

Caracterização de padrões de deslocamentos urbanos em Fortaleza com a utilização de dados de redes sociais georreferenciadas

Characterization of patterns of urban displacements in Fortaleza with the use of data from georeferenced social networks

Sameque Farias Cunha de Oliveira¹ e Carlos Augusto Uchôa da Silva²

¹Universidade Federal do Ceará, Ceará – Brasil, sameque@det.ufc.br

²Universidade Federal do Ceará, Ceará – Brasil, uchoa@det.ufc.br

Recebido:

27 de agosto de 2019

Aceito para publicação:

11 de maio de 2020

Publicado:

15 de dezembro de 2020

Editor de área:

Cira Pitombo

Palavras-chaves:

Redes sociais.

Padrão de deslocamento.

Caracterização socioeconômica.

Regressão espacial.

Keywords:

Social network.

Pattern of displacement.

Socioeconomic characterization.

Spatial regression.

DOI:10.14295/transportes.v28i5.2153

**RESUMO**

Técnicas tradicionais para a obtenção de dados sobre mobilidade vêm sofrendo um processo de defasagem. Nesse cenário, técnicas alternativas, de baixo custo e capazes de incorporar as dinamicidades desses padrões de deslocamentos, têm se mostrado atrativas. Dentre elas, podem ser citadas as bases de dados, provenientes de redes sociais, baseadas em localização. Portanto, o objetivo principal deste trabalho foi a elaboração de um método para caracterizar padrões de mobilidade em Fortaleza com a utilização de dados do *Twitter* e *Instagram*. O método proposto possibilitou a atribuição de viagens a partir dos *check-ins*, identificando os pares OD. Além disso, foi sugerido um método de caracterização socioeconômica dos indivíduos, através de uma regressão espacial. Os resultados indicam que o método foi eficaz na identificação de padrões de deslocamento de pessoas de média e alta renda, principalmente de viagens por motivo lazer. No entanto, uma das limitações mais visíveis deste trabalho é o fato de ele não ter sido capaz de representar o comportamento da população de renda mais baixa na cidade de Fortaleza.

ABSTRACT

Traditional techniques for obtaining data on mobility had suffered a delay process. In this scenario, alternative, low cost techniques capable of incorporating the dynamics of these displacement patterns have been attractive. These include databases from location-based social networks. Therefore, the main objective of this work was the elaboration of a method to characterize mobility patterns in Fortaleza using *Twitter* and *Instagram* data. The proposed method allowed the assignment of trips from the *check-ins*, identifying the OD pairs. In addition, a method of socioeconomic characterization of individuals through spatial regression was suggested. The results indicate that the method was effective in identifying displacement patterns of medium and high income people, mainly for leisure travel. However, one of the results constrains was the inability of representing the lower income population travel behaviour at the city of Fortaleza.

1. CONTEXTUALIZAÇÃO

O conhecimento dos padrões de deslocamentos de pessoas, em especial das origens e destinos de suas viagens, tem se mostrado essencial no processo de planejamento de transportes. Para a obtenção desse tipo de informação é preciso recorrer convencionalmente, a pesquisas domiciliares, a partir das quais são obtidas informações a respeito dos padrões de viagens dos indivíduos (motivo da viagem, origem e destino, modo utilizado, dentre outras), além de suas

características socioeconômicas. O principal produto desse tipo de pesquisa é a estimação de uma matriz origem-destino, com todos os deslocamentos realizados ao longo de um dia típico.

Reconhece-se, entretanto, que tal tipo de abordagem apresenta algumas limitações, relativas tanto à operacionalização da pesquisa quanto à qualidade das informações obtidas. No primeiro caso, a principal limitação está relacionada aos elevados custos, tanto financeiros quanto de tempo (normalmente mais de um ano), para a execução desse tipo de pesquisa. Além disso, é comum verificar erros inerentes ao processo de coleta, tabulação e armazenagem dos dados (Hellings, 1995). Do ponto de vista da qualidade da informação obtida, o que mais se questiona nesse tipo de pesquisa é o fato de ela retratar apenas um recorte temporal específico, normalmente um dia da semana considerado típico. Logo, a matriz obtida apresenta deficiências, já que não consegue incorporar as dinâmicas espaciais e temporais inerentes ao fenômeno analisado (Iqbal *et al.*, 2014).

Tais limitações reforçaram a necessidade de se buscar fontes alternativas de dados de viagens, de modo a preencher as lacunas observadas nas informações providas de pesquisas domiciliares (Ferreira, 2013; Wolf e Thompson, 2003). Dentre algumas alternativas, se destaca a utilização de dados provenientes de redes sociais baseadas em localização (*RSBL*), como o *Twitter* e o *Instagram*, por exemplo. Estes tipos de aplicativos possibilitam a obtenção da localização geográfica dos usuários no momento em que estes compartilham suas atividades através dos chamados *check-ins*. Isto é possível pela incorporação de sistemas globais de posicionamento por satélites (*SGPS*) aos aparelhos celulares. Tais informações apresentam a grande vantagem de serem coletadas em grande escala e em tempo real permitindo dessa forma a incorporação de dinâmicas temporais na análise do padrão de viagens dos indivíduos (Hu e Peter, 2015). Portanto, o uso dessas ferramentas, quando combinadas a outras bases de dados (uso do solo, censo demográfico, por exemplo), despontam como potenciais fontes de informação para se estimar padrões de mobilidade no meio urbano, informação essencial ao processo de planejamento (Agryzkov *et al.*, 2017).

Diante desse contexto, o objetivo principal deste artigo é o desenvolvimento de um método para se estimar o padrão de deslocamentos no meio urbano a partir de dados provenientes de redes sociais baseadas em localização. Para isso, primeiramente foi feita uma revisão da literatura (Seção 2), cujo objetivo foi à identificação de lacunas nessa abordagem a partir do levantamento de trabalhos que utilizaram dados de redes sociais para caracterização de padrões de mobilidade no meio urbano. A Seção 3 introduz o método utilizado para a análise, que vai desde a obtenção e tratamento dos dados, estimação dos motivos das viagens e construção das matrizes origem-destino, além da identificação do perfil socioeconômico dos domicílios amostrados. Já na Seção 4, é apresentada a aplicação do método proposto a partir de um estudo de caso para a cidade de Fortaleza. A partir dos resultados obtidos, foi feita uma discussão do método apresentado, analisando as principais limitações encontradas. Por fim, a Seção 5 traz as principais conclusões do trabalho e recomendações de melhorias para trabalhos futuros.

2. REVISÃO DA LITERATURA

Nesta última década, houve um crescimento no interesse da comunidade científica em explorar a utilização de dados de redes sociais como instrumento de obtenção de padrões de deslocamentos de indivíduos no meio urbano. Os parágrafos a seguir trazem um resumo do estado da arte nesta temática, identificando suas limitações, além de lacunas ainda não exploradas pela literatura.

2.1. Estado da arte no uso de dados de redes sociais para estimação de padrões de deslocamentos

Um dos primeiros trabalhos com temática baseada na utilização de dados de redes sociais como ferramentas para a compreensão de características socioeconômicas de indivíduos foi o estudo de Cheng *et al.* (2011). Os autores utilizaram dados do *Twitter*, referentes a aproximadamente 22 milhões de *check-ins* de usuários do mundo inteiro. O principal objetivo da pesquisa foi à estimação de características socioeconômicas desses usuários a partir de aspectos espaciais e temporais das postagens. Foi verificado que a maior densidade de *check-ins* aconteceu em regiões da América do Norte, Europa Ocidental e Ásia Meridional e Pacífica. Além disso, foi percebida uma maior concentração de postagem às 9h, 12h e 18h. O perfil socioeconômico dos usuários foi determinado com base na sua influência na rede social. Desta forma, ficou evidente que indivíduos com grande quantidade de seguidores pertenciam a classes sociais mais elevadas.

Cranshaw *et al.* (2012) propuseram um modelo de agrupamento para estudar a mobilidade da cidade de Pittsburgh com base nos dados gerados pelas redes sociais. A metodologia foi aplicada e obteve aproximadamente 42787 *check-ins* do *Twitter*. Além disso, um algoritmo foi desenvolvido capaz de varrer uma determinada área e encontrar *clusters*, que posteriormente foram denominados de *lifelihoods* (representação de áreas dinâmicas da cidade). Com o intuito de validar essas áreas, foi realizada uma entrevista com 27 habitantes. Cada entrevista teve duração de uma hora. Os resultados mostraram que os *clusters* podem fornecer um suporte para o entendimento da dinâmica acerca das atividades realizadas pelos usuários dentro do meio urbano.

Hasan *et al.* (2013) avançaram na utilização desses tipos de dados, ao inferir padrões de atividades realizadas por indivíduos a partir de dados do *Twitter* e do *Foursquare*. Foram coletadas informações de *check-ins* dos usuários dessas redes ao longo de 10 meses, em três cidades americanas (Nova Iorque, Chicago e Los Angeles). Com esses dados, os autores construíram um mapa com a distribuição das atividades dos indivíduos da amostra, destacando os locais mais frequentados por esses usuários. Essa amostra foi, então, extrapolada com a utilização de métodos de estimação de densidade de *Kernel*, cujo produto foi a obtenção de mapas de atividades nas três cidades analisadas para períodos temporais distintos, ao longo do dia.

