UNIVERSIDADE FEDERAL DE CAMPINA GRANDE
CENTRO DE ENGENHARIA ELÉTRICA E INFORMÁTICA
UNIDADE ACADÊMICA DE SISTEMAS E COMPUTAÇÃO
CURSO DE BACHARELADO EM CIÊNCIA DA COMPUTAÇÃO

THIERRY SILVA BARROS

ANÁLISE COMPARATIVA ENTRE DIFERENTES FREQUÊNCIAS DE COLETA DE DADOS GEOLOCALIZADOS

CAMPINA GRANDE - PB

2019

THIERRY SILVA BARROS

ANÁLISE COMPARATIVA ENTRE DIFERENTES FREQUÊNCIAS DE COLETA DE DADOS GEOLOCALIZADOS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

Orientador: Professor Dr. Claudio Elízio Calazans Campelo.

CAMPINA GRANDE - PB

2019

B277a Barros, Thierry Silva.
Análise comparativa entre diferentes frequências de coleta de dados geolocalizados. / Thierry Silva Barros. – 2019.

12 f.

Orientador: Prof. Dr. Claudio Elízio Calazans Campelo.
Trabalho de Conclusão de Curso - Artigo (Curso de Bacharelado em Ciência da Computação) - Universidade Federal de Campina Grande; Centro de Engenharia Elétrica e Informática.

1. Geolocalização. 2. Qualidade de dados. 3. Granularidade da informação. 4. Consumo de recursos computacionais. I. Campelo, Claudio Elízio Calazans. II. Título.

CDU:004(045)

Elaboração da Ficha Catalográfica:

Johnny Rodrigues Barbosa
Bibliotecário-Documentalista
CRB-15/626

THIERRY SILVA BARROS

ANÁLISE COMPARATIVA ENTRE DIFERENTES FREQUÊNCIAS DE COLETA DE DADOS GEOLOCALIZADOS

Trabalho de Conclusão Curso apresentado ao Curso Bacharelado em Ciência da Computação do Centro de Engenharia Elétrica e Informática da Universidade Federal de Campina Grande, como requisito parcial para obtenção do título de Bacharel em Ciência da Computação.

BANCA EXAMINADORA:

Professor Dr. Claudio Elízio Calazans Campelo
Orientador – UASC/CEEI/UFCG

Professor Dr. Reinaldo Cézar de Morais Gomes
Examinador – UASC/CEEI/UFCG

Professor Dr. Tiago Lima Massoni
Disciplina TCC – UASC/CEEI/UFCG

Trabalho aprovado em: 25 de novembro 2019.

CAMPINA GRANDE - PB

Análise comparativa entre diferentes frequências de coleta de dados geolocalizados

Trabalho de Conclusão de Curso

Thierry Silva Barros (Aluno), Cláudio Campelo (Orientador)
Departamento de Sistemas e Computação
Universidade Federal de Campina Grande
Campina Grande, Paraíba - Brasil

ABSTRACT

Geolocation data has been widely used for the comprehension of various social phenomena. Nowadays, such data are produced at large scale by people, especially by their smartphones equipped with various types of sensors. However, the capture os this kind of data, in a mobile device, can be expensive as the collection rate increases and may consume device resources. On the other hand, the capture with low frequency may impair the quality and consistency of the information collected. In this context, we conducted a comparative of performance across different data collection frequencies (of 15, 30, 60 and 120 seconds) to analyze the impact on resource consumption and data quality. For this study, we used an application developed in a previous work [1]. This app is able to collect geolocated data from several sensors embedded in the smartphone. The experiment had 10 volunteers, who used the application for a period of 20 days. Afterwards, an evaluation was performed taking into account aspects related to the device resource consumption and aspects regarding the quality of the data captured, in order to show the pros and cons of the different capture frequencies and estimate a frequency best suited for different usage scenarios. As a result, the 30-second frequency collection had the best tradeoff between resource consumption and information generation, and the 60-second frequency collection was the only one that showed no better effectiveness in any specific application compared to the other frequencies.

RESUMO

Dados geolocalizados têm sido utilizados na compreensão de vários fenômenos sociais. Hoje, tais dados têm sua produção realizada em larga escala, especialmente por dispositivos móveis. Porém, a captação desse tipo de dado, em um dispositivo móvel, pode ser custosa, à medida em que se aumenta a frequência de coleta, podendo consumir recursos do aparelho. Por outro lado, a captação com uma baixa frequência pode prejudicar a qualidade e consistência da informação coletada. Nesse contexto, este artigo apresenta um estudo comparativo de desempenho entre diferentes frequências de coleta de dados (15, 30, 60 e 120 segundos), para analisar o impacto no consumo de recursos e na qualidade dos dados captados. Para esse estudo, foi utilizado um aplicativo desenvolvido em um trabalho anterior [1], que tem como funcionalidade a coleta de dados geolocalizados e de sensores embarcados nos *smartphones*. O experimento contou com 10 voluntários, que sujeitaram-se a utilização do aplicativo por um período de 20 dias. Após a captação dos dados, foi realizado uma avaliação levando em consideração aspectos relacionados ao consumo de recursos do dispositivo e aspectos sobre a qualidade dos dados captados, com o objetivo de mostrar os prós e contras das diferentes frequências de captação e estimar uma frequência mais adequada para diferentes cenários de uso. Como resultado, a coleta com frequência de 30 segundos obteve o melhor *tradeoff* entre o consumo de recursos e a geração de informações, e a coleta com frequência de 60 segundos foi a única que não demonstrou nenhuma melhor eficácia em alguma aplicação específica, comparada com as outras frequências.[1]

KEYWORDS

Geolocalização, qualidade de dados, granularidade de informação, *smartphone*, consumo de recursos computacionais, análise de desempenho.

