-se observar que a principal diferença entre os dois clusters se dá nas atividades de duração curta: enquanto os indivíduos do Cluster 1-C realizam atividades deste tipo apenas dentro de Paraisópolis, para os indivíduos do Cluster 2-CF elas acontecem principalmente fora de Paraisópolis.

Outra diferença se dá pelo número de atividades realizadas: para os indivíduos do Cluster 1-C, a mediana dessa variável é igual a 2,1 enquanto para os indivíduos do Cluster 2-CF é igual a 9,1. Essa diferença pode ser um indicativo de que o Cluster 1-C representa indivíduos que não saem de casa por algum tipo de restrição física ou porque não trabalham/estudam, ou ainda deixaram o celular em casa durante todo o período da coleta de dados, enquanto o Cluster 2-CF pode representar indivíduos que trabalham em casa ou que não trabalham e não estudam, mas saem frequentemente de Paraisópolis para realizar atividades pontuais de curta duração.

A distribuição da variável *NumClusters* mostra também que os indivíduos do Cluster 2-CF realizam atividades de tipos variados, mais do que os indivíduos do Cluster 1-C, cuja maioria das atividades se concentram em apenas uma ou duas categorias diferentes.

Clusters 3-PD e 4-PDV

Os indivíduos desses clusters caracterizam-se por realizar a maior parte de suas atividades dentro de Paraisópolis, sendo que em ambos podem ser identificadas atividades principais, caracterizadas por uma duração mais longa que as demais. No entanto, diferem entre si pelo padrão de atividades mais ou menos diverso.

Para os indivíduos do Cluster 3-PD, 40% (mediana) de suas atividades são de duração média a longa e realizadas dentro de Paraisópolis, já para os indivíduos do Cluster 4-PDV, esse valor é igual a 13% (mediana). Contudo, os indivíduos do Cluster 4-PDV realizam mais atividades de duração curta dentro de Paraisópolis (mediana de 29% versus 8% para o Cluster 3-PD) e também mais atividades de duração curta fora de Paraisópolis (mediana de 20% versus 3% para o Cluster 3-PD).

Apesar disso, o número de atividades realizadas não é consideravelmente diferente para os dois grupos, sendo que o Cluster 3-PD ainda apresenta mediana maior, além da variação dos quartis ser também mais ampla. Essas análises permitem inferir que os indivíduos do Cluster 3-PD apresentam padrão mais regular casa-trabalho ou casa-estudo, enquanto os indivíduos do Cluster 4-PDV, apesar de provavelmente trabalharem ou estudarem, também realizam outros tipos de atividades dentro e fora de Paraisópolis.

Clusters 5-PF, 6-PFV e 7-PFA

Os indivíduos desses clusters caracterizam-se por realizar a maior parte de suas atividades fora de Paraisópolis, podendo ser identificadas atividades principais com duração mais longa que as demais. No entanto, diferem entre si pelo padrão de atividades mais ou menos diverso e ativo.

Os indivíduos do Cluster 5-PF apresentam 40% (mediana) de suas atividades com duração média a longa e realizadas fora de Paraisópolis, para o Cluster 6-PFV esse valor é igual a 27% e para o Cluster 7-PFA, 14%. Por sua vez, os indivíduos do Cluster 7-PFA realizam mais atividades curtas dentro e fora de Paraisópolis (mediana de 15% e 32%, respectivamente, versus 11% e 19% para o Cluster 6-PFV, e 9% e 18% para o Cluster 5-PF).

Os indivíduos do Cluster 7-PFA se destacam pelo alto número de atividades realizadas: mediana de 25 atividades, com $Q_{25} = 23$ e $Q_{75} = 26$; para os indivíduos do Cluster 6-PFV esses valores são iguais a 15 atividades, com $Q_{25} = 12$ e $Q_{75} = 16$, e para os indivíduos do Cluster 5-PF, iguais a 11 atividades, com $Q_{25} = 8$ e $Q_{75} = 12,7$.

Como esperado, devido ao maior número de atividades, o Cluster 7-PFA também é o que apresenta maior número de atividades em um mesmo cluster (mediana de 8,5, versus 5,0 para os demais clusters). Além disso, a distribuição da variável *NumClusters* revela padrão mais variado de atividades para os clusters 6-PFV e 7-PFA, cujos quartis estão entre 5 e 6 clusters, em comparação ao Cluster 5-PF, cujos quartis estão entre 3 e 4 clusters.

Outra diferença entre os grupos se dá pela porcentagem de atividades realizadas na residência: enquanto para os indivíduos do Cluster 5-PF esse tipo de atividade é caracterizado principalmente por longas durações, para os indivíduos dos clusters 6-PFV e 7-PFA, as atividades realizadas em casa tendem a ser mais curtas. Ou seja, os indivíduos do Cluster 5-PF tendem a passar longos períodos em casa, enquanto os indivíduos dos clusters 6-PFV e 7-PFA devem usá-la para dormir e outras idas rápidas ao longo do dia.

Essas análises permitem inferir que os indivíduos do Cluster 5-PF apresentam padrão mais regular casa-trabalho ou casa-estudo, enquanto os indivíduos do Cluster 6-PFV, apesar de

trabalharem ou estudarem, também realizam outros tipos de atividades dentro e fora de Paraisópolis. Para os três grupos, as atividades curtas que acontecem fora de Paraisópolis podem representar pausa para o almoço ou outras atividades realizadas durante o expediente de trabalho ou estudo, que se aconteceram, acabaram por segmentar a atividade principal em duas ou mais partes (antes e após as atividades curtas), podendo não revelar a duração total dessas atividades principais.

5.4. Caracterização Socioeconômica dos Grupos de Indivíduos

Neste item serão analisadas as características socioeconômicas dos indivíduos de cada grupo identificado. O formulário usado na entrevista domiciliar, que contém as variáveis utilizadas nesta etapa, pode ser consultado no Anexo 1.

Apresenta-se inicialmente uma caracterização geral da amostra inicial (381 indivíduos) e final selecionada (105 indivíduos), sendo apresentado o perfil socioeconômico desses conjuntos de população, a fim de investigar as questões: a amostra selecionada apresenta algum viés importante em relação à amostra inicial? O perfil dos grupos de indivíduos identificados é consideravelmente distinto do perfil das amostras?

A Tabela 33 apresenta as características socioeconômicas da população de Paraisópolis, considerando os dados do Censo 2010, e os valores para as amostras desse estudo, calculados com base na entrevista realizada. Em comparação com os dados do Censo, a amostra é um pouco mais jovem e possui mais mulheres, enquanto a renda média domiciliar não apresenta diferença considerável.

Tabela 33. Comparação entre as características socioeconômicas da população de Paraisópolis e das amostras inicial e final

Características Socioeconômicas	Censo 2010	Amostra inicial (381 indivíduos)	Amostra final (105 indivíduos)
Feminino	50%	60%	65%
Masculino	50%	40%	35%
16 a 30 anos	45%	50%	46%
30 a 40 anos	27%	26%	30%
40 a 50 anos	16%	16%	21%
Acima de 50 anos	11%	8%	4%
Renda média domiciliar* (Salários Mínimos)	2,2	2,1	2,3

Fonte: Censo (IBGE, 2010) e entrevista domiciliar (2016)

* para as amostras, foram considerados no cálculo apenas indivíduos que aceitaram declarar a renda domiciliar (as taxas de recusa foram 34,4% para a amostra inicial e 30,5% para a amostra final).

Entre todas as variáveis descritivas levantadas na entrevista domiciliar, foram selecionadas aquelas consideradas mais relevantes para os propósitos desse estudo. As distribuições das variáveis para os distintos grupos de indivíduos e amostra inicial e final são apresentadas na Tabela 34 a Tabela 38 e nos gráficos da Figura 49 a Figura 52. Os indivíduos de cada grupo são analisados segundo três eixos principais: pessoal e família, estudo e trabalho, e transporte e exercício físico.

Para facilitar a leitura dos resultados, com base nas descrições da Tabela 30, o padrão de atividades dos indivíduos de cada grupo pode ser resumidamente caracterizado como:

1-C: Casa;

2-CF: Casa e atividades fora de Paraisópolis;

3-PD: Atividade principal dentro de Paraisópolis, padrão regular;

4-PDV: Atividade principal dentro de Paraisópolis, padrão variado;

5-PF: Atividade principal fora de Paraisópolis, padrão regular;

6-PFV: Atividade principal fora de Paraisópolis, padrão variado;

7-PFA: Atividade principal fora de Paraisópolis, padrão variado e ativo.

Tabela 34. Variáveis socioeconômicas para cada grupo de indivíduos e amostras - Pessoal e família (Parte 1)

Grupo de indivíduos	Gênero		Filhos		Gênero e filhos			
	Feminino	Masculino	Mora com filhos/enteados	Não mora com filhos/enteados	Feminino com filhos	Feminino sem filhos	Masculino com filhos	Masculino sem filhos
1-C	73%	27%	36%	64%	36%	36%	0%	27%
2-CF	53%	47%	67%	33%	47%	7%	20%	27%
3-PD	70%	30%	40%	60%	30%	40%	10%	20%
4-PDV	73%	27%	64%	36%	55%	18%	9%	18%
5-PF	60%	40%	60%	40%	30%	30%	30%	10%
6-PFV	68%	32%	52%	48%	32%	36%	20%	12%
7-PFA	50%	50%	42%	58%	17%	33%	25%	25%
Amostra final (105)	65%	35%	53%	47%	37%	28%	16%	19%
Amostra inicial (381)	60%	40%	58%	42%	40%	20%	18%	22%

Grupo de indivíduos	Idade					Estado Civil		
	16 a 20	21 a 30	31 a 40	41 a 50	Acima de 50	Solteira/o	Casada/o	Outros
1-C	18%	27%	18%	36%	0%	45%	55%	0%
2-CF	7%	20%	33%	27%	13%	27%	73%	0%
3-PD	30%	40%	0%	30%	0%	70%	30%	0%
4-PDV	27%	36%	23%	9%	5%	68%	32%	0%
5-PF	20%	30%	30%	20%	0%	50%	40%	10%
6-PFV	16%	16%	44%	20%	4%	64%	32%	4%
7-PFA	17%	25%	42%	17%	0%	50%	42%	8%
Amostra final (105)	19%	27%	30%	21%	4%	55%	42%	3%
Amostra inicial (381)	18%	32%	26%	16%	8%	52%	44%	3%

