UNIVERSIDADE FEDERAL DO CEARÁ

CAMPUS DE QUIXADÁ

CURSO DE GRADUAÇÂO EM CIÊNCIA DA COMPUTAÇÃO

MAURÍCIO CAVALCANTE BRÁZ

IMPLEMENTAÇÃO E AVALIAÇÃO DE ALGORITMOS PARA ANÁLISE DE SIMILARIDADE DE TRAJETÓRIAS NA BIBLIOTECA PYMOVE

QUIXADÁ

2021

MAURÍCIO CAVALCANTE BRÁZ

IMPLEMENTAÇÃO E AVALIAÇÃO DE ALGORITMOS PARA ANÁLISE DE SIMILARIDADE DE TRAJETÓRIAS NA BIBLIOTECA PYMOVE

Monografia apresentada ao Curso de Ciências da Computação do Campus Quixadá da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Bacharel em Ciências da Computação.

Orientadora: Prof[a]. Ma. Lívia Almada Cruz

QUIXADÁ

2021

Dados Internacionais de Catalogação na Publicação
Universidade Federal do Ceará
Biblioteca Universitária
Gerada automaticamente pelo módulo Catalog, mediante os dados fornecidos pelo(a) autor(a)

B839i Bráz, Maurício Cavalcante.
Implementação de algoritmos para análise de similaridade de trajetória na biblioteca PyMove / Maurício Cavalcante Bráz. – 2021.
44 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Ciência da Computação, Quixadá, 2021.
Orientação: Profa. Ma. Lívia Almada Cruz.

1. Análise de trajetórias. 2. Análise de dados. 3. Mineração de dados (Computação). I. Título.
CDD 004

MAURÍCIO CAVALCANTE BRÁZ

IMPLEMENTAÇÃO E AVALIAÇÃO DE ALGORITMOS PARA ANÁLISE DE SIMILARIDADE DE TRAJETÓRIAS NA BIBLIOTECA PYMOVE

Monografia apresentada ao Curso de Ciências da Computação do Campus Quixadá da Universidade Federal do Ceará, como requisito parcial para obtenção do Título de Bacharel em Ciências da Computação.

Aprovada em: __/__/____

BANCA EXAMINADORA

Profª. Ma. Lívia Almada Cruz (Orientadora)
Universidade Federal do Ceará (UFC)

Prof. Dr. Regis Pires Magalhães
Universidade Federal do Ceará (UFC)

Prof. Me. Nicksson Ckayo Arrais de Freitas
Centro Universário Christus (Unichristus)

À Maria Santíssima.

"*Ecce imperio Virginis omnia subiciuntur.*"

AGRADECIMENTOS

Primeiramente agradeço a meu caro amigo Jesus, por todo o caminho que pudemos fazer juntos até aqui, por tantas aventuras, ensinamentos, carinho, providência superabundante, pela missão que podemos compartilhar, e de maneira especial, por ter me apresentado a verdadeira face de sua Igreja, que me educou na virtude e na sabedoria, me ensinando também o precioso valor da ciência, do serviço e da caridade. A toda essa eterna família, minha gratidão.

Agradeço a Maria Santíssima, mãe da Sabedoria, a qual dedico esse humilde trabalho, pelo dom da vida, pelo zelo, pela paciência e por toda a providência de ter chegado até aqui. Agradeço também pela intercessão do meu caro São Tomás de Aquino, por me educar a partir da realidade e da ciência a perceber a Verdade.

Agradeço também ao meu pai, José Mauro, pelo exemplo de honestidade, caridade e sabedoria, à minha mãe Maria do Socorro por todo o zelo, carinho e exemplo de fortaleza e aos meus irmãos por todo o apoio, incentivo e exemplo de determinação.

À minha querida noiva Ana Cecília, por toda a unidade, carinho, zelo, atenção, inspiração, por cada oração para a boa execução deste projeto, suas correções e traduções.

À Comunidade Regina Pacis pela educação na fé, na maturidade, na responsabilidade e por todo o suporte fornecido durante a maior parte do meu curso.