Apesar de trazer avanços no conhecimento do perfil socioeconômico e do padrão de atividades dos usuários de redes sociais, os trabalhos anteriormente citados, pouco contribuíram para a compreensão dos deslocamentos desses indivíduos no meio urbano. Para preencher essa lacuna, Gao *et al.* (2014) utilizaram dados provenientes do *Twitter* para estimar uma matriz origem-destino, no âmbito regional, com os deslocamentos entre cidades da Região Metropolitana da Califórnia, nos Estados Unidos. Para isso, foi coletada uma amostra de aproximadamente 6,8 milhões de *tweets* georreferenciados de mais de 110 mil usuários. O método utilizado para se determinar as viagens era baseado em duas etapas: detecção do deslocamento do indivíduo e agregação da viagem na área. Na primeira etapa, a trajetória do indivíduo era identificada pela sequência de *tweets* geolocalizados. Uma vez definidos o início e o final desse deslocamento, cada viagem terá uma origem e um destino, no nível de agregação considerado.

Outro trabalho cuja proposta foi a estimação de uma matriz origem-destino a partir de dados de redes sociais foi o de Abbasi *et al.* (2017). Em sua pesquisa, eles aplicaram um método probabilístico de estimação de deslocamentos de indivíduos baseado em ranqueamento, através do qual é determinada a probabilidade de determinado destino dada uma origem. A principal dificuldade deste método, entretanto, está justamente em ranquear esses potenciais destinos.

Deste modo, os autores aplicaram três métodos distintos de definição desse ranqueamento: um no qual o peso dos possíveis destinos era determinado pela distância destes à origem; um segundo baseado na quantidade de estabelecimentos entre a origem e o destino; e um terceiro que ponderava a quantidade de estabelecimentos pelo total de *check-ins* realizados nos mesmos. A análise foi realizada em Nova Iorque e utilizou dados provenientes do *Foursquare*. Os resultados dos três modelos foram comparados com os dados de viagens de táxi fornecidos pela Comissão de Táxis e Limousines (TLC). Concluíram que o uso dos dados da rede social melhorou a capacidade de previsão do modelo, em comparação aos dois primeiros métodos, devido à capacidade de se incorporar dinâmicas do comportamento humano através das informações de *check-ins*.

Mollay *et al.* (2017) modelaram a atratividade de destino de viagem em Ontário através dos *check-ins* oriundos do *Foursquare*. A metodologia consistiu na agregação dos *check-ins* como indicadores de atratividade de destino para categorias específicas. Para isso, foi aplicado o modelo *logit multinomial*, e este estimou o potencial dos *check-ins*, em medir a atratividade. Os resultados mostraram que a precisão do modelo em todos os fins de viagem melhorou significativamente, particularmente para viagens de lazer.

Cheng *et al.* (2018) desenvolveram um modelo de regressão *random forest* para estimar o total de viagens entre pares OD da região metropolitana de Sydney, na Austrália. Foram construídos dois modelos distintos: um com a utilização de dados de viagens de uma pesquisa domiciliar, além de informações do censo demográfico; e um segundo modelo incorporando dados de *check-ins* do *Twitter* em substituição à pesquisa domiciliar. Neste segundo modelo, foi utilizada uma amostra de mais de 1 milhão de *tweets* georreferenciados, referentes a mais de 170 mil usuários. Os autores obtiveram uma melhor precisão e estabilidade no modelo de regressão, ao considerar os dados do *Twitter* no modelo.

Por fim, Oliveira *et al.* (2019) obtiveram a posição dos usuários do *Twitter* e do *Instagram*, com o objetivo de identificar o uso do solo da cidade de Fortaleza-CE. O método consistiu na aplicação de um algoritmo e o uso do site *Netlytic* para aquisição dos dados, que foram mapeados e confrontados às informações sobre os diferentes tipos de uso do solo, fornecidas pela prefeitura municipal. Os resultados indicaram que o método apresentou uma alta correlação espacial entre o uso do solo inferido e o solo oficial, permitindo a identificação de áreas domiciliares e comerciais.

2.2. Lacunas metodológicas

A partir da revisão realizada foram identificadas algumas lacunas, principalmente de cunho metodológico, nesses trabalhos. A primeira delas foi a ausência de um método para a identificação dos domicílios dos indivíduos amostrados (Gao *et al.*, 2014; Abbasi *et al.*, 2017; Cheng *et al.*, 2018). Desta forma, não é possível diferenciar viagens de base domiciliar daquelas de base não domiciliar. Além disso, nos trabalhos citados não foram apresentados métodos para identificar as atividades realizadas nos destinos, de modo a classificar as viagens por motivo.

Foram observadas ainda limitações com relação à caracterização socioeconômica dos indivíduos amostrados. Cheng *et al.*, (2011), por exemplo, fizeram essa caracterização apenas com base no número de seguidores dos usuários, em que se assumiu que quanto maior o número de seguidores maior seria a renda do indivíduo.

Apesar de todo o esforço metodológico com a finalidade de suprir essas deficiências, o método proposto neste artigo se baseou em bases de dados agregados e de períodos distintos das coletas de dados realizadas. Dessa forma, existe uma defasagem temporal no método proposto.

3. MÉTODO PROPOSTO

Nesta seção são descritos os passos metodológicos para se estimar padrões de deslocamentos na cidade de Fortaleza com a utilização de dados de rede sociais. Para isso, primeiramente é feita uma descrição da área de estudo e apresentado o zoneamento utilizado. Em seguida, a base de dados utilizada seria apresentada, detalhando as etapas de coleta e tratamento dessas informações. Por fim, foi introduzido o método utilizado para a construção de uma matriz origem-destino para a região, que passa pela identificação das origens e destinos das observações da amostra, dos motivos das viagens, além da caracterização socioeconômica dos indivíduos.

3.1. Descrição da área de estudo e zoneamento utilizado

O método aqui proposto será aplicado na cidade de Fortaleza, localizada no litoral do estado do Ceará, nordeste brasileiro. Em termos populacionais, é a quinta maior cidade do Brasil, com 2,6 milhões de habitantes, os quais ocupam uma área de aproximadamente 315 km². Tais características fazem com que a capital cearense possua a maior densidade demográfica dentre as capitais brasileiras, com quase 7800 hab/km². Em termos territoriais, Fortaleza é composta por 121 bairros, sendo ela a principal cidade de uma região metropolitana composta por outros 14 municípios, os quais também impactam no sistema de transportes da cidade.

Diante de tal configuração, se optou, neste trabalho, por utilizar o zoneamento proposto por Lima (2017). Neste zoneamento, a cidade de Fortaleza foi dividida em 241 ZATs (Zonas de Análise de Tráfego), formadas a partir da agregação dos setores censitários do censo do ano 2000, com base na homogeneidade econômica. Com isso, possui um raio médio de 500 metros por zona, o que leva a inferir que a maior parte das viagens intrazonais são realizadas por modos não motorizados. Vale destacar que, apesar da influência dos municípios da região metropolitana, estes foram desconsiderados no zoneamento proposto, devido à indisponibilidade de dados relativos a essas regiões.

3.2. Bases de dados utilizadas

Neste trabalho, foram utilizadas duas bases de dados provenientes de RSBL: *Twitter* e *Instagram*. A primeira é considerada um *microblog*, no qual os usuários podem interagir através de mensagens de no máximo 280 caracteres, compartilhando, dentre outras informações, atividades realizadas e locais visitados (Bifet e Frank, 2010). Já o *Instagram* tem como base o compartilhamento de fotos e vídeos, através dos quais é possível identificar as atividades realizadas e a localização dos usuários, através dos *check-ins*, ferramenta disponibilizada pelo aplicativo (Piza, 2012). Além destes, foram também utilizados dados de uso do solo, fornecidos pela Secretaria de Finanças do município (SEFIN).

3.3. Obtenção, tratamento e análise dos dados

A Figura 1 apresenta o fluxograma das etapas do método, que engloba a obtenção e tratamento dos dados, estimação de matrizes origem-destino, por motivo, e inferência da classe econômica dos domicílios dos usuários.

