Identificação e Classificação de Pontos de Interesse Individuais com Base em Dados Esparsos

Cláudio G. S. Capanema[1], Fabrício A. Silva[1],Thais Regina M. B. Silva[1]

[1] Departamento de Informática - Universidade Federal de Viçosa

{claudio.capanema,fabricio.asilva,thais.braga}@ufv.br

Resumo. *Dados de localização de dispositivos móveis são fontes importantes para entender o perfil de usuários, ajudando os provedores a oferecerem melhores serviços. Com esse tipo de dado, é possível identificar os pontos relevantes de um usuário, e até mesmo classificar esses pontos como locais de casa ou trabalho. Com esse conhecimento, provedores de serviços móveis podem aumentar o engajamento e a retenção de seus clientes. No entanto, identificar e classificar pontos de interesse (PoI) não são tarefas triviais, e a maioria dos trabalhos existentes assumem que os dados devem ser coletados com uma frequência alta, dificultando e encarecendo o processo. Neste trabalho, são propostas abordagens para identificar e classificar PoIs com base em dados esparsos, ou seja, que foram coletados em intervalos longos de tempo. Os resultados, quando comparados com soluções da literatura, mostram melhorias de pelo menos 13% na precisão para a identificação dos PoIs, e de 10% e 4% na classificação de pontos de casa e de trabalho, respectivamente.*

Abstract. *Geo-spatial data are important sources to understand mobile users profile, helping providers to offer better services. With this type of data, it is possible to identify relevant visiting points of a user, and even to classify these points as home and work locations. With this knowledge, mobile service providers can increase the engagement and the retention of these users. However, identifyng and classifying point of interest (PoI) are not trivial tasks, and the majority of existing works assume that the data have to be collected with a high frequency, making the process harder and more expensive. In this work, we propose approaches to identify and classify PoIs based on sparse data that were collected during long time intervals. The results, when compared with literature solutions, show precision improvements of at least 13% on the identification of PoIs, and 10% and 4% in classification of home work, respectivaly.*

1. Introdução

O advento da utilização em massa de dispositivos móveis, como *smartphones* e *tablets*, trouxe consigo a geração de grandes volumes de dados de localização de usuários. Diversas aplicações fazem o uso do sensor de GPS para fornecer serviços de ofertas baseadas em localização, auxílio na mobilidade, buscas na Web e entrega de conteúdo digital orientados à localização. Além disso, empresas de telefonia coletam dados de registros de acessos, chamados CDR (*Call Detail Records*), que também representam uma fonte importante de dados de localização [Naboulsi et al. 2016].

Além de facilitar a oferta de serviços e conteúdos baseados em localização, dados georreferenciados têm sido utilizados para o entendimento de padrões de mobilidade tanto de indivíduos quanto de grupos de pessoas [Pavan et al. 2015, Naboulsi et al. 2016]. Essa fonte de dados está intimamente relacionada com o conceito de Cidades Inteligentes, uma vez que auxilia, por exemplo, no planejamento urbano [Rathore et al. 2016] e na previsão de volume de tráfego rodoviário [Castro et al. 2012]. Além disso, empresas de diversos ramos utilizam dados georreferenciados para conhecer melhor os seus clientes para, assim, oferecerem serviços mais personalizados.

Um aspecto importante da mobilidade urbana e do perfil de usuários móveis refere-se à identificação e classificação de pontos de interesse (PoIs) dos usuários. Esses pontos correspondem a locais que uma pessoa visita com certa frequência, podendo representar locais de residência, trabalho, lazer, escolas, locais em que se costuma fazer compras e se alimentar, dentre outros. Neste contexto, a identificação de PoIs refere-se a encontrar esses locais (i.e., definir as coordenadas aproximadas), enquanto a classificação visa categorizar um PoI pelo seu tipo (i.e., casa, trabalho, lazer, dentre outros).

A maioria dos trabalhos que visam identificar ou classificar PoIs utilizando informações geradas por GPS baseiam-se em dados densos, ou seja, com alta frequência de coleta (i.e., na ordem de poucos segundos). Com dados densos, é possível observar vários aspectos, como horário de chegada e partida de um local (e consequentemente o tempo de permanência), o trajeto feito de um local a outro e o tempo de deslocamento. No entanto, a coleta intensiva de dados georreferenciados leva a um alto consumo energético dos aparelhos móveis devido à utilização do sensor de GPS, um alto consumo de rede para a transmissão desses dados e uma necessidade maior de capacidade de armazenamento e processamento no servidor. Por isso, geralmente dados densos não são coletados, ou são coletados para amostras pequenas de usuários voluntários. Por outro lado, a coleta de dados georreferenciados de forma esparsa é mais fácil e menos custosa de ser alcançada, sendo uma alternativa viável tanto para os dispositivos móveis quanto para o servidor de armazenamento e processamento desses dados.

Dadas as informações acima, surgem as seguintes perguntas de pesquisa:

- É possível identificar com precisão PoIs de usuários com base em dados esparsos?
- É possível classificar com precisão os PoIs de usuários em *Casa* e *Trabalho* com base em dados esparsos?

Para responder a essas perguntas, o objetivo deste trabalho é propor algoritmos para identificação e classificação de PoIs individuais com base em dados esparsos. O algoritmo de identificação visa inferir os PoIs de um usuário com base em seus locais visitados. Já o algoritmo de classificação visa classificar quais pontos de interesse representam o local de *Casa* e *Trabalho*. Os algoritmos propostos foram comparados com soluções bem conhecidas da literatura utilizando a mesma base de dados esparsos. Observou-se que a proposta atual supera [Montoliu et al. 2013] e [Cuttone et al. 2014] na precisão para a identificação de PoIs em pelo menos 13%. Comparando com [Hoteit et al. 2016] e [Kung et al. 2014], as melhorias encontradas para a classificação de PoIs foram de pelo menos 10% para *Casa* e de 4% para *Trabalho*.