1. INTRODUÇÃO

Geolocalização é a identificação da localização geográfica de um objeto no mundo real. Geralmente é uma informação que pode ser

usada para identificar a localização física de um dispositivo eletrônico. Dados geolocalizados têm sido utilizados na compreensão de vários fenômenos sociais. Atualmente, a geração desses dados acontece em larga escala por pessoas, através de seus *smartphones* equipados com diversos tipos de sensores. Exemplos de dados gerados por *smartphones* que podem conter referências a localizações incluem fotos, vídeos e postagens em redes sociais. Além disso, diversos aplicativos capturam os dados da localização do usuário com certa frequência, com o intuito de oferecer produtos e serviços direcionados e específicos, baseados em suas localizações.

Esses tipos de dados também são objeto de estudo de diversos pesquisadores, que os utilizam para realizar pesquisas de grande impacto social. São dados utilizados em diversos tipos de estudos, tais como aqueles relacionados a políticas de transportes e segurança pública, engenharia de tráfego e outros aspectos associados ao planejamento de cidades. Outra disposição desses tipos de dados está relacionada às trajetórias desempenhadas pelas pessoas, particularmente nos centros urbanos. Dados que estão relacionados à identificação de padrões de mobilidade urbana, os quais expressam características de comportamento humano, possibilitando estudos em diferentes regiões e cidades do mundo, com intuito de identificar esses tipos de padrões [8]. Além disso, serviços baseados em localização têm utilizado esses dados para prever a trajetória dos usuários e poder recomendar novas localizações baseadas na localização atual dos mesmos [4].

Embora os avanços na tecnologia móvel tenham tornado a capacidade dos dispositivos cada vez maiores, tanto em relação aos recursos de processamento quanto aos de dados e bateria, esses recursos continuam sendo limitados. Dentro dessa perspectiva, uma preocupação que tem sido gravemente ignorada é o custo do consumo desses recursos para a captação de dados geolocalizados. A captação com frequência elevada pode ocasionar um impacto significativo nesse consumo, diminuindo o desempenho dos dispositivos e gerando desconforto aos usuários do aplicativo. Por outro lado, a captação com uma baixa frequência pode ocasionar a perda de informação crucial para pesquisas ou empresas que utilizam esse tipo de informação, produzindo resultados com maior nível de imprecisão ou incerteza, quando comparado ao que poderia ser obtido se houvesse uma maior granularidade da informação.

Um aplicativo que capta dados geolocalizados com frequência de 60 segundos, por exemplo, pode perder informações cruciais sobre a locomoção do usuário, nesse intervalo de tempo, e oferecer serviços e informações inadequadas, que não seriam bem aproveitadas pelo usuário. Por outro lado, um aplicativo com alta frequência de coleta pode drenar consideravelmente a bateria do dispositivo do usuário, ou deixá-lo mais lento. Decidir o tempo de frequência para coleta nem sempre é uma tarefa fácil, visto que é necessário analisar o impacto que isso pode ter para o usuário do aplicativo e também o impacto que pode ter no nível de detalhamento dos dados captados. Geralmente o que é decidido é um valor arbitrário que o próprio desenvolvedor do sistema atribui, sem maiores fundamentações teóricas e experimentais.

Sendo assim, neste trabalho, comparou-se a eficácia e a eficiência da utilização de diferentes frequências de coleta de dados com o intuito de observar os prós e contras de cada frequência, e poder estimar uma frequência mais adequada para diferentes cenários de uso. Para esse estudo, foi utilizado um aplicativo capaz de coletar dados geolocalizados e de sensores embarcados do *smartphone*. Foram analisadas capturas em diferentes frequências de coleta e comparadas com base em seus consumos dos recursos do *smartphone* e perda de informação geolocalizada, com a finalidade de verificar se a frequência de coleta tem impacto significativo nessas duas perspectivas e se a alguma das frequências se mostra mais eficiente dentro desse contexto. Com essa comparação, foi realizada uma análise detalhada sobre o *tradeoff* entre a granularidade de coleta e a eficácia na utilização dos dados em algoritmos de análise de dados geolocalizados. O principal algoritmo utilizado para a análise comparativa foi o Dynamic Time Warping [11], um algoritmo para comparar e alinhar duas séries temporais comumente utilizado em pesquisas de geolocalização para comparar a similaridade entre trajetórias. Os outros algoritmos utilizados foram desenvolvidos pelo laboratório de pesquisa, como por exemplo, o algoritmo para encontrar regiões de parada que é uma variação do DBSCAN (density-based spatial clustering of applications with noise [6]. Algoritmos para geração de informações a partir dos dados coletados, como por exemplo, iluminação em determinados locais, regiões em que o usuário ficou parado um determinado tempo e sobre o uso do smartphone, foram desenvolvidos pelo próprio pesquisador.[2]

Participaram do processo de captação dos dados 10 voluntários, durante um período de 20 dias, utilizando o aplicativo durante 5 dias consecutivos, compreendendo apenas dias da semana, para cada frequência de coleta. Ao final de cada dia, os voluntários forneciam os dados relacionados ao consumo de recursos do *smartphone*. O resultado obtido nesta pesquisa pode auxiliar times de desenvolvimento e pesquisadores em suas decisões de qual frequência de coleta utilizar em determinada situação. Por fim, dependendo da situação uma frequência, em específico, pode ser mais indicada que as outras para coleta.

O restante deste artigo está estruturado da seguinte maneira. Na Seção 2, discutimos a metodologia da pesquisa proposta neste trabalho. Em seguida, na Seção 3, apresentam-se os resultados obtidos. Por fim, a Seção 4 conclui o artigo e aponta para trabalhos futuros.

[2] Disponível em https://github.com/ThierryBarros/tcc-algorithms

2. METODOLOGIA

Esta seção descreve a metodologia que foi adotada para analisar o *tradeoff* entre as diferentes frequências de coleta de dados geolocalizados e a qualidade das informações derivadas desses dados. Nessa seção, serão apresentados tópicos relacionados aos critérios utilizados para a análise das informações, à estratégia utilizada para captação dos dados e à geração e análise das informações.