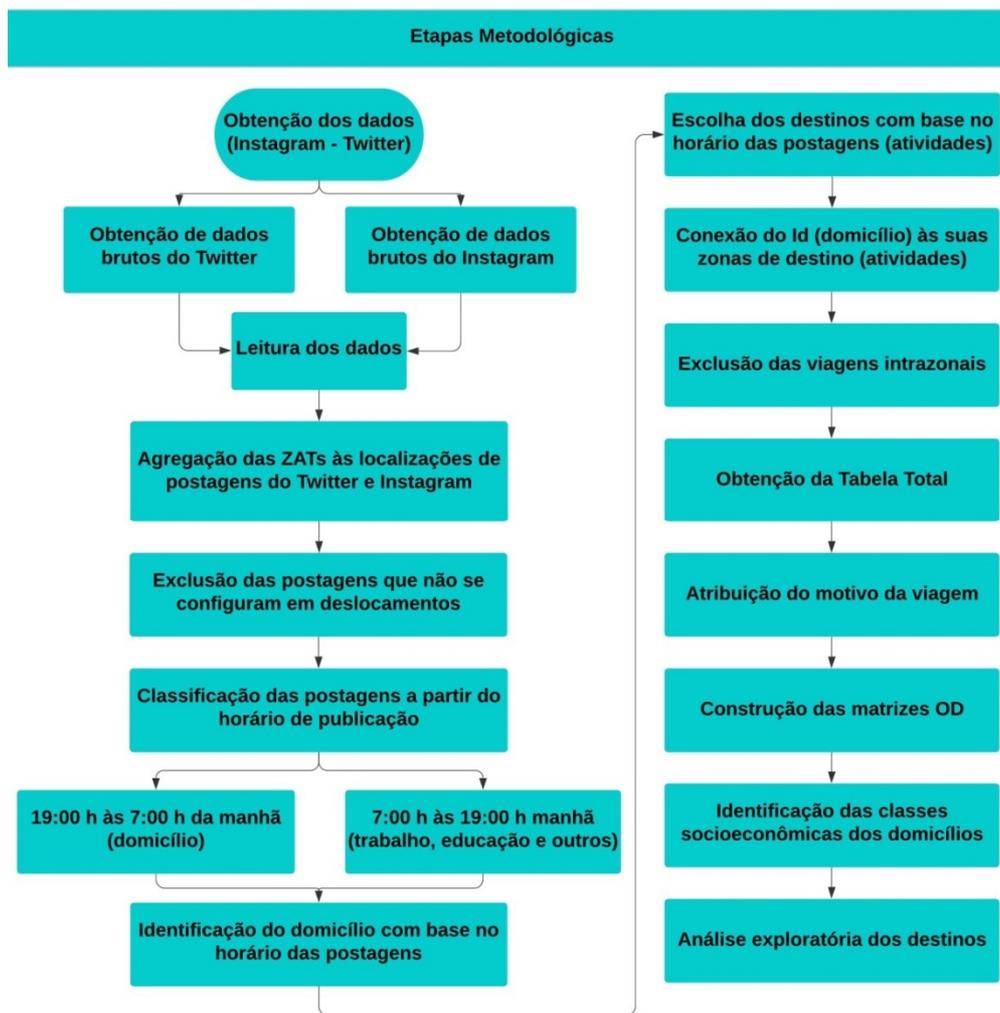


Figura 1. Etapas metodológicas

3.3.1. Obtenção e leitura dos dados

O processo de coleta dos dados do *Twitter* foi iniciado através do *link* a seguir: <https://apps.twitter.com/>. Após o acesso da página, indicada anteriormente, foi necessário realizar um *login* e criar um aplicativo (app). Para criar um app foi necessário preencher algumas informações, tais como: nome ao qual foi nomeado de *Dscan* e descrição do aplicativo. Logo após a criação do app foram obtidas duas chaves de acesso, sendo elas: *Consumer Key* e a *Consumer Secret*. Esses dois *tokens* deram acesso a *Application Programming Interface* - API. Para a coleta ocorrer foi fundamental criar uma rotina em linguagem python e empregar a biblioteca chamada *twython* (Alzahrani, 2018).

Para a execução do *twython* foi necessário inserir na rotina os *tokens* criados. Após a sua implementação, o *twython* utiliza dois parâmetros, sendo eles: a latitude e a longitude, de onde se deseja coletar o *tweet* e o raio.

Para a obtenção dos dados em Fortaleza, aplicou um raio de varredura de 2 km, para cada um dos 241 centroides (latitudes e longitudes) das ZATs (Figura 2), de modo a incorporar toda a área da cidade. Além disso, foram testados diversos valores para o raio de varredura, porém apenas o raio de 2 km forneceu o resultado esperado. Nessa busca, são coletados todos os *tweets* publicados, dentro do raio pré-definido, desde o momento do início da varredura até o momento em que se decidir parar a rotina.

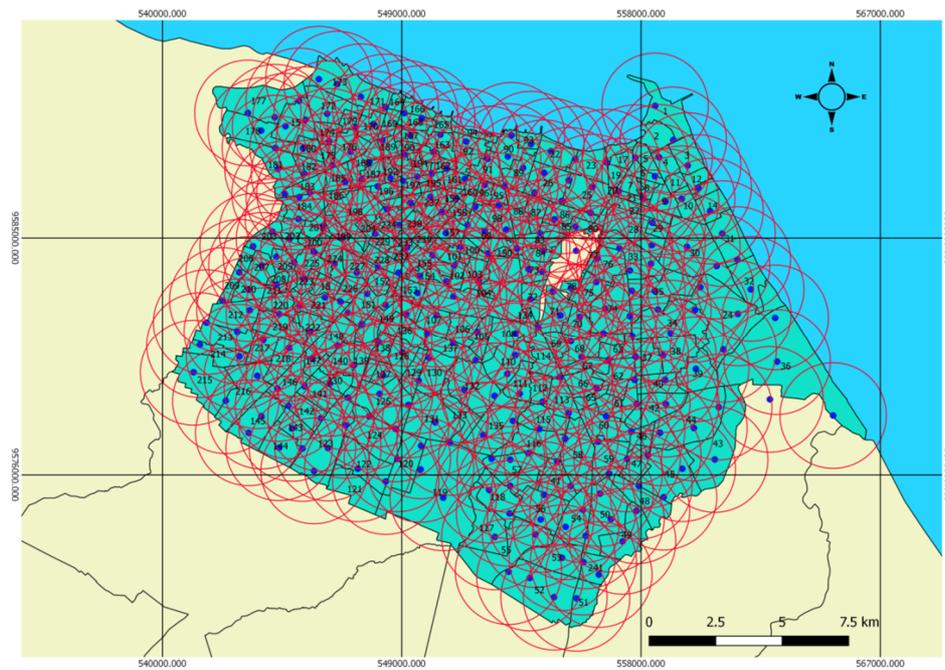


Figura 2. Raio de varredura (2km) utilizado para a coleta de dados do *Twitter*

Em contrapartida, os dados provenientes do *Instagram* foram obtidos através do site <https://netlytic.org>. Em seguida, foram definidos os centroides e raio de varredura (predefinido pelo site) para coleta das informações. O site disponibiliza pontos de coleta, então foram definidos apenas 12 centroides para cobrir a área de estudo de forma integral. A Figura 3 ilustra as áreas definidas para a coleta dos dados do *Instagram*.

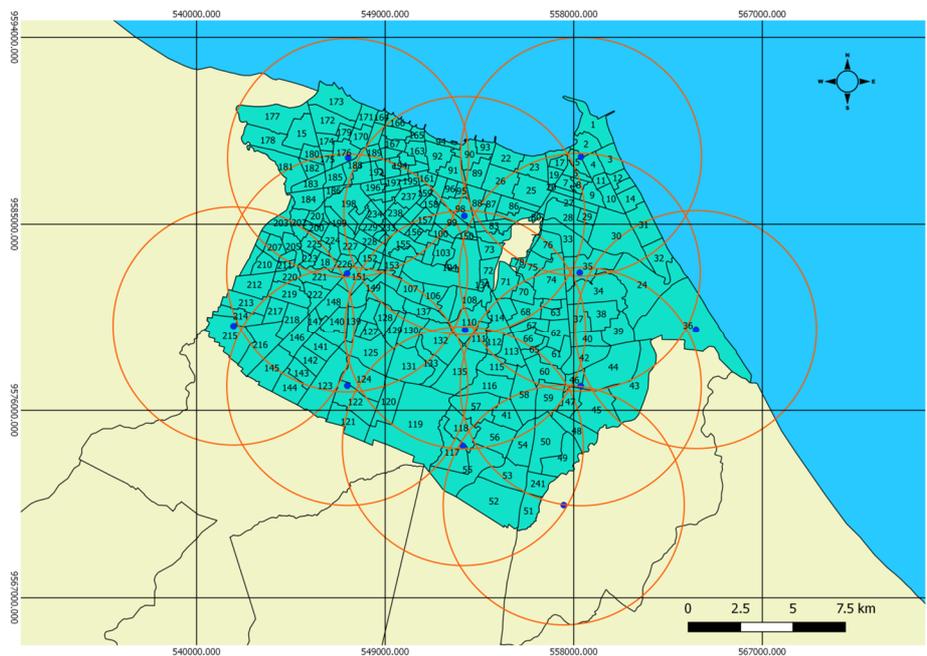


Figura 3. Raio de varredura (5km) utilizado para a coleta de dados do *Instagram*

O tempo de coleta dos dados foi realizado durante o intervalo de um ano, que foi do dia 01/10/2017 a 01/10/2018. Porém, a faixa de tempo utilizada neste trabalho é referente a cinco meses (outubro de 2017 e 22 de março de 2018). Além disso, os dados obtidos pelas duas

redes são advindos de perfis públicos, dessa forma, não existiu a possibilidade de coleta de perfis privados.

Os dados brutos obtidos das duas redes sociais foram, então, convertidos em formato de texto, *.CSV. Dessa forma, foi possível filtrar, através de uma rotina implementada em linguagem R, apenas as informações necessárias para a análise: ID do usuário e data, horário e coordenadas geográficas da postagem. Após a filtragem, as bases de dados foram unificadas. Porém, para evitar erros de duplicação, ou seja, quando um mesmo usuário publica nas duas redes sociais ao mesmo tempo, foi tomado o devido cuidado em verificar se havia publicações realizadas em um mesmo horário (hora, minuto e segundos), latitude e longitude iguais. Assim, caso houvesse uma postagem duplicada, seria considerada apenas uma única.