Este trabalho está organizado da seguinte forma. Inicialmente, na Seção 2 é apresentada uma revisão da literatura contendo os principais estudos relacionados à área de identificação e classificação de PoIs. Em seguida, na Seção 3 são descritas as caracte-

rísticas da base de dados utilizada no trabalho. Na Seção 4 o algoritmo proposto para a identificação de PoIs é apresentado e comparado com dois estudos da literatura. Em seguida, a Seção 5 contém a descrição e avaliação do algoritmo proposto para a classificação de PoIs em *Casa* ou *Trabalho*. Por último, na Seção 6 a conclusão e os trabalhos futuros são apresentados.

2. Trabalhos Relacionados

A identificação e classificação de PoIs é um assunto que faz parte de estudos sobre a mobilidade humana, e que vem crescendo nos últimos anos. Nesta seção, são apresentados os principais trabalhos da área. Inicialmente são apresentadas as soluções para a identificação de PoIs. Em seguida, são citadas propostas de classificação de PoIs em *Casa* e *Trabalho*. A Tabela 1 sumariza os principais trabalhos da literatura.

Para a identificação de PoIs, é comum utilizar-se algoritmos de agrupamento para processar as localizações. O trabalho apresentado em [Csáji et al. 2013] utiliza CDRs e diagrama de Voronoi para associar as torres de telefonia mais frequentemente utilizadas com as suas regiões de cobertura, e em seguida as agrupa utilizando um método de triangulação. O trabalho de [Frias-Martinez et al. 2010] também utiliza diagrama de Voronoi com o mesmo propósito do artigo citado anteriormente. Os autores de [Isaacman et al. 2011, Ranjan et al. 2012] também processam dados de CDRs, e utilizam o algoritmo *Leader* para agrupar as torres de celular próximas. Além disso, [Isaacman et al. 2011] recorre à regressão logística para se obter a relevância de cada local. Já o trabalho [Lee et al. 2015] recorre a uma extensão do algoritmo DBScan, na qual é feito o agrupamento de traços de GPS com base em restrições de distância e velocidade calculadas entre dois pontos para se descobrir locais relevantes. Os autores de [Cuttone et al. 2014] apresentam uma solução baseada no algoritmo *GMM*, e outra que utiliza o DBScan juntamente com uma restrição de distância máxima entre pares de coordenadas de GPS para identificar PoIs. Em [Pavan et al. 2015], são consideradas restrições de tempo, distância e velocidade máximas para filtrar pares de registros consecutivos. Por último, o trabalho [Montoliu et al. 2013] apresenta soluções de agrupamento que utilizam restrições de tempo e distância entre registros consecutivos.

Dentre esses trabalhos, [Csáji et al. 2013], [Trestian et al. 2009], [Kung et al. 2014], [Isaacman et al. 2011], [Järv et al. 2014], [Schneider et al. 2013], [Ranjan et al. 2012] e [Frias-Martinez et al. 2010] utilizam CDRs, que são dados geralmente disponibilizados por operadoras de telefonia e que possuem apenas a localização das torres, e não dos aparelhos. Já [Lee et al. 2015, Pavan et al. 2015, Montoliu et al. 2013] assumem que os dados de GPS são densos, para se conhecer detalhes de deslocamento e permanência dos usuários em cada local. O estudo de [Cuttone et al. 2014] é um dos poucos que considera dados de GPS esparsos e está mais diretamente relacionado ao presente trabalho, sendo utilizado como base de comparação na avaliação dos resultados. Considerando a relevância e a qualidade do trabalho apresentado em [Montoliu et al. 2013], o mesmo também é utilizado como base de comparação, sendo que seus parâmetros foram ajustados para que ele seja melhor adaptado para dados esparsos, que é o foco do trabalho atual.

Outras técnicas além de agrupamentos também são utilizadas para a identificação de PoIs. Os autores de [Järv et al. 2014] utilizaram *Multiple Linkage Analysis* para se obter as localizações com maior número de chamadas em cada mês de um usuário,

que são consideradas suas localizações de interesse. Por outro lado, o tempo em que um usuário permaneceu em cada local visitado é considerado um fator relevante para a identificação de PoIs pelas propostas de [Trestian et al. 2009], [Kung et al. 2014] e [Schneider et al. 2013]. No entanto, esses trabalhos utilizam dados esparsos de CDRs, o que dificulta a estimativa precisa de tempo de permanência em um local.

Para a classificação de PoIs em *Casa* e *Trabalho*, o tempo de permanência em intervalos de horários pré-definidos é uma métrica comumente utilizada. De acordo com [Trestian et al. 2009], o local de maior tempo de permanência entre 22:00h e 6:00h do dia seguinte é classificado como *Casa*; por outro lado, o local de *Trabalho* corresponde aos períodos de 10:00h às 12:00h e de 14:00h às 17:00h. Em [Kung et al. 2014], o local no qual o usuário permaneceu por mais tempo de 8:00h às 20:00h corresponde ao *Trabalho*, e de 20:00h às 8:00h à *Casa*. O artigo [Schneider et al. 2013] considera como *Casa* o local de maior tempo de permanência de meia noite às 06:00h horas da manhã.

A ideia de se classificar os locais como *Casa* e *Trabalho* com base no tempo de permanência em determinadas faixas de horário faz sentido, pois em geral as pessoas passam boa parte do dia em seu local de *Trabalho*, e da noite em *Casa*. Porém, medidas de tempo de permanência não são precisas o suficiente em dados esparsos, uma vez que o usuário pode visitar diversos locais sem gerar registros. Além disso, intervalos de tempo fixos não necessariamente representam todos os usuários, que possuem rotinas diferentes. Neste trabalho, é proposto um algoritmo que define intervalos de horário específicos para cada usuário para a classificação dos locais de *Casa* e *Trabalho*.