2.1 Critérios utilizados na análise comparativa

Para que fosse possível analisar o *tradeoff* entre a granularidade de coleta e a eficácia na utilização dos dados em algoritmos de análise de dados, foram utilizados indicadores relacionados ao consumo de recursos e à qualidade das informações produzidas. Os três indicadores relacionados ao consumo de recursos são:

Consumo de bateria: Este é um indicador referente ao consumo de energia do aplicativo no aparelho. *Smartphones* mais modernos fornecem a informação do consumo de bateria por aplicativo, ou seja, o valor do gasto de bateria por um aplicativo não é enviesado pelo uso de outros aplicativos no *smartphone*. O alto consumo de bateria é um dos principais fatores de resistência para para utilização de aplicativos que realizam coleta de dados de geolocalização..

Consumo de memória RAM: Este indicador é referente ao consumo de memória do aplicativo no aparelho. O alto consumo de memória RAM pode deixar o *smartphone* mais lento, diminuindo assim a capacidade de resposta do aparelho e gerando desconforto ao usuário.

Quantidade de dados transmitidos: Este indicador é referente à quantidade de dados enviados pelo aplicativo a um servidor em nuvem. Altas taxas de transmissão de dados podem impactar diretamente no valor monetário pago por usuários e reduzindo o interesse pelo uso do aplicativo

Então, quando há necessidade de utilização de usuários para prover dados mais detalhados sobre suas localizações, os pesquisadores se deparam com essas dificuldades. Sendo assim, podemos perceber que esses são bons indicadores pois estão diretamente relacionados à resistência das pessoas em se voluntariarem para pesquisas que utilizam aplicativos para captação e análise de dados geolocalizados. Por este motivo, pesquisas de extrema relevância têm sido conduzidas utilizando bases de dados consideravelmente reduzidas, com poucos voluntários. Problemas como esses têm sido apontados por diversos autores [2, 10, 13].

No que se diz respeito à mensuração da qualidade da informação produzida, foram adotados dois indicadores relacionados aos dados geoespaciais (fornecidos pelo sensor de GPS do dispositivo) e outros relacionados a diferentes sensores do *smartphone*. Os dois indicadores baseados em dados obtidos do sensor GPS são a **trajetória** (percurso realizado pelos usuários) e as **regiões de parada** (locais da trajetória onde o usuário ficou parado ou não se distanciou muito por um determinado tempo). O segundo indicador está mais relacionado a aspectos mais semânticos da trajetória, podendo ser usado para inferir tipos locais de interesse do usuário [12]. Esses são bons indicadores para avaliar a qualidade dos dados, pois diversos pesquisadores utilizam essas informações para produzir estudos de grande impacto social, como, por exemplo, predição de trajetórias dos usuários, reconhecimento de atividades humanas, identificação de padrões de mobilidade urbana e também para produção de pesquisas que analisam esses dados para inferir locais turísticos [3, 8, 9]. Outras pesquisas usam esses dados para estudos mais semânticos, como, por exemplo, para inferir regiões de interesse dos usuários ou atividades que eles realizaram [14]. Além disso, a qualidade da informação produzida através desse indicadores está intrinsecamente ligada com os resultados obtidos nessas pesquisas.

Além dos indicadores citados anteriormente, foram selecionados indicadores relacionados às informações dos sensores de iluminação do ambiente e a proximidade que o usuário estava do *smartphone*. Também foram selecionados indicadores do uso do *smartphone*: se a tela estava bloqueada ou desbloqueada e se o áudio estava no modo normal, mudo ou no modo silencioso.

2.2 Estratégia de captação dos dados

Para a realização desse estudo, foi utilizado um aplicativo desenvolvido no laboratório de pesquisa [1], que tem como funcionalidade a coleta de dados geolocalizados e de sensores embarcados do smartphone.

A pesquisa foi dividida em três etapas. A primeira etapa foi a captação dos dados do consumo de recursos do dispositivo móvel e dos dados geolocalizados. Para realização desta etapa, contamos com a participação de 10 voluntários recrutados pelo próprio pesquisador. O tipos de mobilidade utilizada pelos voluntários para se locomover entre regiões da cidade, foram veículos ou ônibus. Cada um dos voluntários foi submetido a um estudo observacional onde eles utilizaram em seu dispositivo o aplicativo para coleta de dados durante um período de 20 dias. A captação foi realizada observando o consumo de recursos do *smartphones* a partir de 4 diferentes frequências de coleta de dados: 15, 30, 60 e 120 segundos. Cada frequência foi observada por um período de 5 dias consecutivos, para poder estimar com maior precisão o intervalo de confiança do consumo de recursos de cada frequência de coleta de dados. A coleta dos dados foi realizada de maneira manual, pois o aplicativo não possui funcionalidades para coleta dos dados do consumo de maneira automática. Ao final de cada dia, os voluntários informavam, através de aplicativo de troca de mensagens, os dados do consumo de cada recurso. Antes do início da fase de aquisição dos dados,

os voluntários foram treinados sobre como obter os dados de consumo e sobre a formatação do relatório diário. Ao final da captação foram produzidos 4 *datasets*, um para cada frequência de coleta, com as informações geolocalizadas de cada frequência de coleta, e 3 tabelas, uma para cada recurso, com as informações do consumo de cada recurso por frequência.

2.3 Geração das informações

A segunda etapa consistiu na geração de informações a partir dos dados captados. Para geração das informações dos recursos, foram calculados os intervalos de confiança de cada recurso por frequência de coleta. Essa estimativa foi utilizada para poder comparar se existe uma diferença significativa no custo, em relação ao consumo dos recursos, das diferentes frequências de coleta. Para calcular o intervalo de confiança, primeiro foi realizada a média dos 5 dias por cada usuário para, a partir disso, ser calculado o desvio padrão amostral e assim poder obter os intervalos de confianças.