3.3.2. Associação das postagens às ZATs

Após a leitura e filtragem dos dados, com o uso de uma rotina em linguagem R, foi realizada a associação dos pontos georreferenciados às respectivas ZATs, dentre as 241 existentes. Tem-se, portanto, em $L_u = (l_u^1, l_u^2, \dots, l_u^n)$ a sequência dos locais (L) dos *tweets* e fotos (*Instagram*) geolocalizados (latitude/longitude) de um determinado usuário (u). Esses pontos são, então, associados espacialmente às suas respectivas ZATs, formando, dessa maneira, um conjunto $Z_u = (Z_u^1, Z_u^2, \dots, Z_u^n)$ para cada usuário.

3.3.3. Exclusão das postagens que não se configuram deslocamentos

Para se configurar um deslocamento era necessário que o usuário tivesse realizado pelo menos duas postagens ao longo de um dia ($n \geq 2$), sendo excluídos da amostra todos os indivíduos que fizeram apenas uma postagem em determinado dia.

3.3.4. Classificação das postagens

Nesta etapa foi realizada a organização e classificação das postagens a partir dos horários em que elas foram publicadas, dividindo-as em dois grupos. Postagens publicadas entre 19h e 7h foram consideradas de base domiciliar, partindo da premissa de que nesta faixa de horário os usuários estão em seus domicílios. Na cidade de Fortaleza um indivíduo costuma sair do seu trabalho às 18 h e leva uma média de 32 minutos do trabalho para casa (Pereira, 2013). Já para as postagens realizadas entre 7 h e 19h, foi constatado que estas estavam relacionadas aos locais de atividades diárias (trabalho, educação e outros) dos indivíduos.

3.3.5. Estimativa do local dos domicílios e identificação dos destinos das viagens

Primeiramente, foi feita a estimativa dos locais dos domicílios dos indivíduos da amostra. Para cada usuário, contabilizou o total de postagens entre 19h e 7h, durante cinco meses, e verificou qual ZAT acumulou a maior quantidade delas, a qual foi considerada como a zona de domicílio do indivíduo. Já os destinos foram identificados a partir da localização das postagens realizadas entre 8h e 11h e entre 13h e 18h, intervalos nos quais verificou que esses usuários estariam realizando atividades, classificadas em trabalho, educação ou outros. Além disso, a faixa de tempo que vai das 12 h as 13 h foi desconsiderada, pois é nesse momento que os indivíduos saem dos seus trabalhos para almoçar, dessa forma, postagens neste intervalo de tempo poderia influenciar a amostra. Finalmente, para cada indivíduo, foi construída a conexão entre o seu domicílio e o local de cada uma das atividades identificadas, obtendo assim suas viagens. Neste caso, somente foram consideradas viagens àqueles casos em que a zona do domicílio fosse diferente da zona da atividade, desconsiderado da análise possíveis deslocamentos intrazonais.

Além disso, apenas considerou como destino a última postagem realizada pelo usuário dentro da faixa de tempo estipulada, excluindo as viagens em cadeia. Considerando que grande parte das viagens intrazonais são realizadas por modos não motorizados, os quais geram menor impacto na rede viária, optou-se, portanto, por desconsiderar as viagens intrazonais na análise. Ao final desse processo, foi obtida uma tabela contendo todas as viagens interzonais, com a localização das origens e destinos, realizados dentro da área de estudo.

3.3.6. Atribuição do motivo da viagem

Construída a tabela com as informações de origem e destino de todas as viagens, o passo seguinte foi à verificação da origem, se realmente é de base residencial e conseqüentemente à atribuição de motivos a esses deslocamentos, os quais foram classificados em trabalho, educação e outros motivos. Para isso, foi implementado uma rotina na linguagem R, cujo objetivo era classificar os pontos nas origens e nos destinos de acordo com o tipo de uso do solo do respectivo lote. As informações acerca do tipo de lote foram obtidas através de dados relativos ao ano de 2015, fornecidos pela SEFIN. Nesta base, os lotes estão classificados nas seguintes categorias: comercial, comunicação, desportivo, do lazer, fechado, hotelaria, industrial, institucional, instrução, misto, prestação, religioso, residencial, residencial e comercial, residencial e serviço, saúde, sem classificação, transporte, fechado, agricultura e trabalho e serviços. A distribuição espacial destes usos está ilustrada na Figura 4. Foram desconsideradas as viagens em que as origens não se encontravam em lotes do tipo residencial.



Figura 4. Lotes desagregados do município de Fortaleza

Esses diversos tipos de usos foram, então, agregados em quatro categorias, correspondentes às origens e aos motivos de viagem considerados na análise: residencial, trabalho, educação e outros. A Figura 5 detalha como foi realizada essa classificação.

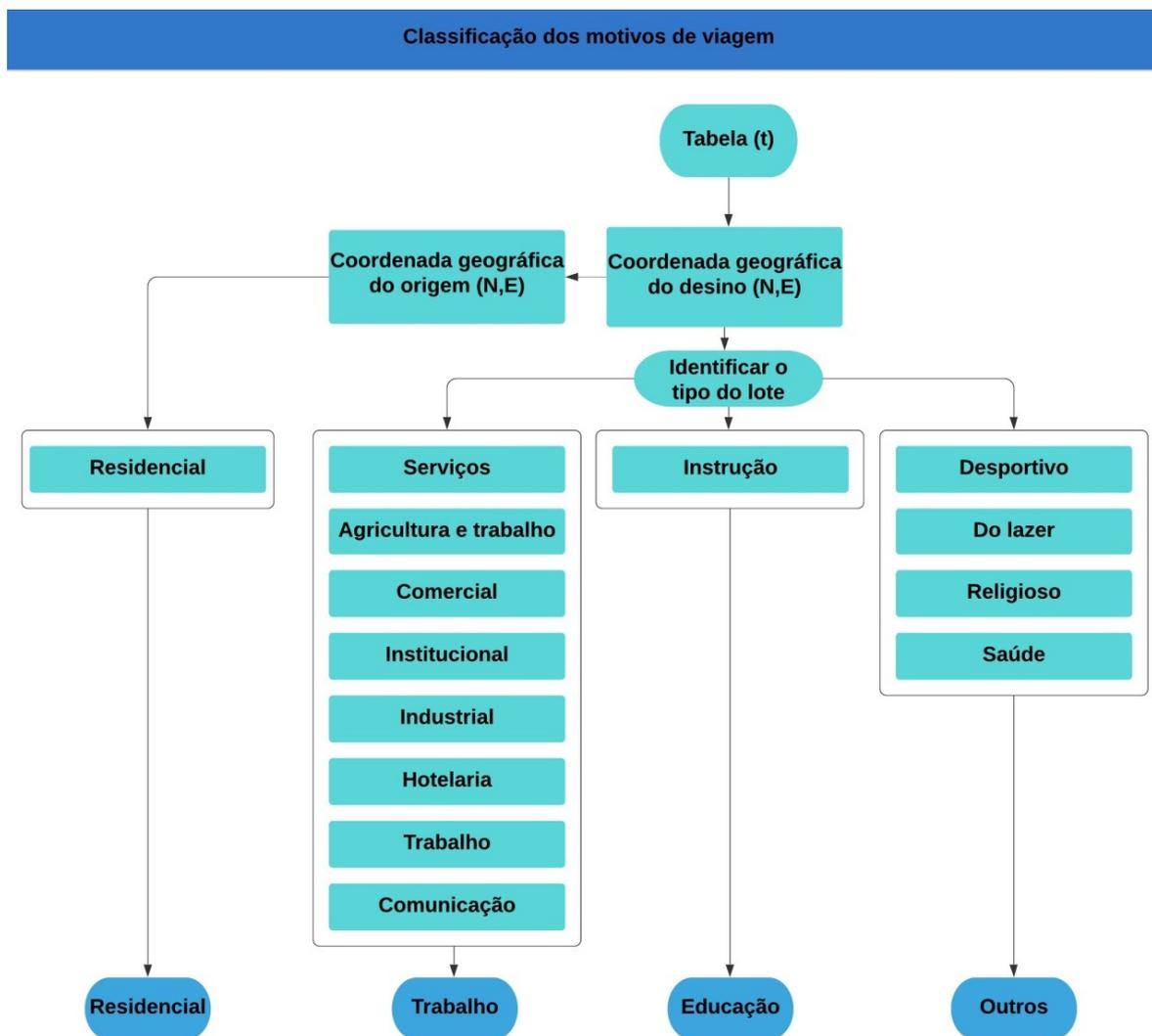


Figura 5. Algoritmo de classificação dos motivos de viagem, com base no tipo de lote da origem e do destino

3.3.7. Obtenção das matrizes origem-destino amostrais

Com as informações de localização das origens e dos destinos das viagens, além da classificação dos motivos destas, foi possível construir três matrizes, para cada motivo de viagem, com os totais de deslocamentos T_{ij} entre as 241 zonas de análise consideradas.