Para [Csáji et al. 2013] é possível discernir claramente os locais de *Casa* e *Trabalho* após a aplicação do algoritmo *k-means* sobre os PoIs. Além disso, trabalhos como [Ranjan et al. 2012, Isaacman et al. 2011, Frias-Martinez et al. 2010, Hoteit et al. 2016] contabilizam a quantidade de registros em determinadas faixas de horários para classificar os locais em *Casa* e *Trabalho*. As soluções [Hoteit et al. 2016, Kung et al. 2014] foram utilizadas neste trabalho como bases de comparação com o nosso algoritmo de classificação de PoI por se tratarem de trabalhos recentes e relevantes da literatura.

Um dos grandes desafios do problema de identificar e classificar PoIs é a validação, já que dados rotulados com essas informações não são facilmente obtidos. Considerando CDRs, empresas de telefonia podem utilizar dados dos contratos de serviços para validar a região da residência e trabalho de seus usuários. No entanto, a precisão obtida ao se utilizar esse tipo de dado é baixa, já que uma antena cobre de centenas de metros a quilômetros. Por outro lado, trabalhos que utilizam dados de GPS necessitam da colaboração de usuários para prover suas informações, e esses dados são em baixa escala, raros e privados. Neste quesito, os trabalhos [Hoteit et al. 2016], [Cuttone et al. 2014], [Lee et al. 2015], [Montoliu et al. 2013] e [Pavan et al. 2015] processaram dados de, respectivamente, 32, 6, 46, 8 e 182 usuários. Já no presente trabalho, foi possível analisar dados reais de 194 usuários, superando as abordagens que utilizaram dados de GPS. A Tabela 1 apresenta os principais aspectos de cada um dos trabalhos relacionados, destacando os tipos de dados, e as técnicas utilizadas por cada um. Sensores como WiFi, *bluetooth* e acelerômetro, além da rede de telefonia (GSM) também podem ser associados às fontes de dados CDR e GPS para se inferir localização de usuários.

Tabela 1. Comparação entre soluções

Solução	Fonte(s) de dado(s)	Técnica(s) para identificação	Técnica(s) para classificação
[Csáji et al. 2013]	CDR	Diagrama de Voronoi e triangulação	*K-means*
[Trestian et al. 2009]	CDR	Tempo de permanência	Tempo de permanência
[Kung et al. 2014]	CDR e GPS	Tempo de permanência	Tempo de permanência
[Isaacman et al. 2011]	CDR	Leader e regressão logística	Quantidade de registros
[Hoteit et al. 2016]	GPS	Quantidade de registros	Quantidade de registros
[Järv et al. 2014]	CDR	*Multiple Linkage Analysis*	*Não classifica
[Schneider et al. 2013]	CDR	Tempo de permanência	Tempo de permanência
[Cuttone et al. 2014]	GPS, Wi-Fi e GSM	*Gaussian Mixture Method* e DBScan	*Não classifica
[Lee et al. 2015]	GPS	DBScan	*Não classifica PoI
[Ranjan et al. 2012]	CDR	*Leader*	Quantidade de registros
[Frias-Martinez et al. 2010]	CDR	Diagrama de Voronoi	Algoritmo genético e quantidade de registros
[Montoliu et al. 2013]	GPS, GSM, Wi-Fi, Bluetooth e acelerômetro	Agrupamento por Grade e DBScan, e restrições de distância e tempo entre registros	*Não classifica
[Pavan et al. 2015]	GPS	Restrições de distância, tempo e velocidade entre registros	*Não classifica
Proposta Atual	GPS	DBScan, e restrições de dias e horas diferentes	Quantidade de registros em horários específicos para cada usuário

3. O Conjunto de Dados

Os dados utilizados neste trabalho foram disponibilizados por uma empresa provedora de serviços móveis sob um acordo de confidencialidade. Foram fornecidos dados de localização de 194 usuários voluntários de dispositivos móveis, durante um período de até 62 dias. Os dados foram coletados mediante a permissão prévia dos usuários.

Para esse conjunto de usuários, foi gerado um registro de localização sempre que o usuário se deslocar para um novo local, distante pelo menos 100 metros do local anterior, e desbloquear a tela do seu *smartphone*. Ou seja, os dados são esparsos pois um único registro de localização é gerado quando há um deslocamento do usuário e um acesso ao *smartphone* no novo local. Um registro de localização contém o identificador do usuário, data e horário, latitude e longitude obtidas pelo sensor de GPS. O identificador do usuário é um número aleatório e único para cada usuário, que não permite a identificação do mesmo de forma alguma.

Os dados abrangem o território brasileiro, e como pode ser observado na Figura 1, há uma maior concentração de registros nas regiões nordeste, sudeste e sul. A Figura 2 exibe a quantidade total de registros gerados em cada hora do dia, sendo possível observar que os usuários são mais ativos entre 8:00h e 10:00h da manhã, e entre 13:00h e 16:00h da tarde. Além disso, a partir da Figura 3 percebe-se que a maioria dos usuários gerou entre 500 e 600 registros durante todo o período de coleta. É importante ressaltar que esse período de coleta varia de usuário para usuário, e os valores de média e mediana correspondentes são ambos de 34 dias. Isso indica que, em geral, os dados dos usuários foram coletados em uma mesma quantidade de dias.