Para a geração das informações geolocalizadas foi utilizado apenas o *dataset* com frequência de 15 segundos, que foi captado nos primeiros 5 dias de coleta, pois é único *dataset* que, a partir del*e*, pode ser simulado *datasets* com frequência de 30, 60 e 120 segundos. Essa metodologia foi adotada para que fosse possível manter o mesmo contexto das informações geradas. Pois, utilizando os diferentes *datasets* produzidos, o contexto muda e as informações produzidas podem ser diferentes. Sendo assim, não seria possível comparar a similaridade entre as informações produzidas. Além disso, a análise comparativa foi realizada utilizando a perda de informação geolocalizada, que foi calculada através do nível de similaridade entre as informações geradas em cada *dataset,* por isso a importância de manter o mesmo contexto. Então, a partir do *dataset* com frequência de 15 segundos, foram simulados *datasets* com frequências de 30, 60 e 120 segundos, apenas removendo valores do conjunto de dados original. A simulação dos conjuntos de dados foi realizada de maneira simples. Como o conjunto original captou dados com uma frequência de 15 segundos, para simular um conjunto de dados com frequência de 30 segundos, é necessário apenas remover metade dos valores do conjunto, ou seja, a cada dois valores em sequência do conjunto original apenas o último valor era mantido. O mesmo se aplica a frequência de 60 segundos, removendo 3 quartos dos valores, e a frequência de 120 segundos, removendo 7 oitavos.

Após a simulação dos *datasets*, foram executados algoritmos sobre os dados para geração das informações geolocalizadas. Os algoritmos produziram informações sobre: identificação de trajetória realizada por usuários, regiões de parada, informações dos sensores embarcados e do uso do smartphone.

2.4 Métricas e análise das informações

Por fim, na última etapa foi realizada a análise das informações geradas na etapa anterior, para poder comparar o *tradeoff* entre o consumo de recursos do *smartphone* e a qualidade dos dados geolocalizados.

Para a análise do consumo de recursos foi realizado apenas um estudo comparativo entre as valores dos intervalos de confiança do consumo de cada recurso por frequência de coleta, para avaliar se existe uma diferença significativa entre os consumos de cada frequência.

Após a análise do consumo de cada recurso do *smartphone* por frequência, foi realizada uma análise comparativa de cada conjunto de dados baseado no nível de similaridade entre as informações produzidas por cada conjunto simulado, comparado com a informações produzidas pelo conjunto original. Sendo assim, foi possível estimar e quantificar o impacto que cada frequência teve na perda de informações geolocalizadas.

A métrica utilizada para comparar a qualidade dos dados geolocalizados em relação às trajetórias produzidas foi o nível de dissimilaridade utilizando o algoritmo *Dynamic Time Warping* [5]. Este algoritmo fornece um valor numérico relacionado à medida de distância entre as trajetórias, que pode ser interpretado como o nível de dissimilaridade, ou perda de informação, entre as trajetórias. Para a identificação das regiões de parada, o algoritmo utilizado foi desenvolvido pelo próprio laboratório de pesquisa. Além disso, para a qualidade das informações das regiões de parada, a métrica utilizada foi calcular a similaridade através do número de regiões identificadas pelo conjunto de dados simulado comparado com o número de regiões identificadas pelo conjunto de dados original. Uma métrica similar foi adotada para poder calcular a similaridade em relação ao uso do *smartphone*. Por exemplo, levando em consideração o conjunto original comparado com o conjunto simulado de 30 segundos, para fazer o cálculo da similaridade entre as informações da tela desligada ou da tela ligada, basta apenas verificar, em cada par, em seqüência, do *dataset* original (que foi mapeado para um único valor do conjunto simulado), a quantidade de ocasiões em que a tela estava desligada e comparar com a quantidade de ocasiões que a tela estava desligada no conjunto simulado, uma vez que cada par de valor do conjunto original é mapeado apenas para um valor no conjunto simulado. Essa estratégia também foi aplicada para calcular a perda de informação em relação ao áudio do smartphone. Para calcular a similaridade para as outras informações de iluminação do ambiente, uso do *smartphone* e identificação do modo de áudio, a métrica utilizada foi, calcular a média de cada sequência de valores do conjunto original, que foi mapeado para o conjunto simulado, e calcular a diferença, em porcentagem, em relação ao valor do conjunto simulado, a soma de todas as diferenças é o valor da dissimilaridade, quanto maior

for esse valor, maior é a perda de informação. A próxima seção apresenta os resultados obtidos.

3. RESULTADOS E DISCUSSÕES

A partir da aplicação da metodologia mostrada anteriormente, realizou-se a análise comparativa dos consumos de cada recurso por frequência de coleta, que foi realizado através dos intervalos de confiança. Além disso, obteve-se também à análise comparativa dos dados geolocalizados comparando o conjunto de dados com maior granularidade (conjunto de dados com tempo de frequência de 15 segundos), com os conjuntos de dados simulados (conjuntos de dados de 30, 60 e 120 segundos que foram extraídos do conjunto de dados com frequência de 15 segundos). Por fim, foram realizadas análises comparativas para os sensores de proximidade e iluminação do ambiente, e para os indicadores do uso do *smartphone*: se a tela estava bloqueada ou desbloqueada e se o áudio estava no modo normal, mudo ou no modo silencioso.

Na Figura 1, é possível perceber que o coleta de dados com frequência de 15 segundos teve o maior consumo em todos os recursos. Por outro lado, o teste estatístico do intervalo de confiança, indicou que não houve diferença significativa no consumo de nenhum dos recursos, entre as frequências de coleta de 30 e 60 segundos, pois em todos os casos seus intervalos de confiança tiveram intersecção de valores. Isto mostra que utilizar a frequência de coleta de 60 segundos não apresenta ganhos significativos, no consumo de recursos, em relação a frequência de coleta de 30 segundos. Em relação à frequência de coleta de 120 segundos, ela teve o melhor desempenho, ou seja, um consumo menor, para os recursos de dados e bateria, comparado com todas as outras frequência analisadas. Porém, não houve diferença significativa para o consumo de processamento comparado com as frequências de 30 e 60 segundos. Sendo assim, foi possível perceber que, para o consumo de recursos, em determinados casos, não existe vantagens em se utilizar frequências de dados com menor granularidade de coleta.