3.3.8. Caracterização socioeconômica dos domicílios

Com o objetivo de se caracterizar os grupos socioeconômicos que compõem a amostra proveniente dos dados de rede sociais, foi realizada uma análise de correlação espacial entre a quantidade de domicílios que tiveram *check-ins* em uma determinada zona de análise e a quantidade de domicílios, por faixa de renda, da respectiva ZAT. As informações de renda dos domicílios foram obtidas através do trabalho de Lima (2017), que considera três estratos de renda distintos: baixa (até 3 salários mínimos), média (entre 3 e 8 salários mínimos) e alta (8 ou mais salários mínimos).

Para avaliar, especialmente, a relação entre as duas variáveis, foi realizado, primeiramente, um teste estatístico univariado de autocorrelação espacial, com o *software GeoDA*, da variável de quantidade de domicílios, para cada estrato de renda. Com isso, foram obtidos os índices

locais de associação espacial (*LISA*) para cada faixa de renda analisada, representados graficamente na forma de mapas. Segundo Ramos (2002), a autocorrelação espacial do índice local de Moran é calculada com base no produto dos desvios em relação à média da variável/indicador observado.

Esse tipo de gráfico de espalhamento classifica as zonas em quatro grupos, com base nos valores da zona relacionados ao valor médio de todas as zonas, cada grupo equivale ao quadrante ao qual a zona pertence no gráfico de Moran (Lima, 2017). Os mapas são apresentados em quatro grupos, o formado pelas zonas *High-High*, ou seja, significa que tanto a zona como as suas vizinhas possuem um valor alto para o indicador. No caso das zonas *Low-Low*, tanto a zona como as suas vizinhas apresentam um valor baixo para o indicador. Por fim, onde essa associação é inversamente equivalente, temos o *High-Low*, com zonas com alto valor indicador e suas vizinhas com baixo. Finalmente, tem o último caso do grupo que são as zonas *Low-High*, ou seja, as zonas com baixo valor indicador e suas vizinhas com alto valor indicador (Anselin, 1995).

Após a classificação dos mapas de espalhamentos, logo em seguida, foi plotado um mapa de densidade *Kernel* com a distribuição espacial dos domicílios da amostra nos quais foram registrados *check-ins*. Esses mapas foram, então, comparados visualmente de modo a identificar a relação entre a concentração desses domicílios e a distribuição espacial dos distintos grupos socioeconômicos. Após essa análise visual, foi feito ainda um teste de regressão espacial (*spatial lag*) entre as variáveis, de modo a quantificar a correlação espacial entre a quantidade de domicílios representativos da amostra de *check-ins* e a quantidade de residências de cada faixa de renda.

4. RESULTADOS E DISCUSSÕES

Nesta seção serão apresentados e analisados os resultados da aplicação do método proposto para o município de Fortaleza.

4.1. Obtenção e leitura dos dados

Para a realização da análise proposta neste trabalho foi coletada uma amostra de 1,4 milhões de *check-ins*, oriundos do *Twitter* e do *Instagram*, referentes ao período compreendido entre 22 de outubro de 2017 e 22 de março de 2018. Para a composição da amostra, só foram consideradas as observações relativas a dias úteis da semana, de segunda a sexta-feira.

O conjunto de dados brutos oriundos do *Twitter* foi coletado no formato *json*. Durante o período definido para a coleta, foram obtidos 30 mil *check-ins* de 10214 usuários distintos. A Figura 6 ilustra todas as informações contidas no dado bruto, muitas das quais são excluídas após uma etapa de filtragem. Com isso, permanecem apenas com as informações consideradas relevantes para a análise, apresentadas no método.

Da mesma maneira, os dados do *Instagram*, coletados de forma automatizada através do site *Netlytic*, passaram pelo mesmo processo de filtragem no algoritmo, de forma a se obter apenas as informações pertinentes (Figura 7). A amostra final, coletada entre outubro de 2017 e março de 2018, foi composta por 1,37 milhão de *check-ins*, associados a 131943 indivíduos.

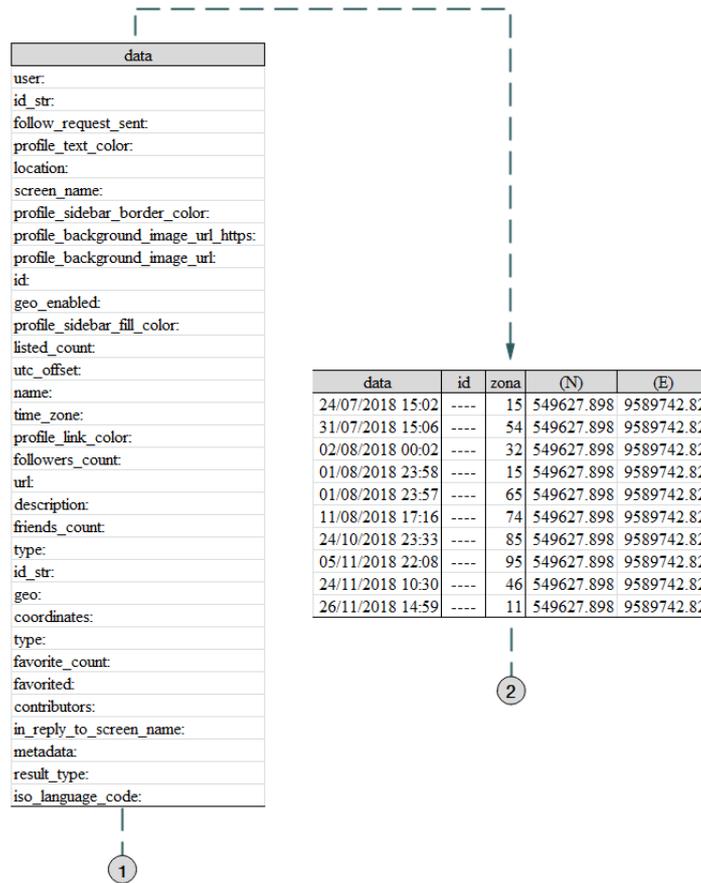


Figura 6. Informações coletadas através do Twitter

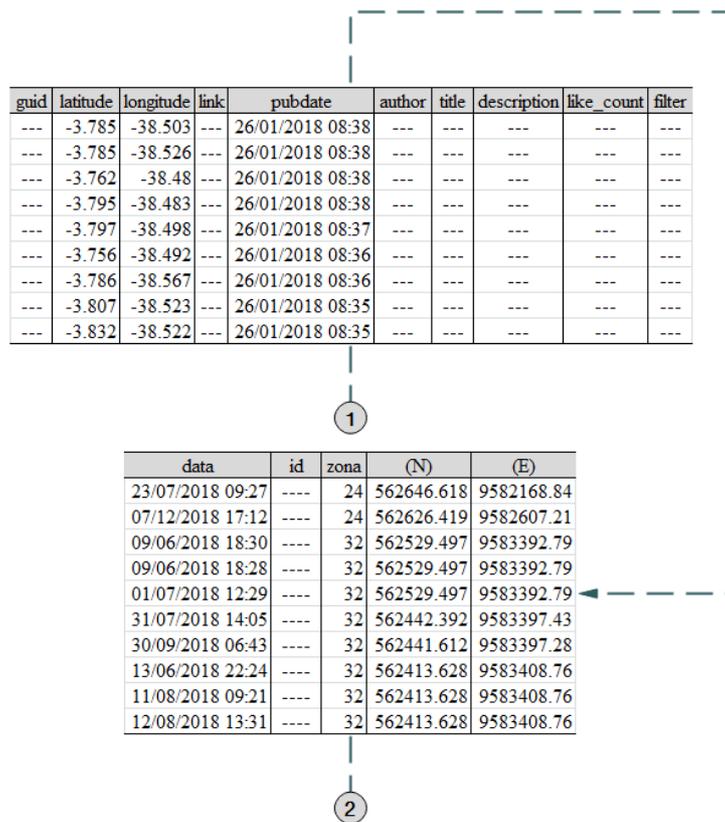


Figura 7. Informações coletadas através do Instagram

4.2. Atribuição das viagens e construção das matrizes origem-destino

As duas tabelas obtidas, com as informações do *Twitter* e do *Instagram*, foram unificadas, resultando em uma única tabela, denominada aqui tabela total. Esta tabela contém apenas os indivíduos que realizaram pelo menos duas postagens ao longo de um dia, como especificado no método. Logo, cada linha corresponde a uma viagem individual, com informações do id do usuário, zona de origem e de destino e horário do *check-in* na origem e no destino. Ao todo, foram obtidas 236264 viagens correspondentes a um total de 48854 indivíduos.