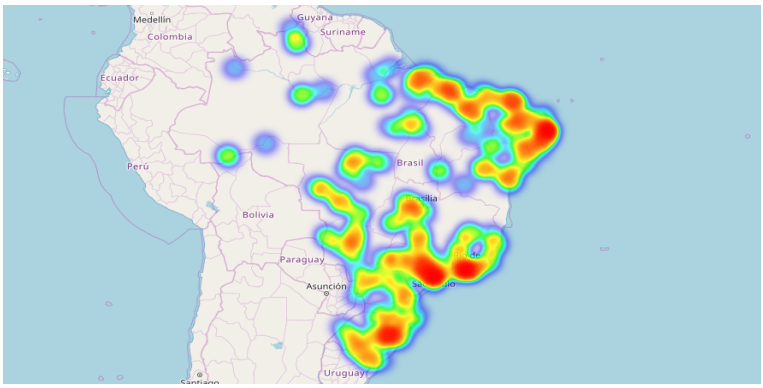

Figura 1. Mapa de calor dos registros coletados dos 194 usuários.

Com o objetivo demonstrar o quão esparsos são os dados, foram gerados gráficos de CDF (Função de Distribuição Acumulada) para a distância (Figura 4) e para o intervalo de tempo (Figura 5) entre pares de registros consecutivos dos usuários. Para as distâncias, a mediana é de 874 metros, a média é de 2.469 metros, e o desvio padrão é de 4.723 metros. Optamos por representar distâncias de até 47.000 metros, para criar um gráfico visualmente mais intuitivo, sendo que 1,31% das distâncias são superiores a 47.000 metros. Já para os intervalos de tempo, a mediana é de 26 minutos, a média corresponde a 66 minutos, e o desvio padrão é de 116 minutos. Além disso, optamos por representar tempos de até 700 minutos, também com o objetivo de criar um gráfico visualmente mais intuitivo, sendo que 1,75% dos tempos são maiores do 700 minutos.

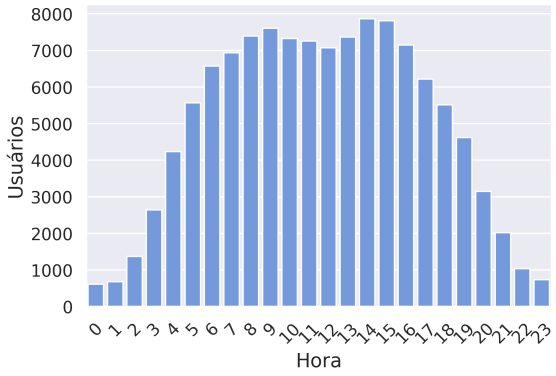

Figura 2. Registros por hora

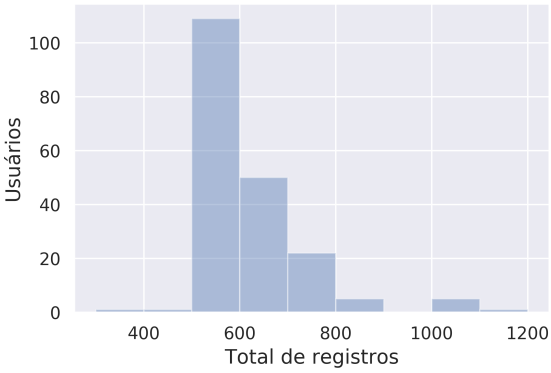

Figura 3. Total de registros por usuário

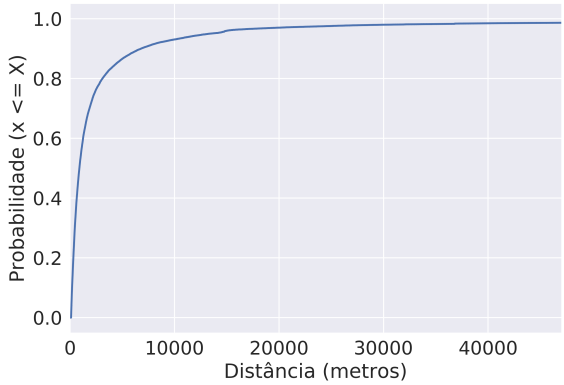

Figura 4. CDF de distância entre cada par de registros consecutivos

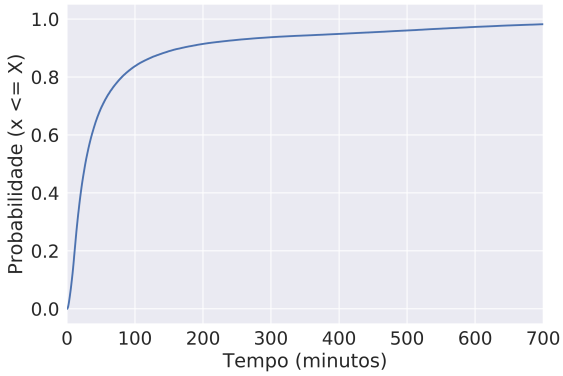

Figura 5. CDF do intervalo de tempo entre cada par de registros consecutivos

4. Identificação de PoI

A identificação de pontos de interesse é uma tarefa essencial para o estudo da mobilidade humana. Com a informação dos PoIs individuais, é possível estudar padrões de mobilidade, perfil de usuários, demanda por infraestrutura computacional e viária, dentre outros. Porém, quando os dados são esparsos, os desafios são maiores pois não é possível ter certeza do horário de chegada em um local, do tempo de permanência, e nem da intensidade de uso de um dispositivo móvel em cada local.

4.1. Solução

Neste trabalho, é proposto e validado um algoritmo para identificar pontos de interesse que seja adequado à característica esparsa dos dados. O algoritmo possui duas etapas: inicialmente, as localizações dos registros são agrupadas utilizando-se um algoritmo de agrupamento por densidade (DBScan), e em seguida os grupos encontrados são filtrados com base em duas métricas possíveis de serem conhecidas em dados esparsos, que são o total de dias visitados e a diversidade de horas de visita. Dessa forma, é possível se obter um conjunto de PoIs mais relevantes, melhorando assim o desempenho do algoritmo.