Nas Figuras 2 e 3, podemos ver a quantidade de regiões de parada que foram identificadas por frequência de coleta. Essas regiões de parada levaram em consideração um raio de distância de no máximo 50 metros, e um tempo mínimo de 5 minutos, ou seja, para que um região de parada fosse identificada o usuário não poderia se distanciar mais do que 50 metros da região e ficar, pelo menos, 5 minutos naquela região. O conjunto de dados com frequência de coleta de 30 segundos conseguiu captar em média 90% de todas as regiões de parada, enquanto que o conjunto com frequência de 60 segundos captou em média 80% e o conjunto com frequência de 120 captou apenas 70% das regiões de parada. Sendo assim, podemos perceber um aumento gradativo do número de regiões de paradas que não foram identificadas pelos conjuntos de dados simulados. Para frequência de 30 segundos, uma perda de apenas 10% do número de regiões de parada, para alguns cenários de uso, pode não ser algo tão problemático. Por outro lado, utilizar uma frequência de coleta de 120 segundos e perder, em média, 30% das regiões pode representar uma lacuna significativa em pesquisas que utilizam esse tipo de informação e nos seus resultados obtidos.

Na Figura 4, é possível perceber a perda de informação das trajetórias produzidas pelos conjuntos de dados simulados com frequências de coleta de 30, 60 e 120 segundos, comparado com a trajetória real produzida pelo conjunto de dados com frequência de coleta de 15 segundos. A perda de informação de trajetória para o conjunto de dados com frequência de 60 segundos comparado com o conjunto de dados com frequência de 30 segundos, cresceu em média 30%. Para a frequência de coleta de 120 segundos comparado com a de 30, a perda de informação cresceu em média em 46%, indicando um crescimento significativo na perda de informações das trajetórias produzidas pelos conjuntos de dados simulados. Além disso, como esse valor da perda de informação foi obtido levando em consideração os 5 primeiros dias de captação, em alguns momentos, como por exemplo, horário em que os voluntários estavam dormindo, eles não se movimentaram por um longo período de tempo, o que influenciou no valor médio da perda de informação, ou seja, se levássemos em conta apenas momentos em que o usuário estava se locomovendo a perda de informação, em média, poderia ser ainda maior.

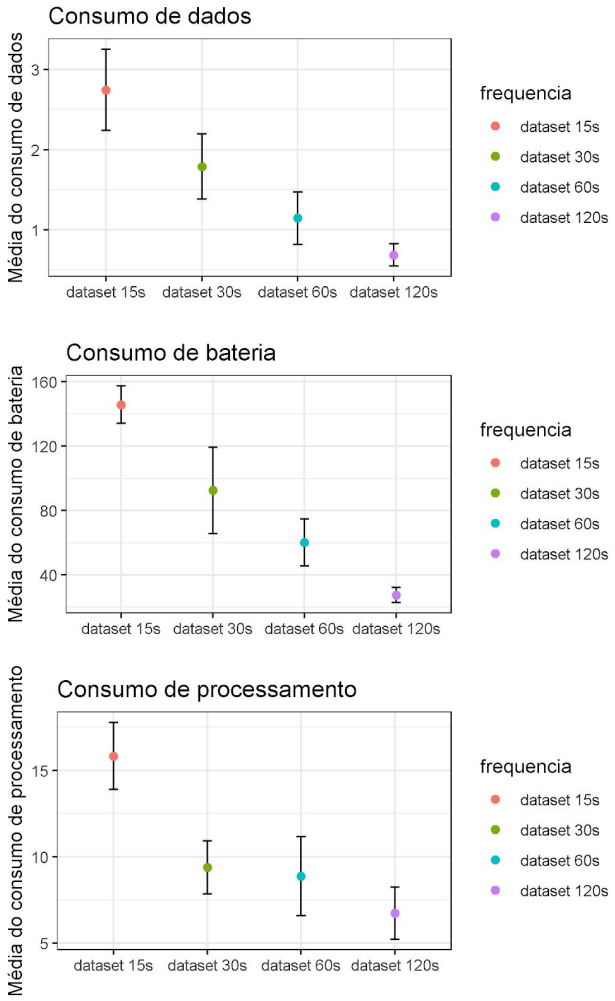

Figura 1: Gráfico dos intervalos de confiança de cada consumo de recurso por frequência de coleta

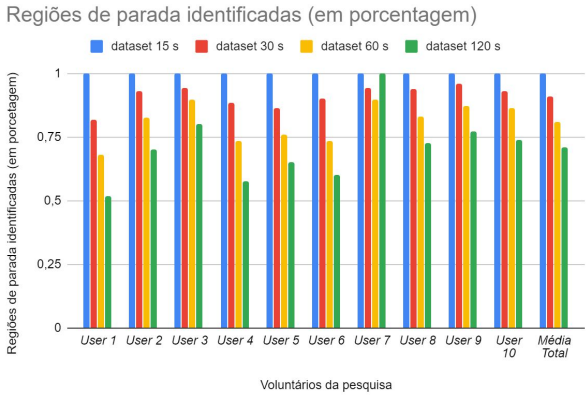

Figura 2: Gráfico da quantidade de regiões de paradas identificadas pelos conjuntos de dados original e simulado