Para cada uma destas viagens, foi verificado se a origem era proveniente de um lote residencial e, posteriormente, determinado um motivo, a partir da identificação do tipo de uso do solo das coordenadas geográficas das origens e dos destinos. Para tipo de lote foi atribuído um motivo, sendo para a origem (residencial) e o da viagem (trabalho, educação ou outros), de acordo com a Figura 5. A Figura 8 apresenta uma parte da tabela total, com a listagem das viagens, já classificadas por motivo.

id	origem	N (O)	E (O)	lote1	destino	data	N (D)	E (D)	diasemana	lote2	motivo
---	68	555277	9580856	residencial	35	03/03/2018 01:12	557643	9582207	6	comercial	trabalho
---	188	547987	9588012	residencial	187	03/03/2018 01:10	548142	9587154	5	comercial	trabalho
---	188	547987	9588012	residencial	187	03/03/2018 01:10	548142	9587154	5	comercial	trabalho
---	31	561195	9585805	residencial	159	29/01/2018 01:18	550973	9586644	6	comercial	trabalho
---	31	561195	9585805	residencial	159	03/01/2018 03:32	550982	9586645	4	comercial	trabalho
---	93	553418	9588687	residencial	32	29/01/2018 01:45	562315	9583519	4	comercial	trabalho
---	93	553418	9588687	residencial	197	03/03/2018 01:10	549787	9587239	6	transportes	trabalho
---	93	553418	9588687	residencial	197	29/01/2018 01:18	549787	9587239	2	transportes	trabalho
---	35	557573	9583395	residencial	55	29/01/2018 01:18	554553	9573834	2	transportes	trabalho
---	35	557573	9583395	residencial	89	03/03/2018 01:10	553498	9587450	4	transportes	trabalho
---	110	552877	9579351	residencial	23	29/01/2018 01:18	556773	9587215	2	religioso	outros
---	110	552877	9579351	residencial	88	03/03/2018 01:10	553506	9585726	5	religioso	outros

Figura 8. Classificação a partir do lote de origem e destino

As viagens realizadas tiveram origem de base residencial. Essas estão exemplificadas na Figura 9. Percebe-se que a distribuição das origens dessas viagens se intensifica em áreas de média e alta renda, como exemplificação tiveram pontos plotados dentro do residencial Landscape. Esse residencial é considerado de luxo e está localizado na zona 23, no norte da cidade. A zona 23 e as suas vizinhas são adensadas por residências e prédios de alto padrão.

Com as informações de viagens, diferenciadas por motivo, foram construídas três matrizes origem-destinos para a cidade de Fortaleza. Neste sentido, alguns pontos merecem ser destacados acerca dos resultados obtidos. Primeiramente, se constatou que mais da metade das viagens identificadas (55%) foram por motivo trabalho, sendo essas viagens realizadas a partir da premissa de que das 8 h às 11 h e das 13 h às 18h os usuários normalmente estão nos seus trabalhos. Por outro lado, as viagens por motivo educação representaram apenas 5% da amostra obtida, enquanto os 40% restantes foram referentes a viagens por outros motivos. Isso indica uma tendência de que esses *check-ins* estejam representando mais fortemente viagens do tipo trabalho, ou seja, os usuários costumam realizar publicações quando estão em seus trabalhos.

A Figura 10 apresenta a distribuição espacial das atividades compulsórias, tipo trabalho e educação. Verifica-se, pela análise do mapa, uma forte concentração destas atividades na região central e na área leste da cidade. A primeira se caracteriza por ser uma região onde predominam

estabelecimentos comerciais e de prestação de serviços, concentrando a maior parte dos empregos ofertados. Já na região leste, foi verificada uma concentração forte de faculdades particulares, o que tende a atrair muitas viagens por motivos educacionais.

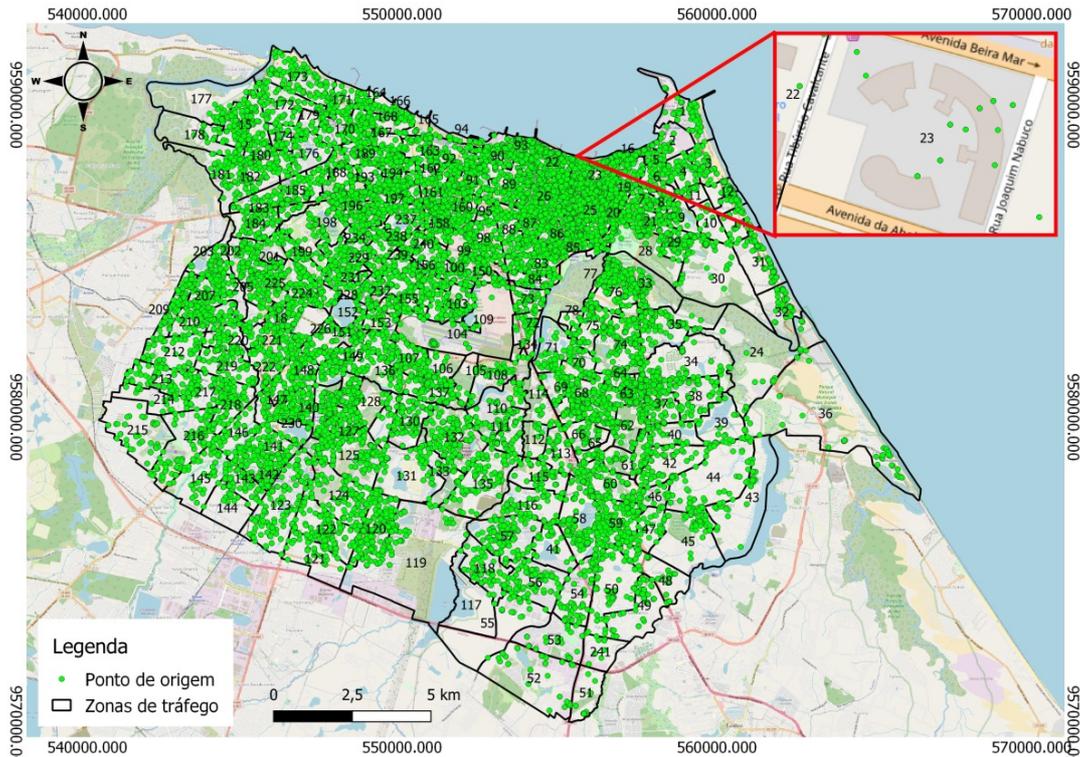


Figura 9. Distribuição espacial das origens das viagens (base residencial)

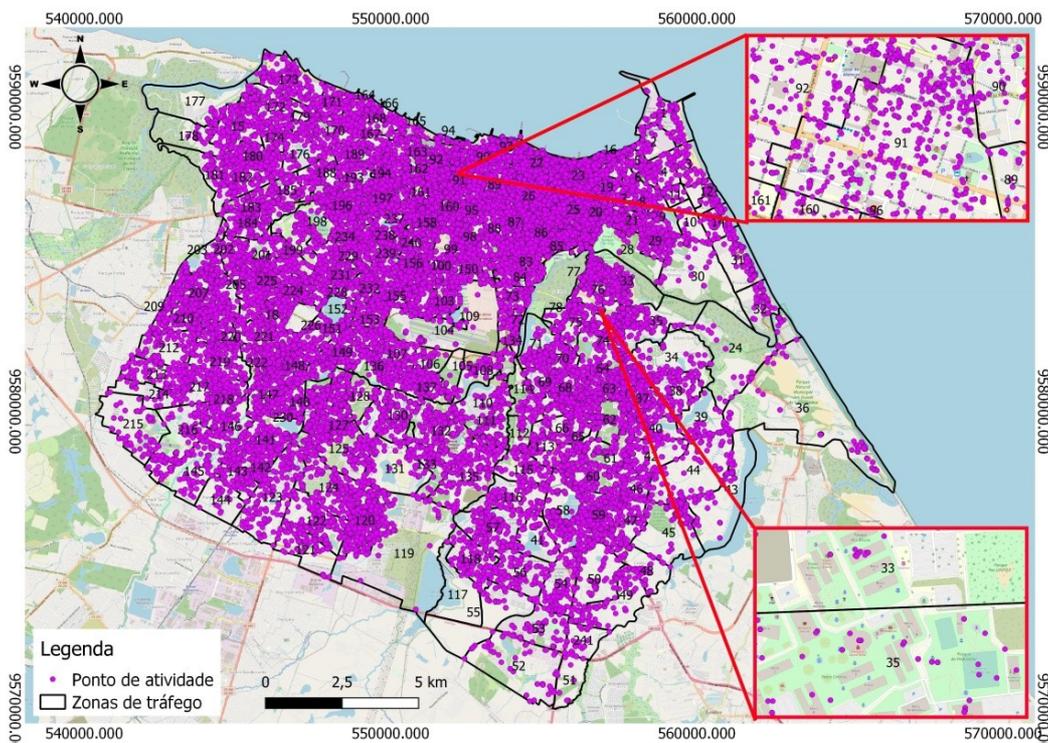


Figura 10. Distribuição espacial das atividades do tipo trabalho e educação

Já a distribuição espacial das atividades realizadas por outros motivos pode ser observada na Figura 11. Aqui, foi percebido uma concentração deste tipo de atividades nas regiões mais ricas da cidade, em especial nas zonas 33 (Edson Queiroz) e 23 (Meireles). A primeira se destaca pela presença de alguns *shopping centers*, enquanto a segunda fica localizada próximo à orla de praia, atraindo uma grande quantidade de indivíduos buscando atividades de lazer.