Para o agrupamento inicial, o DBScan foi escolhido por ser um algoritmo eficiente baseado em densidade, que não exige o número fixo de grupos como entrada e desconsi-

dera pontos que não atendem às restrições estabelecidas por seus parâmetros. Neste caso, o parâmetro de número mínimo de amostras do DBScan foi definido de forma que cada grupo tenha mais de um registro por semana. Em avaliações empíricas, chegou-se a um valor de 18% do total de dias de coleta dos dados de cada usuário, o que corresponde a uma média de 1,26 registros por semana. Além disso, cada registro de um mesmo grupo não pode estar distante mais de 10 metros um do outro. Esses parâmetros foram definidos empiricamente, e garantem que locais visitados muito esporadicamente, e que não estejam próximos a outros locais, não sejam considerados como candidatos a PoIs.

Ao final dessa primeira etapa, cada grupo gerado corresponde a um PoI candidato, e possui um conjunto de coordenadas e seus respectivos horários e datas. Dentre os PoIs candidatos gerados, são selecionados aqueles que obedecem a duas restrições. A primeira seleciona os PoIs que foram visitados em pelo menos 15% dos dias dado todo o período de amostragem do usuário. Isso significa que, por exemplo, para um período de 28 dias de coleta de dados, um PoI deve ter sido visitado pelo usuário pelo menos 4 dias diferentes no período, ou o equivalente a uma vez por semana. É importante ressaltar que o parâmetro de número mínimo de amostras permite que locais visitados em um pequeno número de dias diferentes sejam considerados como candidatos a PoI, enquanto que a restrição descrita impede isso. Por outro lado, a segunda restrição objetiva responder à seguinte pergunta: é possível estabelecer um número mínimo de horas diferentes do dia em que os registros pertencentes a um PoI foram gerados, de modo que se melhore o desempenho do algoritmo para a identificação de pontos de interesse? Após a realização de vários testes empíricos, observou-se que a seleção de PoIs que possuem pontos coletados em pelo menos 7 horas diferentes melhora o desempenho do algoritmo. Essas regras fazem com que apenas locais mais relevantes sejam considerados como PoIs.

O Algoritmo 1 representa a solução para a identificação de PoIs, que recebe como entrada a lista de registros de um usuário e retorna a lista de PoIs desse usuário. Na linha 2, os registros recebidos como entrada são agrupados pelo algoritmo DBScan. Em seguida, a partir da linha 3 até a linha 13, cada um dos grupos gerados são processados para se obter a quantidade de dias e horas diferentes que os seus registros representam. Na linha 14, é obtido o valor, em dias, do período entre o primeiro e o último registros gerados daquele grupo. Se a quantidade de dias diferentes for maior ou igual a 15% do período que o usuário visitou aquele local, e se a quantidade de horas diferentes for maior ou igual a 7, então este grupo é considerado um PoI do usuário, como descrito no algoritmo nas linhas 15 e 16.

4.2. Soluções Base

Com o objetivo de comparar o desempenho do algoritmo proposto com outras soluções da literatura, os algoritmos apresentados por [Cuttone et al. 2014] e [Montoliu et al. 2013] foram implementados. [Cuttone et al. 2014] é um dos poucos trabalhos que consideram dados georreferenciados esparsos para a identificação de PoIs. Foi apresentada uma solução baseada no algoritmo *GMM (Gaussian Mixture Method)*, e uma baseada no algoritmo DBSCan, a qual foi escolhida devido ao melhor desempenho.

[Montoliu et al. 2013] apresentou um algoritmo de agrupamento baseado em grade, e outro baseado no DBSCan, sendo que esse último obteve melhor desempenho sobre a nossa base de dados, e portanto foi escolhido como uma das soluções base. Esse algoritmo primeiramente agrupa os registros em relação ao tempo de permanência no locais,

Algorithm 1 Identificação de PoIs de usuário

Require: $D, r = (r_1, ..., r_n)$: {Período de dias de coleta, e Lista de registros}
Ensure: $PoIs$ {Lista de pontos de interesse}

$PoIs \leftarrow \emptyset$
$Grupos = DBScan(data \leftarrow r, eps \leftarrow 10m, min\_samples \leftarrow |D| * 0.18)$
for $grupo \in Grupos$ **do**
  $dias\_diferentes \leftarrow \emptyset$ {Lista com dias diferentes no grupo}
  $horas\_diferentes \leftarrow \emptyset$ {Lista com horas do dia diferentes no grupo}
  for $ponto \in grupo$ **do**
    if $ponto.dia \notin dias\_diferentes$ **then**
      $dias\_diferentes \leftarrow dias\_diferentes \cup ponto.dia$
    end if
    if $ponto.hora \notin horas\_diferentes$ **then**
      $horas\_diferentes \leftarrow horas\_diferentes \cup ponto.hora$
    end if
  end for
  $periodo \leftarrow Periodo(grupo)$
  if $|dias\_diferentes| \geq periodo * 0.15$ **and** $|horas\_diferentes| \geq 7$ **then**
    $PoIs \leftarrow PoIs \cup grupo$
  end if
end for

para em seguida agrupar utilizando o DBScan. Apesar de se basear em dados densos, essa solução foi escolhida por ser uma referência na literatura, e para verificar se a mesma pode também ser utilizada com dados esparsos, o que ainda não foi feito por outros estudos. Para isso, os parâmetros da solução foram ajustados empiricamente para que sejam o mais propícios a dados esparsos possível.