Tabela 35. Variáveis socioeconômicas para cada grupo de indivíduos e amostras - Pessoal e família (Parte 2)

Grupo de indivíduos	Tamanho da Família					Há quantos anos mora em Paraisópolis				
	1	2	3	4	5 ou mais	Até 1	2 a 5	6 a 10	11 a 20	Acima de 20
1-C	9%	36%	0%	36%	18%	9%	0%	9%	36%	45%
2-CF	7%	13%	47%	7%	27%	20%	33%	13%	7%	27%
3-PD	0%	0%	40%	20%	40%	0%	10%	20%	30%	40%
4-PDV	5%	14%	23%	36%	23%	0%	9%	18%	50%	23%
5-PF	10%	30%	40%	20%	0%	40%	20%	0%	30%	10%
6-PFV	20%	20%	28%	24%	8%	4%	4%	16%	36%	40%
7-PFA	8%	25%	42%	25%	0%	8%	0%	0%	42%	50%
Amostra final (105)	10%	19%	30%	25%	16%	10%	10%	12%	34%	33%
Amostra inicial (381)	5%	23%	30%	21%	22%	7%	12%	14%	35%	32%

Grupo de indivíduos	Renda Pessoal						Escolaridade			
	Recusa	Sem renda	Até 1 SM	1 a 2 SM	2 a 3 SM	Acima de 3 SM	Não estudou	Ens. Fundamental	Ens. Médio	Ens. Superior
1-C	27%	27%	9%	0%	27%	9%	1%	34%	57%	8%
2-CF	40%	27%	13%	7%	7%	7%	1%	25%	62%	12%
3-PD	30%	10%	20%	40%	0%	0%	0%	8%	67%	25%
4-PDV	14%	45%	14%	23%	5%	0%	0%	28%	52%	20%
5-PF	20%	0%	10%	60%	10%	0%	0%	40%	60%	0%
6-PFV	4%	4%	8%	60%	12%	12%	0%	14%	77%	9%
7-PFA	50%	0%	0%	25%	8%	17%	0%	30%	50%	20%
Amostra final (105)	23%	18%	10%	32%	10%	7%	7%	27%	60%	7%
Amostra inicial (381)	27%	26%	12%	24%	7%	4%	0%	36%	64%	0%

Tabela 36. Variáveis socioeconômicas para cada grupo de indivíduos e amostras - Estudo e trabalho

Grupo de indivíduos	Estudando atualmente		Trabalhando atualmente		Ocupação					
	Sim	Não	Sim	Não	Não trabalha	Desempregado	Estudante	Autônomo	S/ carteira assinada	C/ carteira assinada
1-C	18%	82%	45%	55%	27%	18%	9%	45%	0%	0%
2-CF	7%	93%	40%	60%	20%	33%	7%	33%	0%	7%
3-PD	0%	100%	80%	20%	10%	10%	0%	10%	40%	30%
4-PDV	36%	64%	45%	55%	9%	18%	27%	9%	14%	23%
5-PF	0%	100%	100%	0%	0%	0%	0%	10%	30%	60%
6-PFV	20%	80%	92%	8%	4%	0%	4%	8%	20%	64%
7-PFA	33%	67%	92%	8%	0%	0%	8%	25%	8%	58%
Amostra final (105)	19%	81%	70%	30%	10%	11%	10%	18%	15%	36%
Amostra inicial (381)	14%	86%	66%	34%	12%	14%	8%	16%	17%	33%

Grupo de indivíduos	Local de Trabalho			Dias de trabalho por semana				
	Não está trabalhando	Em Paraisópolis	Fora de Paraisópolis	Não está trabalhando	1 a 4	5	6	7
1-C	55%	45%	0%	55%	9%	0%	18%	18%
2-CF	60%	40%	0%	60%	0%	7%	7%	27%
3-PD	20%	70%	10%	20%	10%	30%	30%	10%
4-PDV	55%	23%	23%	55%	9%	14%	18%	5%
5-PF	0%	0%	100%	0%	20%	20%	50%	10%
6-PFV	8%	16%	76%	8%	8%	36%	36%	12%
7-PFA	8%	0%	92%	8%	8%	42%	33%	8%
Amostra final (105)	30%	26%	44%	30%	9%	22%	27%	12%
Amostra inicial (381)	34%	22%	44%	34%	12%	22%	21%	10%

Tabela 37. Variáveis socioeconômicas para cada grupo de indivíduos e amostras - Estudo e trabalho e Transporte

Grupo de indivíduos	Horas de trabalho por dia				Período de início no trabalho		
	Não está trabalhando	Até 7 horas	8 horas	Acima de 9 horas	Não está trabalhando	Matutino	Vespertino ou noturno
1-C	55%	9%	9%	27%	55%	36%	9%
2-CF	60%	0%	13%	27%	60%	40%	0%
3-PD	20%	30%	20%	30%	20%	60%	20%
4-PDV	55%	5%	23%	18%	55%	41%	5%
5-PF	0%	0%	70%	30%	0%	70%	30%
6-PFV	8%	8%	52%	32%	8%	68%	24%
7-PFA	8%	17%	58%	17%	8%	75%	17%
Amostra final (105)	30%	9%	35%	26%	30%	55%	14%
Amostra inicial (381)	34%	10%	32%	23%	34%	49%	17%

Grupo de indivíduos	Posse de carro		Posse de moto		Posse de bicicleta	
	Tem carro	Não tem carro	Tem moto	Não tem moto	Tem bicicleta	Não tem bicicleta
1-C	9%	91%	18%	82%	18%	82%
2-CF	40%	60%	7%	93%	7%	93%
3-PD	10%	90%	30%	70%	0%	100%
4-PDV	14%	86%	18%	82%	5%	95%
5-PF	0%	100%	0%	100%	30%	70%
6-PFV	16%	84%	12%	88%	4%	96%
7-PFA	17%	83%	25%	75%	17%	83%
Amostra final (105)	16%	84%	15%	85%	10%	90%
Amostra inicial (381)	19%	81%	12%	88%	8%	92%

Tabela 38. Variáveis socioeconômicas para cada grupo de indivíduos e amostras - Transporte e exercício físico

Grupo de indivíduos	Modos que usa para sair ou fora de Paraisópolis					Prática de exercício físico	
	A pé ou bicicleta	A pé ou bicicleta; coletivo	Transporte coletivo	Transporte coletivo; individual	Transporte individual	Faz exercício físico	Não faz exercício físico
1-C	9%	0%	73%	0%	18%	45%	55%
2-CF	7%	0%	67%	13%	13%	40%	60%
3-PD	0%	0%	60%	20%	20%	30%	70%
4-PDV	0%	5%	64%	18%	14%	23%	77%
5-PF	0%	10%	90%	0%	0%	50%	50%
6-PFV	4%	0%	84%	4%	8%	16%	84%
7-PFA	0%	8%	67%	17%	8%	42%	58%
Amostra final (105)	3%	3%	72%	10%	11%	31%	69%
Amostra inicial (381)	2%	5%	73%	9%	11%	33%	67%

Figura 49. Distribuição das variáveis socioeconômicas para cada grupo de indivíduos e amostras (Parte 1). Fonte: Elaboração própria.

Figura 50. Distribuição das variáveis socioeconômicas para cada grupo de indivíduos e amostras (Parte 2). Fonte: Elaboração própria.

Figura 51. Distribuição das variáveis socioeconômicas para cada grupo de indivíduos e amostras (Parte 3). Fonte: Elaboração própria.

Figura 52. Distribuição das variáveis socioeconômicas para cada grupo de indivíduos e amostras (Parte 4). Fonte: Elaboração própria.

Pessoal e família

Com relação ao gênero dos indivíduos, a distribuição entre homens e mulheres não difere consideravelmente entre clusters, embora possa ser notada uma maior porcentagem de mulheres nos grupos que realizam a maior parte de suas atividades em suas residências (1-C, 73%), ou que apresentam um padrão característico de uma atividade regular principal dentro de Paraisópolis (3-PD e 4-PDV, 70% e 73%, respectivamente).

A presença de filhos em casa aparenta ter influência em um padrão de atividades mais diverso; destaca-se a diferença entre os dois grupos de indivíduos que realizam mais atividades em suas residências: 67% dos indivíduos do grupo 2-CF, cujo padrão de atividades é mais variado, moram com filhos ou enteados, enquanto esse valor é de apenas 36% para os indivíduos do grupo 1-C, que ficam mais em casa. Essa diferença também pode ser observada entre aqueles que apresentam um padrão característico de uma atividade regular principal dentro de Paraisópolis, sendo que 64% do grupo 4-PDV (padrão mais variado) moram com filhos ou enteados, comparado com 40% do grupo 3-PD (padrão menos variado).

Por sua vez, para aqueles que apresentam um padrão característico de uma atividade regular principal fora de Paraisópolis, quanto mais variado e dinâmico o padrão de atividades, menor a presença de filhos ou enteados em casa: 60% para o grupo 5-PF, 52% para o grupo 6-PFV e 42% para o grupo 7-PFA. Isso pode refletir o fato de que indivíduos que trabalham fora e tem filhos passam mais tempo em casa no tempo extra ao trabalho, em relação àqueles que não tem filhos, que devem ter mais flexibilidade para realizar outros tipos de atividades.

Quando os atributos de gênero e filhos são analisados de forma conjunta, o grupo 4-PDV apresenta a maior porcentagem de mulheres que moram com filhos ou enteados (55% dos indivíduos) e também o grupo 2-CF apresenta alta porcentagem (47%), quando comparada à média para a amostra de 105 indivíduos, igual a 37%. A característica comum às mulheres desses dois grupos pode ser que elas se mantêm próximas de casa, mas fazem atividades diversas. No geral, é possível observar mais mulheres com filhos do que homens com filhos.

Quanto ao estado civil, os resultados apontam que o grupo 2-CF contem mais pessoas casadas (73%) que a amostra selecionada (média de 42%). Por outro lado, aqueles que apresentam um padrão característico de uma atividade regular principal dentro de Paraisópolis (3-PD e 4-PDV) são em sua maioria solteiros e também mais jovens que os demais. Os indivíduos dos grupos que ficam em casa são os mais velhos.