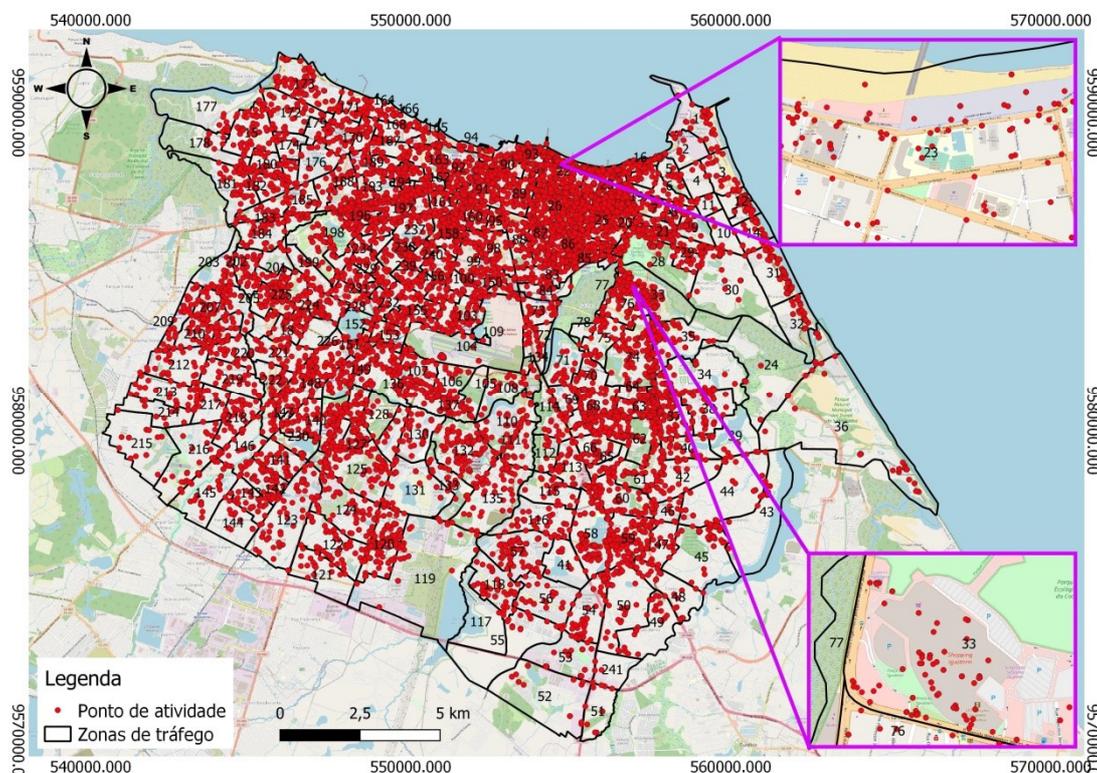


Figura 11. Distribuição espacial das atividades por outros motivos

4.3. Caracterização socioeconômica dos domicílios

Identificadas as viagens realizadas, por motivo, optou por fazer uma análise da situação socioeconômica dos domicílios dos indivíduos amostrados, com base na renda domiciliar. Portanto, primeiramente foi construído um mapa de densidades da distribuição espacial dos domicílios nos quais foram realizados *check-ins* durante o período de coleta. Essa caracterização está ilustrada no mapa da Figura 12. Visualmente é percebido que há uma concentração destes domicílios na região do centro da cidade, ao norte, com alguns aglomerados mais o sul, sudeste e sudoeste.

Analisando a autocorrelação espacial da distribuição dos domicílios Figura 13, concluiu que as maiores densidades de domicílios da amostra de *check-ins* coincidem, de forma mais significativa, com os aglomerados de domicílios de renda média e alta. Então, as variáveis explicativas como tipos de domicílios por faixa de renda foram exemplificadas através do mapa de espalhamento de Moran.

Dessa forma, as zonas *High-High* que possuem em sua maioria domicílios de renda baixa estão clusterizadas ao sudoeste, oeste e noroeste. Por outro lado, as zonas *High-High* de renda média clusterizam a noroeste, sudoeste e nordeste, e por fim, as zonas que possuem em sua maioria domicílios de renda alta estão clusterizadas a norte e nordeste exclusivamente.

Em contrapartida, ao analisar a variável renda baixa, foi observado que as zonas *Low-Low* estão a noroeste e nordeste, enquanto os domicílios de renda média que pertencem ao grupo *Low-Low* se encontram a leste exclusivamente. No entanto, as zonas *Low-Low* de renda alta se localizam a oeste, sul e leste.

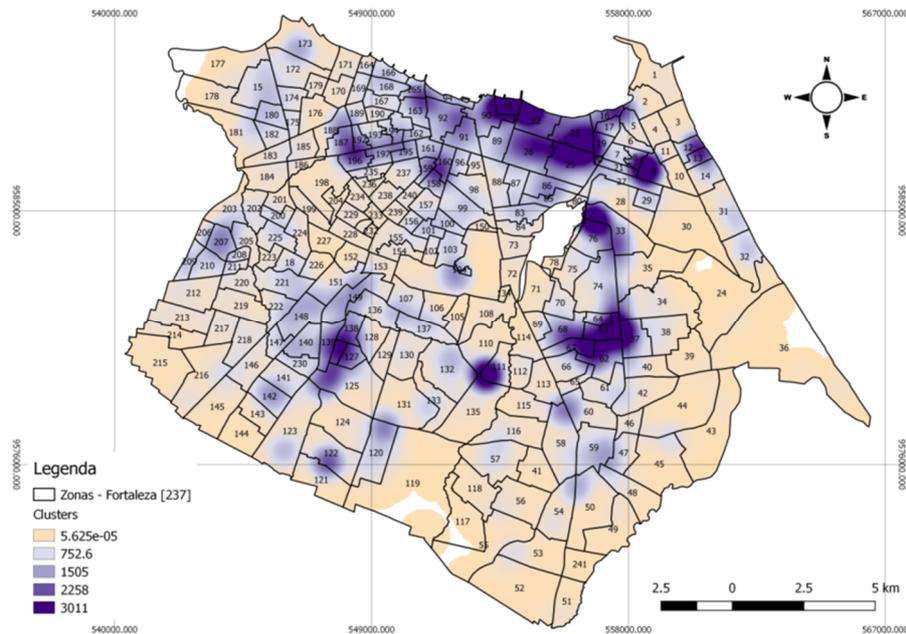


Figura 12. Mapa de densidade dos domicílios componentes da amostra

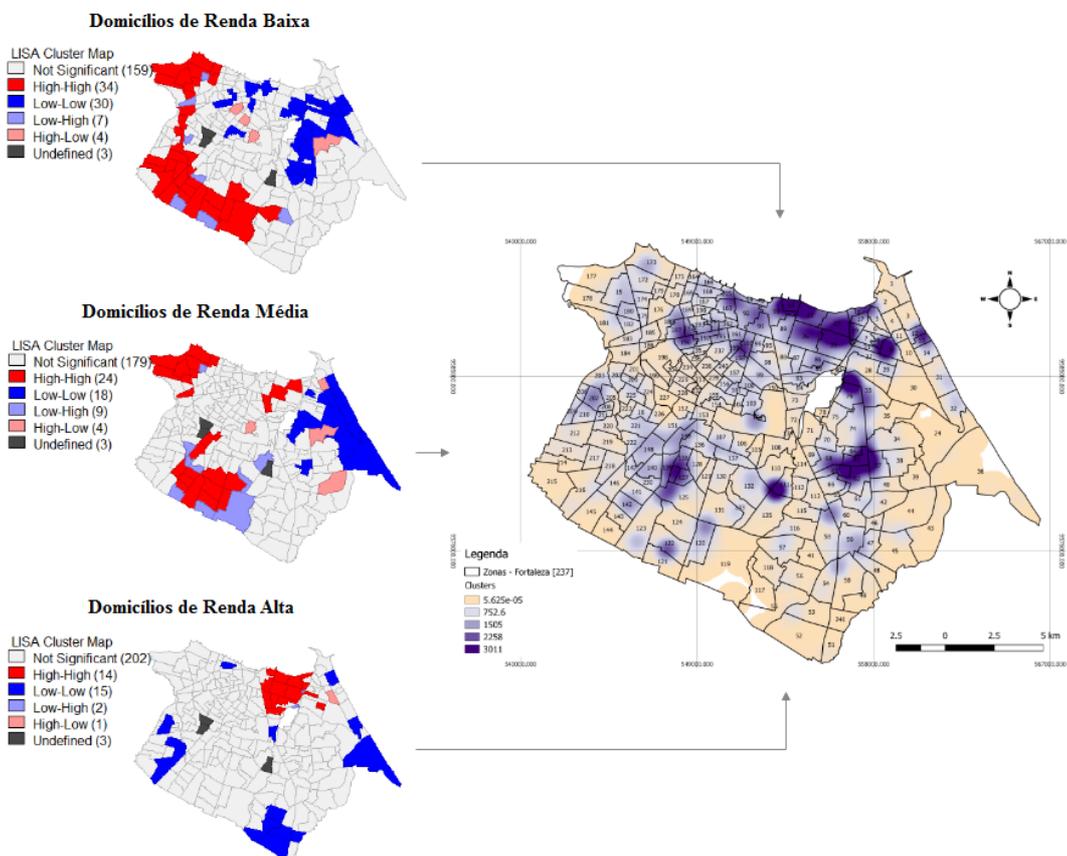


Figura 13. Correlação espacial dos domicílios, por faixa de renda

A partir dessa análise visual entre o mapa de espalhamento de Moran (*Moran Map*) e o mapa da densidade dos domicílios, ficou percebido que existe uma forte relação entre as zonas constituídas por domicílios de média e alta renda com os *clusters* de domicílios “*check-ins*” da amostra. A partir dessa observação, cabe assumir, portanto, que a amostra obtida, proveniente de dados de redes sociais, representa, de forma pouco significativa, as viagens dos grupos pertencentes a estratos de renda menores.