4.3. Resultados

Seja PoI_u o conjunto de PoIs do usuário u rotulados e conhecidos, $PoI_{u,c}$ o conjunto de PoIs do usuário u identificados corretamente e $PoI_{u,i}$ o conjunto de PoIs do usuário u identificados incorretamente por um algoritmo. A precisão ($p = \frac{|PoI_{u,c}|}{|PoI_{u,c}|+|PoI_{u,i}|}$) indica, dentre os pontos identificados, quantos estão corretos. A revocação ($r = \frac{|PoI_{u,c}|}{|PoI_u|}$) indica, dentre todos os PoIs reais de um usuário, quantos foram identificados corretamente pelo algoritmo. Por fim, o *f-score* ($f = 2 * \frac{p*r}{p+r}$) é derivado dessas duas métrica. Para as três métricas, quanto mais próximo de 1, melhor.

Primeiramente, foi avaliado o quão distantes os PoIs identificados pelos algoritmos estão dos PoIs reais dos usuários. Para isso, a Figura 6 ilustra a CDF (Função de Distribuição Acumulada) das distâncias entre todo PoI identificado pelos algoritmos e o PoI real mais próximo. Essa figura mostra o eixo-x com valor máximo de 1.000 metros por questões de visualização, sendo que a proposta atual tem 4,48% das distâncias acima desse valor, enquanto as soluções [Cuttone et al. 2014] e [Montoliu et al. 2013] têm 5,6% e 9,4%, respectivamente. Com base nesse gráfico, é possível perceber que a proposta deste artigo leva à identificação de PoIs mais próximos dos reais.

Para contabilizar os PoIs identificados corretamente, é definida uma distância má-

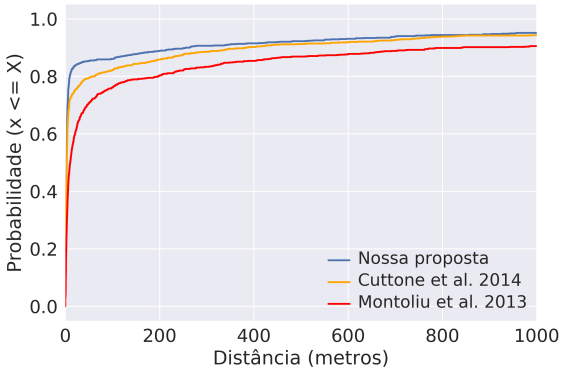

Figura 6. CDF de distâncias entre PoIs identificados e reais

xima para a margem de erro, já que dificilmente um algoritmo irá identificar exatamente as mesmas latitudes e longitudes de um PoI real. Neste trabalho, essa distância de margem de erro foi variada entre 10 e 100 metros. Assim, um PoI identificado é considerado correto se ele está distante a no máximo a distância de margem de erro de um PoI real.

As Figuras 7, 8 e 9 apresentam os resultados médios e o intervalo de confiança de 95% da proposta atual e das soluções base [Cuttone et al. 2014] e [Montoliu et al. 2013]. Os desempenhos variam de acordo com a distância da margem de erro (eixo-x) entre o PoI encontrado por cada solução e o PoI real mais próximo. Pode-se perceber um desempenho superior da proposta atual em relação às soluções base nas métricas precisão, revocação e *f-score*. A precisão na nossa proposta varia de 70% à 74%, enquanto para [Cuttone et al. 2014] o intervalo é de 56% à 61%, e para [Montoliu et al. 2013] a precisão varia de 38% à 52%. Para a métrica revocação, os desempenhos das abordagens citadas anteriormente variam respectivamente, de 80% à 84%, de 73% à 80%, e entre 58% e 79%. Os desempenhos da proposta deste trabalho, de [Cuttone et al. 2014] e [Montoliu et al. 2013] para a métrica *f-score* variaram respectivamente entre 73% e 77%, 61% e 67%, e 45% e 61%.

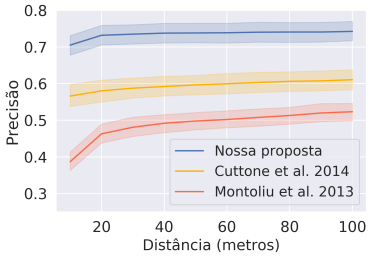

Figura 7. Precisão

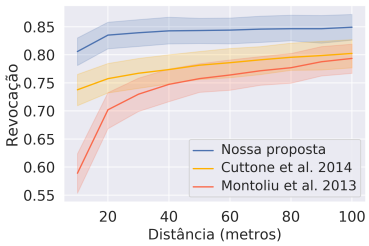

Figura 8. Revocação

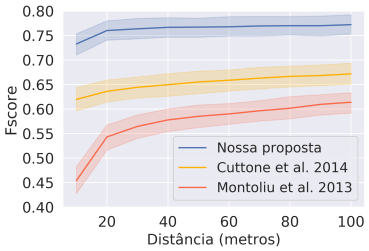

Figura 9. *f-score*

Esses resultados mostram que, apesar de ser uma solução bem elaborada e eficiente para dados densos, a proposta de [Montoliu et al. 2013] não é apropriada para dados esparsos. A solução proposta por [Cuttone et al. 2014], por ser focada em dados esparsos, consegue alcançar resultados melhores que [Montoliu et al. 2013]. A proposta deste trabalho atual consegue melhores resultados ao considerar o número de horas e dias distintos de visitas aos locais, desconsiderando assim locais visitados um número significativo de

vezes (o que faz com que um grupo seja considerado pelo DBScan), mas em poucos dias e horários diferentes (o que faz com que as restrições do algoritmo o desconsidere).