O grupo 6-PFV, caracterizado por atividades variadas fora de Paraisópolis, é o que apresenta maior porcentagem de indivíduos que vivem sozinhos (20%). No geral, os indivíduos que

realizam a maioria de suas atividades dentro de Paraisópolis possuem famílias maiores que aqueles que as fazem fora de Paraisópolis.

Pessoas que vivem há mais tempo na favela parecem realizar suas atividades dentro de Paraisópolis em maior proporção que aquelas que se mudaram há menos tempo, talvez por conhecerem melhor a região: 70% do grupo 3-PD e 73% do grupo 4-PDV vivem em Paraisópolis há mais de 11 anos. Por outro lado, é interessante observar que as pessoas que realizam um maior número de atividades fora da favela também moram lá há mais tempo, o que pode significar que conhecem melhor a cidade: 76% do grupo 6-PFV e 92% grupo do 7-PFA vivem em Paraisópolis há mais de 11 anos.

Os grupos que fazem a maior parte de suas atividades em casa ou dentro da favela apresentam as maiores porcentagens de pessoas que declararam não ter renda. Já entre os grupos que fazem a maior parte de suas atividades fora da favela, a maior parte possui renda pessoal de pelo menos 1SM. Em todos os grupos, muitos indivíduos se recusaram a declarar a sua renda; considerando aqueles que responderam, o grupo 6-PFV é o que apresenta a maior renda pessoal, enquanto o grupo 2-CF é o que apresenta a menor.

Em relação ao grau de escolaridade, as pessoas que realizam a maioria das suas atividades fora de Paraisópolis aparentam ter ensino superior, completo ou incompleto, em maior proporção do que os demais grupos. Isso pode estar ligado ao fato de realizarem atividades mais diversas que os demais. Em todos os grupos a maior parte dos indivíduos possuem ensino médio completo ou incompleto, com média de 62% para a amostra final de 105 indivíduos.

Estudo e trabalho

Os estudantes da amostra estão em sua maioria concentrados no grupo 4-PDV: 36% dos indivíduos desse grupo estão estudando atualmente. Também 20% dos indivíduos do grupo 6-PFV, e 33% dos indivíduos do 7-PFA afirmaram estar estudando atualmente. Pode-se observar que, dentro ou fora de Paraisópolis, são indivíduos que fazem atividades variadas.

Por outro lado, ninguém dos grupos de comportamento de atividades mais regular, 3-PD e 5-PF, declarou estar estudando atualmente, sendo que 80% e 100% dos indivíduos desses grupos declararam estar trabalhando, respectivamente. Por sua vez, 92% dos indivíduos dos grupos 6-PFV e 7-PFA estão trabalhando atualmente, sendo que alguns desses afirmaram também estar estudando.

A ocupação que os indivíduos declararam parece estar bastante relacionada ao grupo em que foram classificados. Os grupos de indivíduos que realizam a maior parte de suas atividades

em suas residências (1-C e 2-CF) são compostos principalmente por pessoas que não estão trabalhando ou são autônomos.

Os grupos de indivíduos que apresentam um padrão característico de uma atividade regular principal, dentro ou fora de Paraisópolis (3-PD e 5-PF), são compostos principalmente por trabalhadores com e sem carteira assinada. Os grupos daqueles que apresentam um padrão característico de uma atividade regular principal fora de Paraisópolis, mas que realizam também outros tipos de atividades (6-PFV) e são mais dinâmicos (7-PFA) concentram principalmente pessoas que trabalham com carteira assinada, embora tenham sido identificados alguns trabalhadores sem carteira assinada no grupo 6-PFV e autônomos no grupo 7-PFA.

O grupo 4-PDV parece ser o mais diversificado quanto às ocupações declaradas, incluindo pessoas que não trabalham ou estão desempregadas, estudantes e trabalhadores com e sem carteira assinada, o que pode justificar padrões de atividades mais variados.

No que se refere ao local de trabalho, a maioria dos indivíduos do grupo 3-PD de fato declararam trabalhar em Paraisópolis; todos os indivíduos do grupo 5-PF e a maioria para os grupos 6-PFV e 7-PFA declararam trabalhar fora de Paraisópolis.

O número de dias de trabalho não diferiu consideravelmente entre os grupos, assim como o período de início no trabalho. A maior parte dos indivíduos dos grupos 5-PF, 6-PFV e 7-PFA declarou trabalhar exatas 8 horas por dia, enquanto a maior parte dos que trabalham dos grupos 1-C e 2-CF o fazem por 8 horas ou mais por dia (talvez pelo fato de serem autônomos).

Transporte e exercício físico

No quesito transporte, o grupo 2-CF apresenta a maior taxa de posse de automóvel (40%), o que pode justificar a participação desses indivíduos em atividades fora de Paraisópolis; já o grupo 3-PD é o que possui a maior taxa de posse de moto (30%), o que parece indicar esta modalidade como uma alternativa comum para os deslocamentos internos à favela. Para o grupo 7-PFA, de indivíduos que realizam diversos tipos de atividades e em grande número, a posse de moto é igual a 25% e de bicicleta, 17%. Enquanto ninguém do grupo 5-PF possui moto, 30% dos indivíduos desse grupo possuem bicicleta.

Quanto ao modo de transporte mais usado para sair de Paraisópolis ou fora da favela, destaca-se o uso de ônibus (ou lotação) em todos os grupos: 72% dos 105 indivíduos o utilizam como modo principal para essa finalidade. Os grupos que realizam a maior parte de suas atividades na residência ou dentro de Paraisópolis aparentam usar o transporte

individual em maior proporção que aqueles que realizam a maior parte de suas atividades fora de Paraisópolis, que se demonstraram mais dependentes do transporte coletivo.

Um total de 31% dos 105 indivíduos da amostra final declarou realizar exercícios físicos regularmente. Essa variável se destaca mais para os grupos 1-C, 5-PF e 7-PFA, não aparentando, contudo, ter relação com o padrão de atividades mais ou menos regular dos indivíduos.

A partir das análises realizadas, a Tabela 39 apresenta a caracterização socioeconômica dos grupos de indivíduos, além de resgatar o conteúdo da Tabela 30, sobre a caracterização dos padrões de atividades de cada grupo.

Tabela 39. Identificação dos padrões de atividades e caracterização socioeconômica dos clusters de indivíduos

Cluster de indivíduos	Padrão de atividade	Caracterização Socioeconômica
Cluster 1	Casa (C) Indivíduos que realizam a maior parte de suas atividades em suas residências.	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 (CF) 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.	Grupo cuja 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 (PD) Indivíduos que apresentam um padrão característico de uma atividade regular principal, como trabalho ou estudo, dentro de Paraisópolis.	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 (PDV) 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.	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 (PF) Indivíduos que apresentam um padrão característico de uma atividade regular principal, como trabalho ou estudo, fora de Paraisópolis.	Grupo cuja grande parte de 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

Cluster de indivíduos	Padrão de atividade	Caracterização Socioeconômica
		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 (PFV) 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.	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 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 (PFA) 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 e bastante ativa, com alto número de atividades observadas no período.	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 haver alguns autônomos, que trabalham fora de Paraisópolis também. Em sua maioria, trabalham 5 ou 6 dias por semana, 8 horas por dia. Alguns tem moto ou bicicleta e praticam exercício físico.

Com base nos resultados, é possível notar que as características socioeconômicas dos indivíduos estão associadas ao seu padrão de atividades.

Os indivíduos que realizam a maior parte de suas atividades diárias fora de Paraisópolis são os que possuem maiores faixas de renda, maior grau de escolaridade e tamanhos de família menores.

Filhos influenciam o padrão de atividades de indivíduos que realizam a maior parte de suas atividades em casa e trabalham em Paraisópolis, tornando-o mais variado; por outro lado, para os indivíduos que trabalham fora de Paraisópolis, os que tem filhos apresentam padrão menos variado.

Entre os indivíduos que realizam a maior parte de suas atividades em casa e trabalham em Paraisópolis, as mulheres que tem filhos apresentam um padrão de atividades mais variado que os homens que tem filhos.

Pessoas casadas demonstraram apresentar padrão de atividades mais ligado à residência do que pessoas solteiras. Ademais, pessoas que fazem a maior parte de suas atividades em casa parecem ser mais velhas; pessoas que realizam atividades em sua maioria dentro de Paraisópolis são mais jovens que aquelas que as realizam em sua maioria fora da favela.

Aqueles que moram na região há mais tempo realizam mais atividades dentro e fora de Paraisópolis, estando também associados a padrões de atividades mais variados.

Em relação à escolaridade, indivíduos que cursam ou cursaram ensino superior realizam a maior parte de suas atividades fora de Paraisópolis; um maior grau de escolaridade também está associado a padrões de atividades mais variados. Como esperado, estudantes apresentam comportamento de atividades mais variado do que pessoas que apenas trabalham.

Quanto à ocupação, indivíduos que trabalham com carteira assinada apresentam padrão mais regular, enquanto pessoas que não trabalham, estudantes, autônomos e pessoas sem carteira assinada apresentam padrões mais variados. A maior parte das pessoas que realizam atividades em sua residência não estão trabalhando atualmente ou são autônomos.

Entre os indivíduos que realizam grande parte de suas atividades na residência, a posse de automóvel aparenta ter relação com a realização de atividades fora da favela.

Essa síntese demonstra a coerência dos resultados aqui obtidos com o que a literatura da análise de demanda por transportes vem revelando ao longo do tempo, quanto ao papel das diversas características socioeconômicas e demográficas e sua relação com o comportamento de atividades individuais.

6. Conclusões

O uso de GPS e outras tecnologias para coleta de dados de transporte facilitou a obtenção de dados de comportamento individual em múltiplos dias, superando algumas das limitações das pesquisas convencionais de demanda por transportes. Por outro lado, a posse de dados de múltiplos dias aumenta a complexidade das análises, o que fez surgir a necessidade de definição de novas métricas para descrever padrões de atividades e viagens, que repensem a lógica do comportamento dos indivíduos.

Esse 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*. Foram identificadas algumas barreiras relacionadas à pesquisa de campo, método de amostragem, coleta e processamento dos dados.