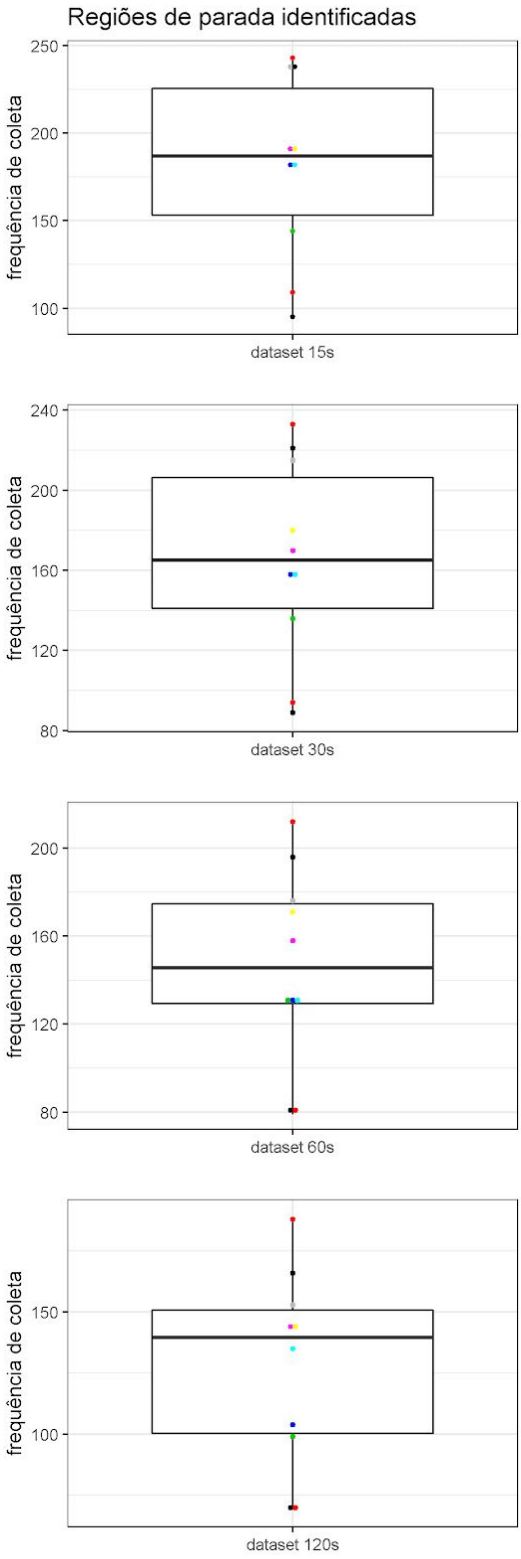

Figura 3: Gráfico dos boxplots dos valores de regiões de parada captadas por frequência de coleta

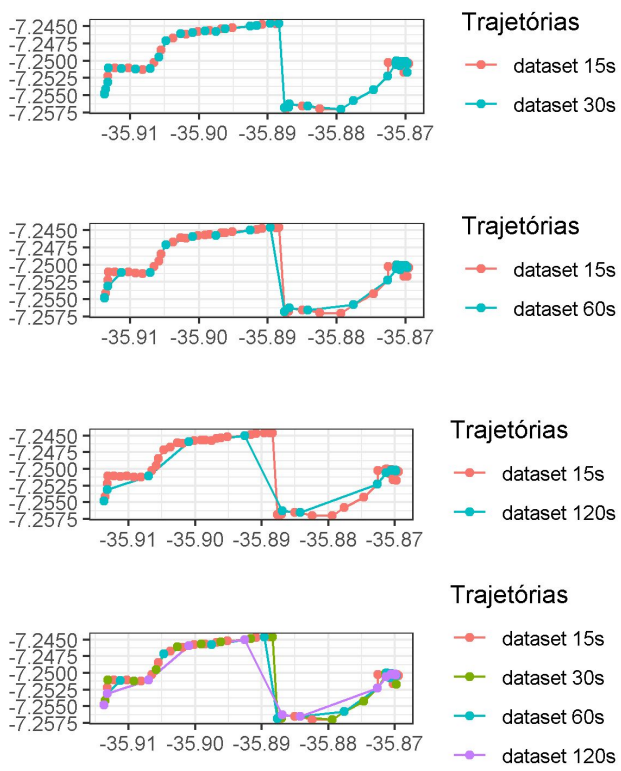

Figura 4: Gráficos de uma trajetória produzida pelo conjunto de dados com frequência de 15 segundos, comparado com a mesma trajetória produzida pelos conjuntos de dados simulados

As médias de acerto da informação captada em relação ao sensor de iluminação do ambiente são exibidas na Figura 5. Como se trata de um indicador com uma grande variação do seu valor ao longo do tempo, a perda de informação foi significativamente grande para os conjuntos de dados simulados, comparado com o conjunto de dados real. Esse valor da média de acerto da informação foi calculado através da diferença média, em porcentagem, do valor estimado pelo conjunto de dados simulado comparado com o valor real do conjunto de dados original. Os conjuntos de dados com frequência de 30, 60 e 120 segundos tiveram, em média, um acerto de 80%, 66%, 55% do valor real, respectivamente, indicando uma perda de informação significativa. Além disso, se retirarmos da análise os dados em horários que o usuário estava dormindo, onde a iluminação se mantém praticamente constante por um longo período, a perda de informação é ainda maior. Para pesquisas que utilizam esse tipo de informação, essa diferença pode representar um erro, de estimativa, significativo na pesquisa. Por exemplo, um pesquisador pode tentar determinar se o usuário estava em um ambiente de trabalho ou não, através do valor da iluminação, pois a Norma Brasileira[3], determina que nos escritórios e demais ambientes de trabalho os valores ideais de iluminância devem ser de 500 a 1000 lux. Portanto, se o for captado um dado de iluminação com uma taxa de acerto de apenas 55% do valor real, o valor estimado pode estar, facilmente, fora dessa faixa de valores indicando uma informação possivelmente equivocada, de que o usuário não estaria em um ambiente de trabalho.