5. Classificação de PoIs

A classificação dos tipos de locais que cada usuário frequenta também corresponde a um passo relevante no processo de análise da mobilidade humana. Ao se conhecer os locais de casa e trabalho, e os padrões de deslocamento entre esses locais, provedores de serviços podem personalizar seus serviços de acordo com o perfil do usuário. Neste trabalho, são classificados os locais de *Casa* e *Trabalho* de cada usuário. Tanto a proposta apresentada quanto as soluções base selecionadas processam apenas pontos de GPS coletados durante dias de semana. Isso porque, em geral, a rotina humana é melhor definida durante esses dias, o que torna mais eficaz a classificação desses tipos de PoIs.

5.1. Solução

A *Casa* e o *Trabalho* são definidos como os locais onde mais registros foram gerados em horários específicos. O grande diferencial desta proposta é que esses horários variam de usuário para usuário, e não são fixos como nos outros trabalhos.

Inicialmente, identifica-se o maior período de inatividade do usuário, ou seja, o maior intervalo de horário de 00:00h até as 23:00h no qual não se geraram registros, considerando todos os dias de amostragem. O horário para a identificação da *Casa* corresponde ao intervalo de 2 horas anteriores até as 2 horas posteriores ao intervalo de inatividade encontrado na etapa anterior. A justificativa para isso é que, em geral, as pessoas começam e terminam os seus dias no local de sua moradia, e portanto é mais provável que sejam gerados registros em casa no intervalo próximo ao período de inatividade. Caso não seja encontrado um período de inatividade, consideramos intervalos fixos iguais aos da solução base [Hoteit et al. 2016], que será descrita posteriormente, uma vez que empiricamente esses foram os intervalos fixos que trouxeram os melhores resultados. O intervalo de horário para a classificação do local de *Trabalho* corresponde ao período do dia oposto ao intervalo de horário definido para a classificação da *Casa*. Dessa forma, supomos que o *Trabalho* é o local onde uma pessoa gera mais registros quando está fora de casa.

Em outras palavras, são definidos dois intervalos de horários para contabilizar registros em cada local. O local que possui mais registros no período definido para classificar a casa é classificado como *Casa*, e o mesmo processo é utilizado para classificar o local que representa o *Trabalho* do usuário. Portanto, caso o intervalo de horário de inatividade de um dado usuário seja das 22:00h até as 05:00h do dia seguinte, por exemplo, então o período para a classificação em *Casa* é das 20:00 (22:00h - 2h) às 07:00h (05:00h + 2h). Dessa forma, o período para a classificação do PoI em *Trabalho* é das 08:00h às 19:00h.

O Algoritmo 2 apresenta a solução proposta para a classificação de PoIs para cada usuário. A partir da linha 1 até a linha 9, são obtidas as horas que os registros de todos os PoIs foram gerados. Com essas informações, é possível obter o intervalo de inatividade através da função utilizada na linha 10. Como mencionado anteriormente, caso não exista um intervalo de inatividade para o usuário, considera-se o mesmo intervalo utilizado pela solução de [Hoteit et al. 2016]. Nas linhas 11 e 12, as funções utilizadas retornam o PoI

que representa o local de *Casa*, e o PoI que representa o local de *Trabalho*. Para todo usuário, o algoritmo sempre define os locais de casa e trabalho.

Algorithm 2 Classificação de PoIs de usuário

Require: $PoIs = (p_1, ..., p_n)$: {Lista de pontos de interesse}
Ensure: $casa \in PoIs, trabalho \in PoIs$ {Pontos de interesse classificados}

```
horas ← ∅
for poi ∈ PoIs do
    for ponto ∈ poi do
        hora ← ponto.hora
        if hora ∉ horas then
            horas ← horas ∪ hora
        end if
    end for
end for
intervalo ← Intervalo_Inativo(horas)
casa ← Classifica_Casa(PoIs, intervalo)
trabalho ← Classifica_Trabalho(PoIs, intervalo)
```

5.2. Resultados

Foram utilizadas as propostas de [Kung et al. 2014] e [Hoteit et al. 2016] como soluções base. A primeira foi adaptada para contabilizar a quantidade de registros de cada PoI entre intervalos de horários, e não o tempo de permanência. Esta solução considera o intervalo de horário de 20:00h às 08:00h para determinar o local de *Casa*, e 08:00h às 20:00h para determinar o local de *Trabalho*. A segunda proposta classifica a *Casa* como o local onde o usuário gerou mais registros no período de tempo de 22:00h às 7:00h do dia seguinte, e o *Trabalho* como o local de maior número de registros de 9:00h às 17:00h. Nesta avaliação, foi considerada uma distância de 100m de margem de erro. Em outras palavras, se o local de *Casa* classificado pelo algoritmo estiver a menos de 100m do local real da casa do usuário, então considera-se um acerto. Vale destacar que os PoIs considerados como reais são os originais dos dados rotulados, e não os obtidos pelo algoritmo proposto neste trabalho. Isto permite que a validação seja feita de modo independente, assumindo que a identificação de PoIs já foi realizada. Além disso, somente foram considerados na avaliação os PoIs rotulados como *Casa* ou *Trabalho* nos dados originais, visando assim avaliar a eficiência das soluções em distinguir entre um tipo de local e outro. A precisão é dada pela divisão do número de acertos pelo número de usuários, pois foram classificados os locais de todos os usuários.