Com relação à pesquisa de campo, foi importante a contratação de uma empresa com experiência no local, o que facilitou o acesso à favela e a seus moradores. O engajamento da comunidade durante as fases de planejamento da pesquisa foi importante para abordar as questões associadas ao crime organizado dentro da favela. Por outro lado, a falta de engajamento dos indivíduos que desinstalaram o aplicativo logo após concluir a entrevista pode indicar uma falta de confiança nos propósitos da pesquisa.

A escolha do intervalo de 5 minutos entre dois registros consecutivos de GPS, definidos a fim de preservar a bateria dos *smartphones*, não limitou a possibilidade de identificar atividades e locais de interesse. No entanto, foi necessário um dispendioso procedimento para limpeza dos dados e seleção dos indivíduos para análise, o que acarretou em uma grande redução da amostra. Além disso, devido à falta de sinal de GPS, por vezes os intervalos entre registros foram maiores que 5 minutos, levando a uma perda de precisão da informação sobre os horários das atividades, embora isso não pareça ter afetado o processo de classificação das mesmas.

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.

Em estudos de transporte, a segmentação dos indivíduos é importante para reduzir a complexidade da análise, em particular no caso de uso de dados de comportamento em múltiplos dias. Ademais, os grupos de atividades e indivíduos identificados neste estudo podem ser usados em processos de modelagem da demanda por viagens. Foi constatado que a maior parte dos indivíduos apresenta padrão de atividades variado (aproximadamente

56% dos indivíduos), o que reflete a importância de se considerar outros tipos de padrões de atividades, além dos padrões regulares casa-trabalho, nos modelos de viagens.

Pode-se dizer que o método empregado facilitou a coleta de dados de múltiplos dias, permitindo o conhecimento da localização das pessoas continuamente. Além disso, destaca-se atualmente a crescente posse de *smartphones* entre os mais diversos segmentos da população. Até mesmo em uma favela como Paraisópolis, *smartphones* são comuns e a internet 3G é amplamente disponível.

Com isso, é ressaltada a importância do uso da tecnologia para obter melhores informações sobre os padrões de atividade dos indivíduos, em especial de populações de baixa renda residentes em favelas, 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.

6.1. Recomendações futuras

Diversos passos dos procedimentos de coleta e processamento dos dados merecem maior exploração e a consideração de hipóteses e parâmetros alternativos, em particular nos algoritmos de clusterização utilizados para identificação de atividades e locais de interesse dos indivíduos.

Acredita-se ainda que o maior potencial para pesquisas futuras reside na aplicação de dados de GPS de forma combinada com fontes de dados tradicionais, como as pesquisas origem e destino domiciliares, visando fornecer insumos para modelos de planejamento de transportes.

Em relação ao método de classificação das atividades utilizado neste estudo, pode ser interessante estudar a relação entre os grupos de atividades obtidos e os tipos de atividades usados em pesquisas tradicionais. Seria interessante também reaplicar o método utilizado, considerando apenas um dia de dados dos indivíduos, e comparar o resultado com a classificação de atividades e indivíduos obtida para múltiplos dias, visando analisar se há diferenças significativas entre o comportamento dos indivíduos em um único dia e em múltiplos dias.

7. Referências Bibliográficas

- Ampt, E., Hickman, M. (2015). Workshop synthesis: Survey methods for hard-to-reach groups and modes. *Transportation Research Procedia*, 11, 475-480.
- Anda, C., Erath, A., Fourie, P. J. (2017). Transport modelling in the age of big data. *International Journal of Urban Sciences*, 21(sup1), 19-42.
- Arbex, R. O., Alves, B. B., Giannotti, M. A., 2016. Comparing accessibility in urban slums using smart card and bus gps data. In Transportation Research Board 95th Annual Meeting (No. 16-5614).
- Axhausen, K. W., Zimmermann, A., Schönfelder, S., Rindsfuser, G., Haupt, T. (2002). Observing the rhythms of daily life: A six-week travel diary. *Transportation*, 29(2), 95-124.
- Axhausen, K. W., Löchl, M., Schlich, R., Buhl, T., Widmer, P. (2007). Fatigue in long-duration travel diaries. *Transportation*, 34(2), 143-160.
- Barry, B. (2002). Social Exclusion, Social Isolation, and the Distribution of Income. In: Understanding Social Exclusion (pp. 13-29). Nova Iorque: Oxford University Press.
- Behrens, R., Freedman, M., McGuckin, N. (2009). The challenges of surveying 'hard to reach' groups: Synthesis of a workshop. In *Transport Survey Methods: Keeping up with a Changing World* (pp. 145-152). Emerald Group Publishing Limited.
- Bianco, R. (2018). Smart data e medidas espaço-temporais de acessibilidade de residentes da comunidade de Paraisópolis (São Paulo). Dissertação de mestrado do Departamento de Engenharia de Transportes da Escola Politécnica da Universidade de São Paulo.
- Bolbol, A., Cheng, T., Tsapakis, I., Haworth, J. (2012). Inferring hybrid transportation modes from sparse GPS data using a moving window SVM classification. *Computers, Environment and Urban Systems*, 36(6), 526-537.
- Bricka, S., Bhat, C. (2006). Comparative analysis of global positioning system-based and travel survey-based data. *Transportation Research Record: Journal of the Transportation Research Board*, (1972), 9-20.
- Carvalho, E. T. (2009). Plano de Urbanização da Comunidade Paraisópolis – Primeira Etapa: Resultados no Setor de Transporte Público, São Paulo Transporte S.A.
- Casas, I. (2007). Social exclusion and the disabled: An accessibility approach. *The Professional Geographer*, 59(4), pp. 463-477.
- Casas, I., Horner, M. W., Weber, J. (2009). A comparison of three methods for identifying transport-based exclusion: A case study of children's access to urban opportunities in Erie and Niagara counties, New York. *International Journal of Sustainable Transportation*, 3(4), pp. 227-245.

- Chen, C., Ma, J., Susilo, Y., Liu, Y., Wang, M. (2016). The promises of big data and small data for travel behavior (aka human mobility) analysis. *Transportation research part C: emerging technologies*, 68, 285-299.
- Ciccantell, P. S., Bunker, S. G. (Eds.). (1998). *Space and Transport in the World-system* (No. 191). Greenwood Publishing Group.
- Dharmowijoyo, D. B., Susilo, Y. O., Karlström, A., Adiredja, L. S. (2015). Collecting a multi-dimensional three-weeks household time-use and activity diary in the Bandung Metropolitan Area, Indonesia. *Transportation Research Part A: Policy and Practice*, 80, 231-246.
- Duncan, J. S., Badland, H. M., Schofield, G. (2009). Combining GPS with heart rate monitoring to measure physical activity in children: A feasibility study. *Journal of Science and Medicine in Sport*, 12(5), 583-585.
- Ester, M., Kriegel, H. P., Sander, J., Xu, X., 1996. A density-based algorithm for discovering clusters in large spatial databases with noise. In *Kdd* (Vol. 96, No. 34, pp. 226-231).
- Ewing, R., Cervero, R. (2010). Travel and the built environment: a meta-analysis. *Journal of the American planning association*, 76(3), 265-294.
- Farber, S., Páez, A. (2011). Running to stay in place: the time-use implications of automobile oriented land-use and travel. *Journal of Transport Geography*, 19(4), pp. 782-793.
- Favero, L., Fávero, P. (2017). *Análise de Dados: Técnicas multivariadas exploratórias com SPSS e Stata*. Elsevier Brasil.
- França, E. e Costa, K. P. (2012). Política Municipal de Habitação: Uma Construção Coletiva, Vol. 2 - Do Plano ao Projeto: Novos Bairros e Habitação Social em São Paulo, Prefeitura Municipal de São Paulo, Secretaria de Habitação.
- Golob, T. F., Meurs, H. (1987). A structural model of temporal change in multi-modal travel demand. *Transportation Research Part A: General*, 21(6), 391-400.
- Hanson, S. (1980). Spatial diversification and multipurpose travel: implications for choice theory. *Geographical Analysis*, 12(3), 245-257.
- Hanson, S., Huff, J. O. (1981). Assessing day-to-day variability in complex travel patterns. *Transportation Research Record*, 891, 18-24.
- Hanson, S., Huff, J. (1986). Classification issues in the analysis of complex travel behavior. *Transportation*, 13(3), 271-293.
- Hanson, S., Huff, O. J. (1988). Systematic variability in repetitious travel. *Transportation*, 15(1), 111-135.
- Herfindahl, O. C. (1950). *Concentration in the steel industry* (Doctoral dissertation, Columbia University.).

- Hernandez, D. O., Titheridge, H. (2016). Mobilities of the periphery: Informality, access and social exclusion in the urban fringe in Colombia. *Journal of transport geography*, 55, 152-164.
- Huff, J. O., Hanson, S. (1986). Repetition and variability in urban travel. *Geographical Analysis*, 18(2), 97-114.
- Huff, J. O., Hanson, S. (1990). Measurement of habitual behaviour: Examining systematic variability in repetitive travel. *Developments in Dynamic and Activity-Based Approaches to Travel Analysis*, 229-249.
- IBM (2011). IBM SPSS Statistics 20 Algorithms.
- IBM, (2016). Guia do Usuário do Sistema Principal do IBM SPSS Statistics 24.
- Jiang, S., Ferreira, J., González, M. C. (2012). Clustering daily patterns of human activities in the city. *Data Mining and Knowledge Discovery*, 25(3), 478–510. <https://doi.org/10.1007/s10618-012-0264-z>.
- Joh, C. H., Arentze, T. A., Hofman, F., Timmermans, H. J. P. (2002). Activity-travel pattern 6 similarity: a multidimensional alignment method. *Transportation Research B*, 7(36), 385-403.
- Kang, H., Scott, D. M. (2010). Exploring day-to-day variability in time use for household members. *Transportation Research Part A: Policy and Practice*, 44(8), 609-619.
- Kantar Worldpanel (2016). Disponível em: <https://canaltech.com.br/noticia/smartphones/57-da-populacao-brasileira-possui-smartphone-diz-pesquisa-78051/>. Acesso em 10 de abril de 2017.
- Kitamura, R. (1986). Linear panel analysis of travel behavior. *Netherlands Institute of Transport, Rijswijk, the Netherlands*.
- Kitamura, R. (1987). A panel analysis of household car ownership and mobility. *Doboku Gakkai Ronbunshu*, 1987(383), 13-27.
- Kitamura, R. (1998). An analysis of weekly activity patterns and travel expenditure. *Institute of Transportation Studies*.
- Kitamura, R., Bovy, P. H. (1987). Analysis of attrition biases and trip reporting errors for panel data. *Transportation Research Part A: General*, 21(4-5), 287-302.
- Kitamura, R., Hoorn, T. (1987). Regularity and irreversibility of weekly travel behavior. *Transportation*, 14(3), 227-251.
- Koch, J., Lindau, L. A., Nassi, C. D. (2013). Transportation in the Favelas of Rio de Janeiro. *Lincoln Institute (www.lincolninst.edu)*.
- Kwan, M. P. (1999). Gender and individual access to urban opportunities: A study using space-time measures. *The Professional Geographer*, 51(2), pp. 211-227.
- Lacerda, N. (2007). Intervenções no bairro do Recife e no seu entorno: indagações sobre a sua legitimidade. *Sociedade e Estado*, 22(3).