Para avaliar de forma quantitativa a relação entre os domicílios da amostra e a distribuição das residências por faixa de renda, foi realizada uma regressão espacial do tipo *spatial lag*. Pelos valores do coeficiente de correlação espacial R^2 , foi percebida, de fato, uma maior correlação espacial entre a localização dos domicílios da amostra com a localização dos domicílios de média e alta renda, com valores de R^2 de 0,4 e 0,5 respectivamente. Enquanto os domicílios de baixa renda apresentaram um coeficiente de correlação espacial de 0,28.

Uma das hipóteses levantadas para explicar a baixa representatividade dos domicílios de baixa renda na amostra de viagens, proveniente de dados de redes sociais, é o fato da dificuldade de acesso à *internet* por parte dos indivíduos deste grupo, apesar de que nos últimos anos houve uma maior disseminação dos celulares para a maioria da população de baixa renda. Segundo dados do IBGE de 2018, apenas 65,9% dos domicílios com renda menor ou igual a 3 salários mínimos têm *internet* em casa, seja móvel ou fixa.

5. CONCLUSÕES

Neste trabalho, foram identificados padrões de deslocamento em Fortaleza através de dados obtidos de redes sociais (*Twitter* e *Instagram*). A partir dos *check-ins* georreferenciados, realizados pelos usuários desses aplicativos, foi possível identificar suas viagens cuja base fosse o domicílio. Além disso, com o auxílio da base de dados de uso do solo desagregada ao nível do lote, foi possível ainda identificar os motivos dessas viagens, discriminadas em trabalho, educação ou outros motivos. Por fim, foi realizada uma análise de caracterização socioeconômica, com base na renda, dos domicílios que compuseram a amostra. Diante dos resultados obtidos, foi possível concluir que a base de dados provenientes de redes sociais georreferenciadas se apresenta como uma alternativa viável aos métodos tradicionais de obtenção de dados de viagens, em especial as pesquisas domiciliares, de alto custo de obtenção e estacionários no tempo.

Apesar das vantagens identificadas na utilização deste tipo de dado, foram encontradas também algumas limitações. Uma delas foi a pequena representatividade de domicílios de faixa de renda mais baixa na composição da amostra de indivíduos. Com isso, os padrões de deslocamento observados só refletem as características de uma parcela com níveis de renda mais elevados. Além disso, foi verificada uma elevada proporção de viagens do tipo outros e poucas viagens por motivos educacionais. Isso leva a crer que os usuários de redes sociais tendem, de modo geral, a realizar *check-ins* quando estão realizando atividades que não sejam compulsórias.

No entanto, o trabalho proposto possui uma limitação no que se refere à representatividade dos dados (comportamento da população como um todo). Pois, foi observada que boa parte da população ainda não tem acesso às redes sociais baseadas em localizações, ou seja, indivíduos de baixa renda. Essa limitação, no entanto, exigirá novas pesquisas. Para estudos futuros, será proposto um estudo sobre a variabilidade de viagens, além das etapas de geração e distribuição, e dessa forma, conseqüentemente, uma comparação entre a matriz OD proposta pelo método

e uma matriz externa (gerada por alguma pesquisa). Portanto, os resultados amostrais obtidos só são expansíveis em sua maioria para viagens não compulsórias de determinados grupos socioeconômicos, de faixa de renda mais elevada.

REFERÊNCIAS

- Abbasi, O., Alesheikh, A., Sharif, M. (2017) *Ranking the city: the role of location-based social media check-ins in collective human mobility prediction*. ISPRS International Journal of Geo-Information, v. 6, n. 5, p. 136. <https://doi.org/10.3390/ijgi6050136>
- Alzahrani, S., M. (2018) *Development of IoT mining machine for Twitter sentiment analysis: mining in the cloud and results on the mirror*. In: 2018 15th Learning and Technology Conference (L&T). IEEE, 2018. p. 86-95.
- Agryzkov, T., Martí, T., Tortosa, L. (2017) *Measuring urban activities using Foursquare data and network analysis: a case study of Murcia (Spain)*. International Journal of Geographical Information Science, v. 31, n. 1, p. 100-121. <https://doi.org/10.1080/13658816.2016.1188931>
- Anselin, L. (1995) *Local indicators of spatial association - LISA*. Geographical analysis, 27, pp. 93-115.
- Bifet, A., Frank, E. (2010) *Sentiment knowledge discovery in twitter streaming data*. In International conference on discovery science, pp. 1-15. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-642-16184-1_1
- Capdevila, J., Cerquides, J., Nin, J., & Torres, J. (2017) *Tweet-SCAN: An event discovery technique for geo-located tweets*. Pattern Recognition Letters, v. 93, p. 58-68.
- Cheng, Z., Caverlee, J., Lee, K., Sui, D. (2011) *Exploring millions of footprints in location sharing services*. ICWSM, pp. 81-88.
- Cheng, Z., Jian, S., Maghrebi, M., Rashidi, T., Waller, S. (2018) *Is Social Media an Appropriate Data Source to Improve Travel Demand Estimation Models?* In: Transportation Research Board.
- Cranshaw, J., Schwartz, R., Hong, J. I., Sadeh, N. (2012) *The livehoods project: Utilizing social media to understand the dynamics of a city*. Sixth International AAAI Conference on Weblogs and Social Media.
- Ferreira, J. (2013) *Mobi-System: towards an information system to support sustainable mobility with electric vehicle integration*. Doctoral thesis. Minho's University (Portugal). <http://hdl.handle.net/1822/28871>
- Gao, S., Jiue, Y., Bo, Y., Yingjie, H., Krzysztof, J. (2014) *Detecting origin-destination mobility flows from geotagged Tweets in greater Los Angeles area*. In: Eighth International Conference on Geographic Information Science (GIScience'14).
- Hanson, S., Huff, J. (1988) *Systematic variability in repetitious travel*. Transportation, 15(1-2), pp. 111-135. <https://doi.org/10.1007/BF00167983>
- Hanson, S., Huff, J. *Assessing day-to-day variability in complex travel patterns*. Transportation Research Record, 891, pp. 18-24, 1981.
- Hasan, S., Zhan, X., Ukkusuri, S. (2013) *Understanding urban human activity and mobility patterns using large-scale location-based data from online social media*. In Proceedings of the 2nd ACM SIGKDD international workshop on urban computing. p. 6. <https://doi.org/10.1145/2505821.2505823>
- Hellinga, B. (1995) *Estimating dynamic origin-destination demands from link and probe counts*. Queen's University at Kingston.
- Hu, W., Peter, J. (2017) *An adaptive Hawkes process formulation for estimating time-of-day zonal trip arrivals with location-based social networking check-in data*. Transportation Research Part C: Emerging Technologies, v. 79, p. 136-155. <https://doi.org/10.1016/j.trc.2017.02.002>
- IBGE. (2018) *Pesquisa Nacional por Amostra de Domicílios Contínua*. Diretoria de Pesquisas, Coordenação de Trabalho e Rendimento.
- Iqbal, M. D. Shahadat, C., Charisma, F., Wang, P., Gonzalez, M. (2014) *Development of origin-destination matrices using mobile phone call data*. Transportation Research Part C: Emerging Technologies, v. 40, p. 63-74. <http://hdl.handle.net/1721.1/108682>
- Lima, L. S. (2017) *Espreiamento Urbano por Autossegregação e seus Impactos na Acessibilidade Urbana de Fortaleza*. Tese de Doutorado. Dissertação. Universidade Federal do Ceará, Fortaleza - CE.
- Molloy, J., Moeckel, R. (2017) *Improving destination choice modeling using location-based big data*. ISPRS International Journal of Geo-Information, v. 6, n. 9, p. 291.
- Oliveira, S.F.C., Silva C. A. U., Silva, M. Y. O (2019). *O uso do instagram e do twitter para identificação e mapeamento de usos do solo em áreas urbanas: o caso de Fortaleza-CE*. Brazilian Journal of Development, v. 5, n. 9, p. 16128-16149. [tps://doi.org/10.34117/bjdv5n9-172](https://doi.org/10.34117/bjdv5n9-172).
- Pereira, R. H. (2013). *Tempo de deslocamento casa-trabalho no Brasil (1992-2009): diferenças entre regiões metropolitanas, níveis de renda e sexo*. Texto para Discussão, Instituto de Pesquisa Econômica Aplicada (IPEA).
- Piza, M. (2012) *O fenômeno Instagram: considerações sob a perspectiva tecnológica*. Monografia (Graduação em Sociologia). Universidade de Brasília, Brasília.
- Ramos, F. R. (2002) *Análise espacial de estruturas intra-urbanas: o caso de São Paulo*. São José dos Campos: INPE.
- Wolf, J., Oliveira, M., Thompson, M. (2003) *Impact of underreporting on mileage and travel time estimates: Results from global positioning system-enhanced household travel survey*. Transportation Research Record: Journal of the Transportation Research Board, n.1854, p. 189-198. <https://doi.org/10.3141/1854-21>.