O outro sensor captado, foi o sensor de proximidade, que capta a distância que a parte frontal do *smartphone* está de um determinado objeto. Neste caso, o erro médio dos conjuntos de dados simulados foi menor comparado ao erro do sensor de iluminação. Isso se deve, principalmente, por ser um dado que tem menor variação ao longo do tempo. A média de acerto para as coletas com frequência de 30, 60 e 120 segundos foram de 89%, 83% e 76%, respectivamente. Esse dado pode ser importante, por exemplo, para determinar situações em que a tela do *smartphone* estava próxima do rosto do usuário, indicando que ele poderia estar em uma possível ligação.

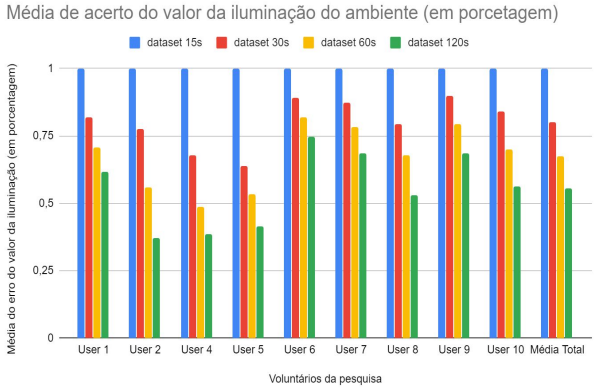

Figura 5: Gráfico da média dos valores da iluminação estimado pelos conjuntos de dados simulados comparado com os valores reais do conjunto de dados original

Por fim, na Figura 6, é possível observamos a média de acerto em relação à informação do uso do *smartphone*, se a tela estava bloqueada ou desbloqueada. Em média o valor de acerto, foi bem alto, para todos os conjuntos de dados simulados, ficando sempre perto dos 90%. Isso se deve ao fato dos usuários, em geral, deixarem a tela bloqueada por um período de tempo superior a 1 ou 2 minutos, ou utilizar o *smartphone*, com a tela desbloqueada, por um longo período de tempo. O outro dado captado, em relação ao uso do *smartphone*, foi o modo do áudio (se estava no modo normal, mudo ou silencioso). Da mesma forma, como podemos ver na Figura 7, a média de acerto também

[3] Norma NBR 5413 – Iluminância de Interiores http://ftp.demec.ufpr.br/disciplinas/TM802/NBR5413.pdf

foi bem alta, ficando, em média, por volta dos 95% para todos os conjuntos de dados simulados. Estes resultados demonstram, portanto, que a frequência de coleta não possui impacto significativo na perda de informação do uso do *smartphone*.

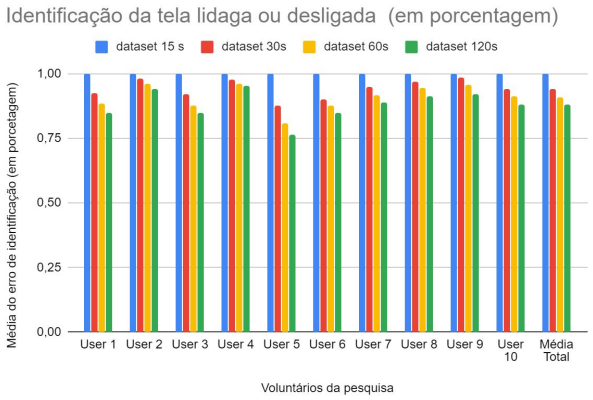

Figura 6: Gráfico da média de acerto da identificação da tela ligada ou desligada

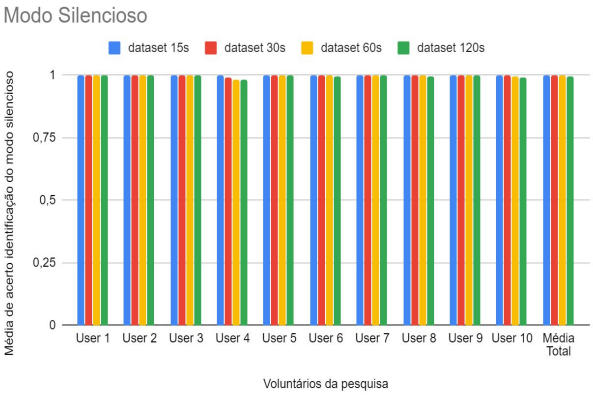

Figura 7: Gráfico da média de acerto da identificação do modo de áudio silencioso

4. CONCLUSÃO

A frequência de coleta de dados geolocalizados pode impactar diretamente no desempenho dos smartphones e também na qualidade de aplicativos que utilizam esses tipos de dados para diferentes necessidades. O resultado obtido pela análise comparativa apresentada neste artigo é um importante artefato para subsidiar times de desenvolvimento e pesquisadores em suas decisões quanto à frequência de coleta de dados geolocalizados, de modo que mantenham a satisfação dos usuários sem comprometer a qualidade dos dados captados. Neste trabalho foi possível observar as vantagens e desvantagens que cada frequência de coleta apresentou em comparação às outras. Em relação ao consumo de recursos, como já era esperado, a frequência com maior granularidade (frequência de coleta 15 segundos) obteve o pior desempenho, ou seja, o maior custo no consumo de recursos. Porém, as frequências de 30 e 60 segundos não mostraram diferença significativa no consumo de recursos entre elas. Este resultados indicam que não existe vantagens, em relação ao consumo de recursos, de se utilizar qualquer uma das duas frequências. A frequência com menor granularidade de coleta (120 segundos) obteve o melhor desempenho em relação aos consumos de bateria e dados, mas não apresentou ganhos de desempenho significativos em relação aos consumo de processamento, comparado com as frequência de 30 e 60 segundos. Neste ponto de vista, é possível perceber que, em relação ao consumo de recursos do smartphone, em determinadas situações, não existe vantagens de se utilizar uma frequência de coleta ou outra.