- Lima, M. L. C. C., Somekh, N. (2013). Análise urbanística e diagnóstico de assentamentos precários: um roteiro metodológico. *Ambiente Construído*, 13(1), 109-127.
- Löchl, M., Axhausen, K. W., Schönfelder, S. (2005, March). Analysing Swiss longitudinal travel data. In *5th Swiss Transport Research Conference, Ascona*.
- Lucas, K. (2012). Transport and social exclusion: Where are we now? *Transport Policy*, 20, pp. 105-113.
- Lucas, K., Porter, G. (2016). Mobilities and livelihoods in urban development contexts: introduction. *Journal of transport geography*, 55, 129-131.
- Maia, M. L., Lucas, K., Marinho, G., Santos, E., de Lima, J. H. (2016). Access to the Brazilian City—from the perspectives of low-income residents in Recife. *Journal of Transport Geography*, 55, 132-141.
- Maia, R.A.I., Orrico Filho, R.D., 2007. Uma Análise para Integração de Veículos de Baixa Capacidade ao Sistema de Transporte Convencional. XIV CLATPU e XXI ANPET, 2007, Rio de Janeiro Congresso Latino-Americano de Transporte Público e Urbano—XIV CLATPU.
- Martell, L. (2016). *The sociology of globalization*. John Wiley & Sons.
- Meyer, J. F., Haddad, E., Tanaka, R. M. H., Silva, A. P., Barbon, A. L., Santos, G. M. Residential Real Estate Market in Paraisópolis: What Has Changed in the Last Ten Years? Faculty of Architecture and Urbanism of the University of São Paulo (FAU/USP) and Lincoln Institute of Land Policy.
- Moiseeva, A., Timmermans, H., Choi, J., Joh, C. H. (2014). Sequence Alignment Analysis of Variability in Activity Travel Patterns Through 8 Weeks of Diary Data. *Transportation Research Record: Journal of the Transportation Research Board*, (2412), 49-56.
- Ortega-Tong, M. A. (2013). Classification of London's public transport users using smart card data, 163.
- Ortúzar, J. D. D., Willumsen, L. G. (2002). Modelling transport. John Wiley & Sons. West Sussex, England.
- Pas, E. I. (1988). Intrapersonal variability and model goodness-of-fit. *Transportation Research Part A: General*, 21(6), 431-438.
- Pas, E. I., Koppelman, F. S. (1987). An examination of the determinants of day-to-day variability in individuals' urban travel behavior. *Transportation*, 13(2), 183-200.
- Pas, E. I., Sundar, S. (1995). Intrapersonal variability in daily urban travel behavior: some additional evidence. *Transportation*, 22(2), 135-150.
- Pas, E. I. (1987). Intrapersonal variability and model goodness-of-fit. *Transportation Research Part A: General*, 21(6), 431-438.

- Pizzol, B., Bianchi, B., Abrantes, M., Strambi, O. (2017). Travel survey tools and methods : challenges on surveys at slums. 11th International Conference on Transport Survey Methods Travel, Quebec, Canada.
- Santos, J. (2012). Disponível em: <http://jornal.paraisopolis.org/maria-teresa-diniz-coordenadora-do-prog-de-urbanizacao-de-paraisopolis-memorias-da-coordenacao/>.
- Schönfelder, S., Axhausen, K. W. (2003). Activity spaces: measures of social exclusion?. *Transport policy*, 10(4), 273-286.
- Schönfelder, S., Antille, N. (2002). *Exploring the potentials of automatically collected GPS data for travel behaviour analysis: A Swedish data source*. ETH, Eidgenössische Technische Hochschule Zürich, Institut für Verkehrsplanung, Transporttechnik, Strassen-und Eisenbahnbau IVT.
- Shapcott, M. (1978). Comparison of the use of time in Reading, England with time use in other countries. *Transactions of the Martin Centre for Architectural and Urban Studies*, 3, 231-257.
- Shen, L., Stopher, P. R. (2014). Review of GPS travel survey and GPS data-processing methods. *Transport Reviews*, 34(3), 316-334.
- Sila-Nowicka, K., Vandrol, J., Oshan, T., Long, J. A., Demšar, U., Fotheringham, A. S. (2016). Analysis of human mobility patterns from GPS trajectories and contextual information. *International Journal of Geographical Information Science*, 30(5), 881-906.
- Stopher, P. R., Metcalf, H. M. (1996). *Methods for household travel surveys* (Vol. 236). Transportation Research Board.
- Stopher, P., FitzGerald, C., Xu, M. (2007). Assessing the accuracy of the Sydney Household Travel Survey with GPS. *Transportation*, 34(6), 723-741.
- Stopher, P., FitzGerald, C., Zhang, J. (2008). Search for a global positioning system device to measure person travel. *Transportation Research Part C: Emerging Technologies*, 16(3), 350-369.
- Susilo, Y. O., Axhausen, K. W. (2014). Repetitions in individual daily activity–travel–location patterns: a study using the Herfindahl–Hirschman Index. *Transportation*, 41(5), 995-1011.
- Susilo, Y. O., Kitamura, R. (2005). Analysis of Day-to-Day Variability in a Individual's Action Space: Exploration of 6-Week Mobidrive Travel Diary Data. *Transportation Research Record: Journal of the Transportation Research Board*. 1902, pp. 124-133. Washington, D.C.: Transportation Research Board of the National Academies.
- Vlassenroot, S., Gillis, D., Bellens, R., Gautama, S. (2015). The use of smartphone applications in the collection of travel behaviour data. *International Journal of Intelligent Transportation Systems Research*, 13(1), 17-27.
- Widhalm, P., Yang, Y., Ulm, M., Athavale, S., González, M. C. (2015). Discovering urban activity patterns in cell phone data. *Transportation*, 42(4), 597-623.

Yamamoto, T., Kitamura, R. (1999). An analysis of time allocation to in-home and out-of-home discretionary activities across working days and non-working days. *Transportation*, 26(2), 231-250.

Apêndices

1. Processo de Classificação dos Indivíduos

Neste Apêndice são apresentados os resultados obtidos para os métodos de agrupamento Hierárquico e K-means, utilizados de forma complementar ao método *TwoStep* para a classificação dos residentes de Paraisópolis segundo o seu padrão de atividades individual.

Método Hierárquico

Para o processo de aglomeração hierárquico foram usados o método de encadeamento completo e a distância quadrática euclidiana. A Figura 53 apresenta o dendrograma resultante. É possível identificar um grande salto de distância, indicando a formação de 6 ou 7 clusters; o número de 6 clusters foi o escolhido para a análise dos resultados, uma vez que o sétimo cluster seria composto por apenas um indivíduo. A Tabela 40 apresenta a estatística descritiva dos clusters de indivíduos, incluindo a média, desvio padrão, valores mínimo e máximo, enquanto nos gráficos da Figura 54, para cada variável, são apresentadas as médias para os diferentes clusters.

Pode-se observar que o Cluster 6 é caracterizado por indivíduos que fizeram poucas atividades no período (média de 1,2), curtas e dentro de Paraisópolis. O Cluster 3 é caracterizado por indivíduos que realizaram pouco mais de 1 atividade por dia, com padrão mais variado, havendo atividades de todos os tipos, mas com predominância de atividades longas na residência, médias ou longas fora de Paraisópolis e curtas dentro de Paraisópolis.

O Cluster 2 se destaca por mais da metade das atividades de seus indivíduos, na média, serem atividades curtas realizadas fora de Paraisópolis. Os demais tipos de atividades que surgem nesse grupo são atividades longas na residência (em média, 23%) e algumas atividades curtas realizadas dentro de Paraisópolis (em média, 9%).

Os Clusters 1 e 5 se aproximam quanto ao número de atividades realizadas e padrão variado de tipos de atividades, mas se diferenciam especificamente quanto às atividades de duração média ou longa: enquanto 16%, em média, das atividades dos indivíduos do Cluster 1 são de duração média ou longa fora de Paraisópolis, 41%, em média, das atividades dos indivíduos do Cluster 5 são de duração média ou longa dentro de Paraisópolis. Além disso, o Cluster 1 também apresenta mais atividades curtas, dentro e fora de Paraisópolis, do que o Cluster 5.

O Cluster 4 aparenta ser o cluster de indivíduos mais ativos, que realizaram em média 30 atividades no período, ou seja, média de 4,3 atividades por dia. Além disso, suas atividades são distribuídas, em média, entre 5 dos 6 clusters, o que os caracteriza também com um padrão de atividades bastante variado. De fato, uma baixa porcentagem das atividades,

apenas 6% em média, são atividades na residência de duração longa, o que indica ser um grupo de indivíduos que não passam tanto tempo em casa, pois também 46% das atividades são de duração curta e fora de Paraisópolis.

Por fim, a análise de variância (ANOVA), apresentada na Tabela 41, testa a hipótese de que a média de cada variável é igual entre os grupos. Pode-se observar que todos os *valores-P* são consideravelmente menores que 0,05 (*valor-P* = 0,000). Portanto, é possível rejeitar a hipótese nula de que, para cada variável, a média para os diferentes clusters são iguais, ao nível de significância de 5%.

Figura 53. Dendrograma para o método de encadeamento completo – Método Hierárquico.