A Figura 10 apresenta o desempenho do algoritmo proposto em comparação com as soluções base [Kung et al. 2014] e [Hoteit et al. 2016], para classificar PoIs em *Casa* e *Trabalho*. A utilização de intervalos variáveis de horário levou a um desempenho melhor, de 95% para a classificação do local de *Casa* e de 50% para a classificação do local de *Trabalho*. Os respectivos desempenhos alcançados por [Kung et al. 2014] foram de 85% e 44%. Por outro lado, a solução [Hoteit et al. 2016] obteve valores de 84% para classificação de *Casa* e 46% para *Trabalho*. Além disso, foi possível utilizar intervalos variáveis de horário para 112 dos 194 usuários, demonstrando que a proposta atual pode

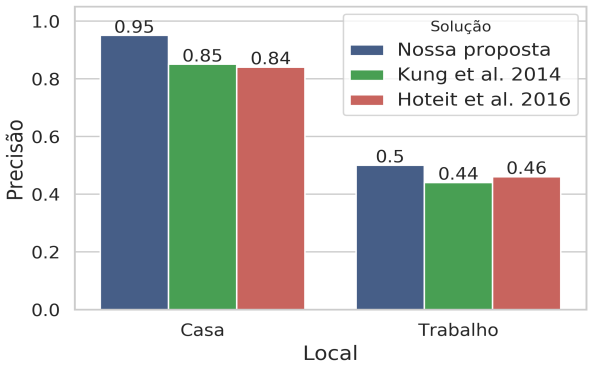

Figura 10. Precisão de classificação de PoIs

ser aplicada de maneira abrangente.

6. Conclusões e Trabalhos Futuros

O presente trabalho apresentou duas contribuições para a área de análise de mobilidade, tendo como diferencial a utilização de dados esparsos. A primeira contribuição corresponde à identificação de pontos de interesse (PoIs) de usuários de dispositivos móveis, e a segunda corresponde à classificação de PoIs em casa e trabalho. Os algoritmos apresentados foram validados comparando-se os resultados gerados com soluções conhecidas da literatura. As avaliações comparativas mostraram que foi possível alcançar melhores resultados, demonstrando que as hipóteses assumidas em termos da diversidade de visitas (para identificação) e da personalização dos intervalos de horários (para classificação) foram eficientes.

Como trabalhos futuros, pretende-se melhorar a classificação de locais de trabalho, além de classificar outros tipos de PoIs, como locais de refeição e lazer. Além disso, também é relevante analisar aspectos de deslocamento entre os PoIs, para se prever possíveis demandas por recursos computacionais e viários.

7. Agradecimento

Este trabalho contou com o apoio da Fapemig, CNPq e CAPES.

Referências

Castro, P. S., Zhang, D., e Li, S. (2012). Urban traffic modelling and prediction using large scale taxi gps traces. In *International Conference on Pervasive Computing*, pages 57–72. Springer.

Csáji, B. C., Browet, A., Traag, V. A., Delvenne, J.-C., Huens, E., Van Dooren, P., Smoreda, Z., e Blondel, V. D. (2013). Exploring the mobility of mobile phone users. *Physica A: statistical mechanics and its applications*, 392(6):1459–1473.

Cuttone, A., Lehmann, S., e Larsen, J. E. (2014). Inferring human mobility from sparse low accuracy mobile sensing data. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication*, pages 995–1004. ACM.

Frias-Martinez, V., Virseda, J., Rubio, A., e Frias-Martinez, E. (2010). Towards large scale technology impact analyses: Automatic residential localization from mobile phone-call data. In *Proceedings of the 4th ACM/IEEE international conference on information and communication technologies and development*, page 11. ACM.

Hoteit, S., Chen, G., Viana, A., e Fiore, M. (2016). Filling the gaps: On the completion of sparse call detail records for mobility analysis. In *Proceedings of the Eleventh ACM Workshop on Challenged Networks*, pages 45–50. ACM.

Isaacman, S., Becker, R., Cáceres, R., Kobourov, S., Martonosi, M., Rowland, J., e Varshavsky, A. (2011). Identifying important places in people's lives from cellular network data. In *International Conference on Pervasive Computing*, pages 133–151. Springer.

Järv, O., Ahas, R., e Witlox, F. (2014). Understanding monthly variability in human activity spaces: A twelve-month study using mobile phone call detail records. *Transportation Research Part C: Emerging Technologies*, 38:122–135.

Kung, K. S., Greco, K., Sobolevsky, S., e Ratti, C. (2014). Exploring universal patterns in human home-work commuting from mobile phone data. *PloS one*, 9(6):e96180.

Lee, S., Choi, Y., Lim, S., e Park, J. (2015). A spatio-temporal distance based clustering approach for discovering significant places from trajectory data.

Montoliu, R., Blom, J., e Gatica-Perez, D. (2013). Discovering places of interest in everyday life from smartphone data. *Multimedia tools and applications*, 62(1):179–207.

Naboulsi, D., Fiore, M., Ribot, S., e Stanica, R. (2016). Large-scale mobile traffic analysis: a survey. *IEEE Communications Surveys & Tutorials*, 18(1):124–161.

Pavan, M., Mizzaro, S., Scagnetto, I., e Beggiato, A. (2015). Finding important locations: A feature-based approach. In *Mobile Data Management (MDM), 2015 16th IEEE International Conference on*, volume 1, pages 110–115. IEEE.

Ranjan, G., Zang, H., Zhang, Z.-L., e Bolot, J. (2012). Are call detail records biased for sampling human mobility? *ACM SIGMOBILE Mobile Computing and Communications Review*, 16(3):33–44.

Rathore, M. M., Ahmad, A., Paul, A., e Rho, S. (2016). Urban planning and building smart cities based on the internet of things using big data analytics. *Computer Networks*, 101:63–80.

Schneider, C. M., Belik, V., Couronné, T., Smoreda, Z., e González, M. C. (2013). Unravelling daily human mobility motifs. *Journal of The Royal Society Interface*, 10(84):20130246.

Trestian, I., Ranjan, S., Kuzmanovic, A., e Nucci, A. (2009). Measuring serendipity: connecting people, locations and interests in a mobile 3g network. In *Proceedings of the 9th ACM SIGCOMM conference on Internet measurement*, pages 267–279. Acm.