Além disso, em relação à qualidade dos dados produzidos, foi possível perceber que, referente aos dados espaciais, a perda de informação foi consideravelmente grande para as frequências de 60 e 120 segundos. Por exemplo, a frequência de coleta de 120 segundos só conseguiu captar 70% de todas as regiões de parada. Adicionalmente, foi possível perceber que em relação aos dados dos sensores do smartphone, como o sensor de iluminação do ambiente, a perda de informação foi ainda maior, podendo ter acertos de apenas 55% do valor real. Por fim, em relação aos dados do uso do smartphone, a perda de informação foi consideravelmente menor, mostrando não existir uma grande diferença de perda de informação dependendo da frequência de coleta.

Nesta perspectiva, a análise comparativa mostrou que, dependendo da situação, uma frequência de coleta pode ser mais indicada do que outra. Por exemplo, em cenários onde se necessite de dados espaciais ou dados de sensores, mas não necessite de um baixo consumo de recursos do smartphone, utilizar uma frequência com coleta de 15 segundos seria a opção mais adequada para se obter os dados com maior qualidade e melhor precisão. Por outro lado, em cenários em que são utilizados apenas dados do uso do smartphone, adotar frequências com menor granularidade não iria impactar significativamente na qualidade dos dados, sendo portanto uma boa alternativa para diminuir o custo do consumo de recursos. Porém, em geral, a frequência de coleta de 30 segundos foi a que obteve o melhor tradeoff entre às frequências analisadas, pois apresentou um desempenho médio no consumo de recursos (semelhante ao consumo com frequência de 60 segundos), ao passo que obteve um bom desempenho em relação a qualidade dos dados captados (inferior apenas à coleta com frequência de 15 segundos).

Como trabalhos futuros, pretende-se replicar o experimento contando com um número maior de voluntários, pelo menos 20 voluntários, acompanhando-os durante um período de tempo mais longo, de 1 a 2 meses. Pretende-se ainda analisar outros fatores, como por exemplo, atividades físicas desempenhadas pelo usuários em horários específicos, e outras frequências de coleta, bem como outras métricas de qualidade de informação produzida a partir dos dados coletados.

5. AGRADECIMENTOS

Primeiramente aos meus pais por todo o suporte ao longo desses anos e durante toda a minha vida.

Aos meus amigos, em especial Lavínia que me deu todo suporte nesses últimos meses, e aos meus parentes que sempre me apoiaram e se disponibilizaram a participar do estudo.

Ao meu orientador, Cláudio Campelo, por todos os ensinamentos, conselhos e orientações durante esses dois anos de projeto e pesquisa.

Por fim, mas não menos importante, a todas as pessoas que me ajudaram, de forma direta ou indiretamente, à chegar até aqui.

6. REFERÊNCIAS

[1] Barros T., Campelo, C. (2019). Plataforma para fomentar a produção e a disponibilização de dados geolocalizados. *XVI Congresso de Iniciação Científica da Universidade Federal de Campina Grande*. ISSN 2177-112X.

[2] El Faouzi, N. E., Leung, H., & Kurian, A. (2011). Data fusion in intelligent transportation systems: Progress and challenges–A survey. Information Fusion, 12(1), 4-10.

[3] Feng, Z., Zhu, Y. (2016). A Survey on Trajectory Data Mining: Techniques and Applications. *IEEE Acces*s, vol. 4, (May 2016), 2056-2067.

[4] Herder, E., Siehndel, P., & Kawase, R. (2014). Predicting user locations and trajectories. *Proceedings of the International Conference on User Modeling, Adaptation, and Personalization*. Springer, Cham, (Jul. 2014), 86-97.

[5] Lerato, L. & Niesler, T. J AUDIO SPEECH MUSIC PROC. (2019) 2019: 6. https://doi.org/10.1186/s13636-019-0149-9.

[6] Luo, Ting & Zheng, Xinwei & Xu, Guangluan & Fu, Kun & Ren, Wenjuan. (2017). An Improved DBSCAN Algorithm to Detect Stops in Individual Trajectories. *ISPRS International Journal of Geo-Information*. 6. 63. 10.3390/ijgi6030063.

[7] Noulas, A., Scellato, S., Lambiotte, R., Pontil, M., & Mascolo, C. (2012). A tale of many cities: universal patterns in human urban mobility. *PloS one*, 7(5), e37027.

[8] Kong, Xiangjie & Li, Menglin & Ma, Kai & Tian, Kaiqi & Wang, Mengyuan & Ning, Zhaolong & Xia, Feng. (2018). Big Trajectory Data: A Survey of Applications and Services. *IEEE Access*. DOI 10.1109/ACCESS.2018.2873779.

[9] Mazimpaka, Jean Damascene & Timpf, Sabine. (2016). Trajectory data mining: A review of methods and applications. *Journal of Spatial Information Science*. 13. 61–99.

[10] Parent, C., Spaccapietra, S., Renso, C., Andrienko, G., Andrienko, N., Bogorny, V. & Theodoridis, Y. (2013). Semantic trajectories modeling and analysis. *ACM Computing Surveys (CSUR)*, 45(4), 42.

[11] Vaughan, N., Gabrys, B. (2016). Comparing and combining time series trajectories using Dynamic Time Warping. *Proceedings of the 20th International Conference on Knowledge Based and Intelligent Information and Engineering Systems*, 465-474.

[12] Xiang, L., Gao, M., Wu, T. (2016). Extracting Stops from Noisy Trajectories: A Sequence Oriented Clustering Approach. *ISPRS Int. J. Geo-Inf*. 2016, 5, 29.

[13] Zheng, Y. (2015). Trajectory data mining: an overview. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 6(3), 29.

[14] Ziaeefard, M. Bergevin, R. (2015). Semantic human activity recognition: A literature review. *Pattern Recognition*, vol. 48 (Aug. 2015), 2329-2345.