Tabela 40. Estatística descritiva dos clusters de indivíduos – Método Hierárquico

Variáveis	Cluster de indivíduos	1	2	3	4	5	6	Total
	Nº de Indivíduos	51	18	16	4	10	6	105
Num Atividades	Média	15,4	9,9	7,9	30,3	17,0	1,2	13,2
	Desvio padrão	5,5	5,2	4,3	6,7	6,7	0,4	7,5
	Mínimo	4	1	3	26	8	1	1
	Máximo	26	20	14	40	26	2	40
NumClusters	Média	4,8	3,3	3,1	5,0	4,1	1,0	4,0
	Desvio padrão	0,8	1,2	0,8	1,4	1,1	0,0	1,4
	Mínimo	3	1	2	3	2	1	1
	Máximo	6	5	4	6	5	1	6
MaxRep Clusters	Média	5,2	5,3	3,6	13,5	7,2	1,2	5,2
	Desvio padrão	2,0	2,8	1,7	1,3	2,4	0,4	2,9
	Mínimo	1	1	1	12	3	1	1
	Máximo	10	12	6	15	10	2	15
%Cluster 1 (Residência – Longa)	Média	17%	23%	36%	6%	17%	0%	20%
	Desvio padrão	8%	10%	5%	4%	14%	0%	12%
	Mínimo	0%	0%	25%	0%	0%	0%	0%
	Máximo	33%	40%	45%	10%	38%	0%	45%
%Cluster 2 (Resid. – Curta-média)	Média	21%	4%	2%	24%	27%	0%	15%
	Desvio padrão	7%	7%	3%	13%	14%	0%	12%
	Mínimo	7%	0%	0%	12%	7%	0%	0%
	Máximo	38%	21%	9%	38%	46%	0%	46%
%Cluster 3 (Fora – Média-longa)	Média	16%	5%	23%	13%	1%	0%	13%
	Desvio padrão	17%	8%	20%	9%	3%	0%	16%
	Mínimo	0%	0%	0%	0%	0%	0%	0%
	Máximo	55%	25%	46%	19%	11%	0%	55%
%Cluster 4 (Dentro – Média-longa)	Média	6%	2%	2%	4%	41%	0%	8%
	Desvio padrão	8%	5%	8%	3%	10%	0%	13%
	Mínimo	0%	0%	0%	0%	23%	0%	0%
	Máximo	28%	14%	33%	8%	54%	0%	54%
%Cluster 5 (Dentro – Curta)	Média	21%	9%	27%	8%	8%	100%	23%
	Desvio padrão	14%	11%	28%	9%	9%	0%	26%
	Mínimo	0%	0%	0%	0%	0%	100%	0%
	Máximo	50%	40%	75%	20%	27%	100%	100%
%Cluster 6 (Fora – Curta)	Média	18%	57%	10%	46%	6%	0%	22%
	Desvio padrão	11%	16%	12%	11%	6%	0%	21%
	Mínimo	0%	40%	0%	30%	0%	0%	0%
	Máximo	43%	100%	33%	54%	15%	0%	100%

Figura 54. Média das variáveis para cada cluster – Método Hierárquico.

Tabela 41. Análise de variância (ANOVA) dos clusters de indivíduos – Método Hierárquico

Variáveis		Soma dos Quadrados	gl	Erro Quadrático Médio	F	Valor-P
NumAtividades	Entre clusters	3054,892	5	610,978	21,792	0,000
	Intra cluster	2775,622	99	28,037		
	Total	5830,514	104			
NumClusters	Entre clusters	114,453	5	22,891	28,543	0,000
	Intra cluster	79,395	99	0,802		
	Total	193,848	104			
MaxRepClusters	Entre clusters	456,152	5	91,230	20,839	0,000
	Intra cluster	433,410	99	4,378		
	Total	889,562	104			
%Cluster 1 (Residência – Longa)	Entre clusters	0,760	5	0,152	21,087	0,000
	Intra cluster	0,714	99	0,007		
	Total	1,474	104			
%Cluster 2 (Residência – Curta-média)	Entre clusters	1,016	5	0,203	34,334	0,000
	Intra cluster	0,586	99	0,006		
	Total	1,602	104			
%Cluster 3 (Fora – Média-longa)	Entre clusters	0,573	5	0,115	5,304	0,000
	Intra cluster	2,138	99	0,022		
	Total	2,711	104			
%Cluster 4 (Dentro – Média-longa)	Entre clusters	1,289	5	0,258	48,639	0,000
	Intra cluster	0,525	99	0,005		
	Total	1,814	104			
%Cluster 5 (Dentro – Curta)	Entre clusters	4,283	5	0,857	34,205	0,000
	Intra cluster	2,480	99	0,025		
	Total	6,763	104			
%Cluster 6 (Fora – Curta)	Entre clusters	3,320	5	0,664	51,394	0,000
	Intra cluster	1,279	99	0,013		
	Total	4,599	104			

Método K-means

Para o processo de aglomeração com o método K-means, a distância usada foi a euclidiana. A partir do resultado do método hierárquico, o número de clusters de indivíduos foi predefinido como 6 para entrada no método. A Tabela 42 apresenta a estatística descritiva dos 6 clusters de indivíduos, incluindo a média, desvio padrão, valores mínimo e máximo.

Destaca-se o Cluster 1, pelo baixo número de atividades realizadas, sendo que a distribuição entre tipos de atividades indica que são indivíduos que ficaram em casa (12%) e realizaram atividades curtas, dentro de Paraisópolis (88%). Os indivíduos do Cluster 3 realizaram em média 1 atividade por dia, distribuídas entre atividades longas na residência (27%), ou outras atividades curtas, dentro (9%) e fora (57%) de Paraisópolis.

Os Clusters 4, 5 e 6 caracterizam-se por um alto número de atividades realizadas e também notável diversidade dos tipos de atividades. Os indivíduos dos Clusters 4 e 5 se diferenciam em relação às atividades médias ou longas: enquanto 34% das atividades do Cluster 4 são atividades médias ou longas realizadas fora de Paraisópolis, 35% das atividades do Cluster 5 são atividades médias ou longas realizadas dentro de Paraisópolis; já o Cluster 6 se destaca pela alta porcentagem de atividades de curta duração realizadas tanto dentro (30%) quanto fora de Paraisópolis (19%).

Os indivíduos do Cluster 2 aparentam ser os mais ativos, já que realizam em média, 3,5 atividades por dia, e também bastante diversos quanto ao tipo de atividades, pois em média apresentam atividades de 5 tipos diferentes. A maioria de suas atividades são curtas (48%), realizadas dentro (15%) e fora de Paraisópolis (33%), seguidas por atividades de duração média ou longa, realizadas fora da favela (15%).

A Tabela 43 apresenta a análise de variância dos clusters de indivíduos obtidos pelo método K-means. Todos os *valores-P* são consideravelmente menores que 0,05 (*valor-P* = 0,000); com isso, é possível rejeitar a hipótese nula de que, para cada variável, a média para os diferentes clusters são iguais, ao nível de significância de 5%.

Tabela 42. Estatística descritiva dos clusters de indivíduos – Método K-means

Variáveis	Cluster de indivíduos	1	2	3	4	5	6	Total
	Nº de Indivíduos	10	14	16	29	14	22	105
Num Atividades	Média	2,0	24,4	7,9	13,0	18,1	12,3	13,2
	Desvio padrão	1,2	5,8	4,5	4,1	5,6	4,3	7,5
	Mínimo	1	14	1	4	8	3	1
	Máximo	4	40	18	20	26	20	40
NumClusters	Média	1,4	5,1	3,0	4,6	4,3	4,4	4,0
	Desvio padrão	0,5	0,9	1,0	0,9	1,0	0,7	1,4
	Mínimo	1	3	1	3	2	3	1
	Máximo	2	6	5	6	5	6	6
MaxRep Clusters	Média	1,6	9,5	4,6	4,6	7,0	4,5	5,2
	Desvio padrão	0,7	2,8	3,1	1,6	2,0	1,6	2,9
	Mínimo	1	6	1	1	4	1	1
	Máximo	3	15	12	8	10	7	15
%Cluster 1 (Residência – Longa)	Média	12%	9%	27%	23%	15%	24%	20%
	Desvio padrão	16%	6%	11%	11%	11%	7%	12%
	Mínimo	0%	0%	0%	0%	0%	14%	0%
	Máximo	33%	21%	43%	45%	38%	40%	45%
%Cluster 2 (Resid. – Curta-média)	Média	0%	24%	0%	16%	27%	16%	15%
	Desvio padrão	0%	8%	2%	10%	11%	8%	12%
	Mínimo	0%	12%	0%	0%	7%	0%	0%
	Máximo	0%	38%	6%	38%	46%	29%	46%
%Cluster 3 (Fora – Média-longa)	Média	0%	15%	5%	34%	1%	2%	13%
	Desvio padrão	0%	11%	9%	11%	3%	5%	16%
	Mínimo	0%	0%	0%	17%	0%	0%	0%
	Máximo	0%	35%	33%	55%	11%	14%	55%
%Cluster 4 (Dentro – Média-longa)	Média	0%	3%	1%	2%	35%	9%	8%
	Desvio padrão	0%	4%	4%	4%	13%	10%	13%
	Mínimo	0%	0%	0%	0%	12%	0%	0%
	Máximo	0%	10%	11%	14%	54%	33%	54%
%Cluster 5 (Dentro – Curta)	Média	88%	15%	9%	11%	13%	30%	23%
	Desvio padrão	16%	14%	11%	9%	14%	14%	26%
	Mínimo	67%	0%	0%	0%	0%	7%	0%
	Máximo	100%	45%	40%	29%	47%	50%	100%
%Cluster 6 (Fora – Curta)	Média	0%	33%	57%	15%	9%	19%	22%
	Desvio padrão	0%	15%	18%	9%	10%	14%	21%
	Mínimo	0%	10%	29%	0%	0%	0%	0%
	Máximo	0%	54%	100%	29%	30%	43%	100%

Tabela 43. Análise de variância (ANOVA) dos clusters de indivíduos – Método K-means

Variáveis		Soma dos Quadrados	gl	Erro Quadrático Médio	F	Valor-P
NumAtividades	Entre clusters	3814,856	5	762,971	37,474	0,000
	Intra cluster	2015,658	99	20,360		
	Total	5830,514	104			
NumClusters	Entre clusters	116,524	5	23,305	29,838	0,000
	Intra cluster	77,324	99	0,781		
	Total	193,848	104			
MaxRepClusters	Entre clusters	463,052	5	92,610	21,496	0,000
	Intra cluster	426,510	99	4,308		
	Total	889,562	104			
%Cluster 1 (Residência – Longa)	Entre clusters	0,386	5	0,077	7,026	0,000
	Intra cluster	1,088	99	0,011		
	Total	1,474	104			
%Cluster 2 (Residência – Curta-média)	Entre clusters	0,909	5	0,182	25,979	0,000
	Intra cluster	0,693	99	0,007		
	Total	1,602	104			
%Cluster 3 (Fora – Média-longa)	Entre clusters	2,069	5	0,414	63,802	0,000
	Intra cluster	0,642	99	0,006		
	Total	2,711	104			
%Cluster 4 (Dentro – Média-longa)	Entre clusters	1,284	5	0,257	47,919	0,000
	Intra cluster	0,530	99	0,005		
	Total	1,814	104			
%Cluster 5 (Dentro – Curta)	Entre clusters	5,237	5	1,047	67,947	0,000
	Intra cluster	1,526	99	0,015		
	Total	6,763	104			
%Cluster 6 (Fora – Curta)	Entre clusters	3,038	5	0,608	38,550	0,000
	Intra cluster	1,560	99	0,016		
	Total	4,599	104			

2. Distribuição de Variáveis por Grupo de Indivíduo

Neste Apêndice são apresentadas as distribuições das variáveis consideradas no processo de classificação, para cada cluster e para todos os indivíduos. Em vermelho forte estão as curvas de distribuição das variáveis para os indivíduos de cada cluster; as curvas de cor mais opaca ao fundo apresentam a distribuição das variáveis para todos os indivíduos, de todos os clusters.

Figura 55. Distribuição da variável *NumAtividades* para cada cluster.

Figura 56. Distribuição da variável *NumClusters* para cada cluster.

Figura 57. Distribuição da variável *MaxRepClusters* para cada cluster.

Figura 58. Distribuição da variável *%Cluster 1 (Residência – Longa)* para cada cluster.

Figura 59. Distribuição da variável *%Cluster 2 (Residência – Curta-média)* para cada cluster.

Figura 60. Distribuição da variável *%Cluster 3 (Fora – Média-longa)* para cada cluster.

Figura 61. Distribuição da variável *%Cluster 4 (Dentro – Média-longa)* para cada cluster.

Figura 62. Distribuição da variável *%Cluster 5 (Dentro – Curta)* para cada cluster.

Figura 63. Distribuição da variável *%Cluster 6 (Fora - Curta)* para cada cluster.

Anexos

1. Formulário da entrevista domiciliar

PESQUISA MOBILIDADE URBANA E TRANSPORTE EM PARAISÓPOLIS

Pesquisa realizada na data: ____ / ____ / ____ Hora: ____:____ Nome do pesquisador: _____

Identificação (nº do selo ou apartamento): _____

Qual o número do celular que será usado na pesquisa? (____) _____ Operadora (pré ou pós pago) _____

Tem plano com pacote de dados de internet? (S/N) ☐ Modelo: _____ Hora de Início: _____

1. PESSOAL

1.1 Nome: _____

1.1 Endereço: _____ 1.3 Idade: _____

1.4 Gênero: ☐ 1.Feminino ☐ 2.Masculino ☐ 3.Outro: 1.4.1 _____

1.5 Estado civil: ☐ 1.Solteira/o ☐ 2.Casada/o ou em relacionamento estável
☐ 3.Viúva/o ☐ 4.Divorciada/o ☐ 5.Outro: 1.5.1 _____

2. ESCOLARIDADE

☐ 1.Não estudou ☐ 2.Fundamental incompleto ☐ 4.Médio incompleto ☐ 6.Superior incompleto
☐ 3.Fundamental completo ☐ 5.Médio completo ☐ 7.Superior completo

2.1 Fez algum outro curso (ETEC, EJA, outros)? ☐ 1.Sim. Qual? 2.1.1 _____ ☐ 2.Não

2.2 Está estudando atualmente? ☐ 1.Sim ☐ 2.Não (siga p/ 3.1)

2.2.1 Se sim, qual o período? ☐ 1.Manhã ☐ 2.Tarde ☐ 3.Noite ☐ EAD ☐ EAD Semi presencial

2.2.2 Se sim, qual o curso/série? ☐ 1.Ano ens. fundamental ☐ 2.Ano ens. médio ☐ 3.Ano ens. superior
☐ 4.ETEC ☐ 5.Outro: 2.2.2.1 _____

3. ORIGEM

3.1 Você nasceu em Paraisópolis? ☐ 1.Sim (siga p/3.3) ☐ 2.Não, cidade/estado: 3.1.1 _____

3.2 Se não, há quanto tempo mora em Paraisópolis? ____ anos

I. Menos de 1 ano ☐ 3.2.1 Onde morava antes? _____

3.3 Você mora em Paraisópolis por quê? (admita múltiplas respostas):

☐ I. Tinha família aqui. ☐ IV. Não tenho condições de morar em outro lugar.
☐ II. Tive uma oportunidade de emprego na região. ☐ V. Paraisópolis tem acesso fácil por transporte público.
☐ III. Tive uma oportunidade de moradia. ☐ VI. Outros 3.3.1 _____

3.4 Se não nasceu no Paraisópolis, veio morar por quê? (admita múltiplas respostas)

☐ I. Tinha amigos aqui. ☐ VI. Tinha família aqui.
☐ II. Porque tem emprego próximo. ☐ VII. Teve uma oportunidade de emprego na região.
☐ III. Teve uma oportunidade de moradia ☐ VIII. Não tinha condições de morar em outro lugar.
☐ IV. Paraisópolis tem acesso fácil por transporte público ☐ IX. Não fui eu que tomei essa decisão.
☐ V. É fácil chegar em outros lugares da cidade. ☐ X. Outros 3.4.1 _____

4 SOCIOECONÔMICO

4.1 Você já foi beneficiado por algum programa habitacional? ☐ Sim. Qual? 4.1.1 _____ ☐ Não
(CDHU, MCMV, COHAB, etc.) (siga p/ 4.3)

4.2 Se não, recebe aluguel social (auxílio moradia)? ☐ Sim ☐ Não

4.3 Alguém que mora no mesmo domicílio já se beneficiou de algum programa habitacional? ☐ 1.Sim. Qual? 4.3.1 _____ ☐ 2.Não
☐ 3.Não sei

4.4 Você ou alguém que mora no seu domicílio se beneficiam de algum outro programa social? ☐ 1.Sim. Qual? 4.4.1 _____ ☐ 2.Não

(Bolsa Família, renda mínima, renda cidadã, Bolsa escola, etc.)

5 FAMÍLIA

5.1 Mora com quantas pessoas? (colocar número em cada categoria)

☐	I. Esposa/marido ou companheira/o	☐	V. Netos
☐	II. Pai	☐	VI. Mãe
☐	III. Filha/o ou enteada (se não tiver siga p/6.1)	☐	VII. Irmãos
☐	IV. Avós	☐	VIII. Outros: 5.1.1 _____

5.2 Deslocamento dos menores de idade do domicílio para a escola (se tiver filhas/os ou esteadas/os que estudam)

	5.2.1 Menor	5.2.2 Menor	5.2.3 Menor	5.2.4 Menor	5.2.5 Menor
I. Qual é a idade?					
II. Masculino ou Feminino? (M/F)					
III. Qual é a escola? *					
IV. Em qual período estuda? (M/V/N)					
V. Ela/e vai sozinha/o (sem um responsável acompanhando)? (S/N)					
VI. Como vai para a escola? **					
VII. Ela/e volta sozinha/o (sem um responsável acompanhando)? (S/N)					
VIII. Como volta da escola? **					

* Creche Cedrinho(1), Creche Anglicana(2), Cei Paraisópolis(3), Emei Burle Marx(4), Emei Paraisópolis(5), Emei Perimetral II(6), Ceu Paraisópolis (Cei, Emei E Emei)(7), Emef Paulo Freire(8), Emef Casarão(9), Emef Perimetral I(10), Emef D. Veremundo Toth(11), Ee Homero(12), Ee Miguel Arrees(13), Ee Zilda Gamba Datel(14), Ee Etelvina De Góes(15), Etec Abdias Do Nascimento(16), Colégio Porto Seguro(17), Outros Dentro do Paraisópolis(18), Outros fora do Paraisópolis(19)

**Perua(1) / ônibus(2) / a pé(3) / carro(4) / bicicleta(5) / outro(6)

6 OCUPAÇÃO

6.1 Qual a sua ocupação? (admite respostas múltiplas)

☐	I. Não trabalha	☐	VII. Tem negócio próprio
☐	II. Dona/o de casa	☐	VIII. Estudante
☐	III. Trabalhos pontuais ("faz bicos")	☐	IX. Desempregado
☐	IV. Trabalha SEM carteira assinada	☐	X. Aposentado
☐	V. Trabalha COM carteira assinada	☐	XI. Outro: 6.1.1 _____
☐	VI. Autônomo		(se não trabalha ir p/7.1)

6.2 (Para quem trabalha) Você trabalha em Paraisópolis (S/N)? ____ 6.2.1 Se sim, em qual rua? _____

6.2.2 Se não, em qual cidade e bairro trabalha? _____ 6.2.3 Há quanto tempo está no seu trabalho atual? ____

6.3 Trabalha em geral quantos dias na semana? ____ 6.3.1 Quando? ☐ I. Dias úteis ☐ II. Sábado ☐ III. Domingo
(pode ter respostas múltiplas) =

6.3.2 Em qual período entra? ☐ 1. Manhã ☐ 2. Tarde ☐ 3. Noite 6.3.3 Trabalha quantas horas por dia? ____

7 TRANSPORTE

7.1 Tem carro? ☐ 1. Sim ☐ 2. Não

7.2 Se sim, onde estaciona o carro dentro do Paraisópolis? ☐ 1. Rua ☐ 2. Garagem ☐ 3. Estacionamento ☐ 4. Outro: 7.2.1 ____

7.3 Planeja adquirir nos próximos 6 meses? ☐ 1. Sim ☐ 2. Incerto ☐ 3. Não

(se já tem, planeja adquirir outro?)

7.4 Se sim, qual tipo de carro?

☐ 1.Novo

☐ 2.Usado

☐ 3.Não sabe

7.5 Se sim, onde vai estacionar?

☐ 1.Rua

☐ 2.Garagem

☐ 3.Estacionamento

☐ 4.Não sei

7.6 Tem motocicleta?

☐ 1.Sim

☐ 2.Não

7.7 Planeja adquirir nos próximos 6 meses? (se já tem, planeja adquirir outra?)

☐ 1.Sim

☐ 2.Incerto

☐ 3.Não

7.8 Tem bicicleta própria?

☐ 1.Sim

☐ 2.Não

7.9 Planeja adquirir nos próximos 6 meses? (se já tem, planeja adquirir outra?)

☐ 1.Sim

☐ 2.Incerto

☐ 3.Não

7.10 Em relação às afirmações abaixo responder de 1 a 5, sendo 1 que discorda totalmente e 5 concorda totalmente.

	1.Dicorda Totalmente	2.Discorda parcialmente	3.Neutro	4.Concorda parcialmente	5.Concorda Totalmente
7.10.1 Você sente/sentiria satisfação de possuir um carro.					
7.10.2 Você acha que seus amigos e familiares te admiram/admirariam por possuir um carro					
7.10.3 Possuir um carro te torna/tornaria uma pessoa mais importante na tua comunidade					
7.10.4 De carro você consegue/conseguiria fazer todas as atividades diárias					
7.10.5 É sempre mais fácil e rápido ir de carro do que de outro modo de transporte					
7.10.6 Mesmo se houver trânsito, de carro você consegue/conseguiria chegar mais rápido onde você precisa					

7.11 Utiliza transporte público?

☐ 1.Sim

☐ 2.Não (siga p/7.12)

7.11.1 Se sim, utiliza alguma dessas modalidades de Bilhete Único ou BOM? (resposta múltipla):

I.Bilhete Único Comum

☐

VI.Bilhete Único Temporal (Mensal/Semanal/Diário)

☐

II.Bilhete BOM (Intermunicipal)

☐

VII.Bilhete Estudantil Comum

☐

III.Vale transporte

☐

VIII.Bilhete Estudantil Temporal

☐

IV.Bilhete de Idoso / Deficiente

☐

IX.Bilhete Estudantil Gratuito

☐

V.Gratuidade desempregado (ônibus)

☐

X.Gratuidade desempregado (METRO/CPTM)

☐

7.11.2 Quanto gasta por mês com transporte público (não incluindo benefícios)? R\$ _____

7.12 Como você costuma se locomover pela comunidade? (admite respostas múltiplas)

☐

I.A pé

☐

II.Lotação/ônibus. Linhas 7.12.1 _____

☐

III.Moto

☐

IV.Moto táxi

☐

V.Bicicleta

☐

VI.Carro

☐

VII.Carona

☐

VIII.Outros: 7.12.2 _____

☐

7.13 Quais os meios de locomoção que mais usa para sair ou fora da comunidade?

☐

I.A pé

☐

II.Lotação/ônibus. Linhas 7.13.1 _____

☐

III.Moto

☐

IV.Metro/CPTM. Linhas 7.13.2 _____

☐

V.Bicicleta

☐

VI.Carro

☐

VII.Carona

☐

VIII.Taxi / Uber

☐

IX.Outros: 7.13.3 _____

☐

7.14 Qual é a sua **principal dificuldade** no transporte diário?

7.15 Quais são as **principais dificuldades**, para os meio que você utiliza, enfrentadas nos seus deslocamentos diários em **Paraisópolis**? Se não utiliza, por quê?

Caminhada 7.15.1 utiliza para deslocamento? ☐

- ☐ I. Falta de calçadas
- ☐ II. Obstáculo na calçada
- ☐ III. Carros estacionados nas calçadas
- ☐ IV. Falta de faixas de pedestre
- ☐ V. Falta de sinal para pedestre
- ☐ VI. Má iluminação
- ☐ VII. Perigo de atropelamento
- ☐ VIII. Perigo de queda
- ☐ IX. Muita subida
- ☐ X. Receio de ser assediada(o) (assobios, assédio verbal ou físico)
- ☐ XI. Perigo de ser roubada/o

Bicicleta 7.15.2 utiliza para deslocamento? ☐

- ☐ I. Falta de ciclovias
- ☐ II. Obstáculos na rua
- ☐ III. Não tem onde deixar a bicicleta
- ☐ IV. Muita subida
- ☐ V. Carros estacionados na rua
- ☐ VI. Motoristas de outros veículos agressivos
- ☐ VII. Má iluminação
- ☐ VIII. Perigo de sofrer quedas
- ☐ IX. Perigo de atropelamento
- ☐ X. Perigo de ser roubada/o
- ☐ XI. Receio de ser assediada(o) (assobios, assédio verbal ou físico)

Ônibus

7.15.3 utiliza para deslocamento? ☐

- ☐ I. Distância ao ponto de ônibus
- ☐ II. Falta de linhas de ônibus
- ☐ III. Alto preço da passagem
- ☐ IV. Espera muito no ponto
- ☐ V. Ônibus lotados

- ☐ VI. Falta de conforto (falta de assentos, muito quente, balança muito, corre muito)
- ☐ VII. Congestionamento dentro da comunidade
- ☐ VIII. Demora muito para chegar no destino
- ☐ IX. Receio de ser assediada(o) (assobios, assédio verbal ou físico)
- ☐ X. Perigo de ser roubada/o

Carro 7.15.4 utiliza para deslocamento? ☐

- ☐ I. Ruas estreitas
- ☐ II. Pedestres na rua
- ☐ III. Não tem onde estacionar
- ☐ IV. Muita subida
- ☐ V. Congestionamento
- ☐ VI. Motoristas de outros veículos agressivos
- ☐ VII. Carros estacionados na rua
- ☐ VIII. Estacionamento caro
- ☐ IX. Receio de ser assediada(o) (assobios, assédio verbal ou físico)
- ☐ X. Perigo de ser roubada/o
- ☐ XI. Perigo de atropelar alguém

Moto 7.15.5 utiliza para deslocamento? ☐

- ☐ I. Ruas estreitas
- ☐ II. Pedestres na rua
- ☐ III. Não tem onde estacionar
- ☐ IV. Muita subida
- ☐ V. Congestionamento
- ☐ VI. Motoristas de outros veículos agressivos
- ☐ VII. Carros estacionados na rua
- ☐ VIII. Estacionamento caro
- ☐ IX. Receio de ser assediada(o) (assobios, assédio verbal ou físico)
- ☐ X. Perigo de ser roubada/o
- ☐ XI. Perigo de atropelar alguém

8 SAÚDE

8.1 Onde você é atendida/o? (admiti respostas múltiplas)

- ☐ I. UBS I
- ☐ II. UBS II

- ☐ IV. AMA
- ☐ V. Consultas em hospitais/médicos especialistas

☐ III.UBS III☐ VI.Outro. Qual? 8.1.1 _____8.2 Você tem dificuldades em comparecer nas consultas marcadas? ☐ 1.Sim ☐ 2.Não (siga p/8.3)

8.2.1 Se sim, por quais motivos? (respostas múltiplas)

☐ I.Morar longe☐ II.Horário das consultas☐ III.Não ter com quem deixar o filho☐ IV.Esquecer o compromisso☐ V.Outro. Qual? 8.2.1.1 _____

8.3 Já sofreu algum acidente de trânsito no Paraísopolis? Se sim, qual a data, período e veículo (quando se aplicar)?

	8.3.1 Acidente	8.3.2 Acidente	8.3.3 Acidente	8.3.4 Acidente
I.Tipo de acidente*				
II.Data				
III.Período (M/V/N)				
IV.Rua				
V.Veículo**				

*Queda em calçada(1), queda em rua(2), atropelamento(3), acidente com outro veículo(4)

** Carro(1), moto(2), ônibus(3), bicicleta(4), outros(5), não se aplica(6).

8.4 Sofre de alguns desses problemas de saúde abaixo? (resposta múltipla)

☐ I.Obesidade (percepção)☐ II.Pressão alta☐ III.Sobrepeso (percepção)☐ IV.Diabetes☐ V.Estresse☐ VI.Problemas respiratórios☐ VII.Outro. Qual? 8.4.1 _____

8.5 Quais fatores do seu dia a dia você acha que influenciam a sua saúde?

☐ I.Dormir pouco ou mal☐ II.Pressão de horários☐ III.Falta de exercício físico☐ IV.Má alimentação☐ V.Muita poluição☐ VI.Barulho☐ VII.Outros

8.5.1 _____

8.6 Faz exercício físico? ☐ 1.Sim ☐ 2.Não (siga p/8.7)8.6.1Se sim, com qual frequência ☐ 1.1x por semana ☐ 2.2x por semana ☐ 3.3x ou mais por semana8.7 Anda pela comunidade? ☐ 1.Sim ☐ 2.Não (siga p/9.1)

(para ir ao comércio, serviços, casa de amigos ou parentes, etc.)

8.7.1Se sim, com qual frequência ☐ 1.1x por semana ☐ 2.2x por semana ☐ 3.3x ou mais por semana

9. Renda

9.1 Renda familiar R\$ _____

☐ 1.Recusa☐ 2.Sem renda

9.2 Renda Pessoal R\$ _____

☐ 1.Recusa☐ 2.Sem renda

10. É cliente de quais operadoras?

	10.1 Tim	10.2 Oi	10.3 Claro	10.4 Vivo	10.5Nextel	10.6 Outra
1.Pré-pago						
2.Pós-Pago						

Você concordaria que entrassem em contato posteriormente para que fossem respondidas algumas perguntas complementares? (S/N)

☐

TABELA – QUAIS SÃO OS LOCAIS QUE COSTUMA FREQUENTAR

I.Atividade*	II.Local (se dentro do Paraisópolis a Rua)	Meio de Transporte						X.Frequência			Período		Tempo de Viagem [min]	
		III.A pé	IV.Bici	V.Ônibus/ lotação	VI.Metr ô/trem	VII.Carr o	VIII.Mot o	IX.Outro s. Qual?	1. Diário	2. 1 vez/ semana	3. 1 vez/ mês	XI.Semana	XII.Fim de semana	XIII.Ida

*Atividades: Trabalho(1) / Estudos(2) / Estudos de filhos (3) / Compras (4) / Lazer (5) / Médico (Consultas UBS) (6) / Médico (Consultas em especialistas) (7) / Visitar parentes (8) / Outros (9)