Àos meus amigos com os quais tenho a honra de partilhar a bela jornada da vida e buscar, em unidade, os mesmos ideais.

À professora Lívia pela perfeita orientação, disponibilidade, paciência e por demonstrar sempre um grande zelo por este trabalho.

Ào professor Regis e ao Nicksson Arrais, pelo suporte no desenvolvimento, pelas dicas, disponibilidade e avaliação.

Cada verdade é um reflexo: por trás do reflexo, e a dar lhe o valor, está a Luz. Cada ser é um testemunho; cada fato é um segredo divino: para além está o objeto da revelação, o herói do testemunho. Toda a verdade se destaca no Infinito como vinda de um fundo de perspectiva; ela se assemelha a esse infinito, pertence a ele. Porém, uma verdade particular, por mais que queira estar em primeiro plano sempre terá as imensidóes no fundo, distantes. Poderíamos dizer que uma verdade particular é apenas um símbolo, um símbolo real, um sacramento do Absoluto. Ela se faz presente e ela existe, mas não existe por si só, não se basta; vive em função de algo, e morreria, se fosse relegada à sua própria inconsistência. Para a alma totalmente desperta, toda verdade é, portanto, um ponto de encontro. O Pensamento soberano convida o nosso para este lugar: e vamos perder esse encontro sublime? *(A.D Sertillanges)*

RESUMO

Com o crescimento do volume de dados de trajetórias, inúmeros estudos têm sido feitos para detectar novos insights e compreender acerca de padrões de movimento, classificações, predições e o comportamento de objetos no espaço geográfico. Nota-se uma grande quantidade de pesquisas sobre dados de trajetórias, porém, se percebe que não são muitas as ferramentas capazes de auxiliar pesquisadores a extrairem desses dados todas as informações que eles podem oferecer. Entre os diversos mecanismos utilizados nesse serviço, a medida de similaridade entre trajetórias demonstra grande eficiência na análise de dados de trajetória. Por isso, este trabalho realiza uma implementação de operações de recuperação de trajetórias baseadas em similaridade no tempo e no espaço, disponilizando a qualquer pessoa as operações desenvolvidas, tanto para o uso, quanto para o aperfeiçoamento delas. Para a análise das operações, diversos tipos de conjuntos de dados são utilizados. Também apresenta algumas dificuldades dessas recuperações e contextos que podem alterar o modo como podemos basea-las, além de realizar comparações com outras já desenvolvidas

Palavras-chaves: Similaridade entre trajetórias. Análise de trajetórias. Medidas de distância.

ABSTRACT

With the growth in the volume of trajectory data, numerous studies have been done to detect new insights and understand about movement patterns, classifications, predictions and the behavior of objects in geographic space. There is a large number of research on trajectory data. However, it is noticed that there are not many tools capable of helping researchers to extract from this data all the information they can offer. Among the various mechanisms used in this service, the measure of similarity between trajectories demonstrates great efficiency in the analysis of trajectory data. For this reason, this work implements operations to recover trajectories based on similarity in time and space, making available to any person the operations developed, both for use and for their improvement. For the analysis of operations, several types of data sets are used. It also presents some difficulties of these recoveries and contexts that can change the way we can base them, in addition to making comparisons with others already developed.

Key-words: Trajectory Similarity. Trajectory Analisis. Distance measures.

LISTA DE ILUSTRAÇÕES

LISTA DE QUADROS

LISTA DE ABREVIATURAS E SIGLAS

KNN *K Nearest Neighbors*

DTW *Dinamic Time Warping*

LCSS *Longest Common Subsequence*

MOTA *Multi Object Tracking Accuracy*

MEDP *Mean Euclidean Distance Predictive*

MEDT *Mean Euclidean Distance Trajectory*

SUMÁRIO

1 INTRODUÇÃO

Com o crescimento do volume de dados de trajetórias, inúmeros estudos geográficos, sociológicos e computacionais têm sido feitos para detectar novos insights e compreender acerca de padrões de movimento, classificações, predições e o comportamento de objetos no espaço geográfico, como em Zheng (2015) e Kong *et al.* (2018).

Esses estudos têm se aplicado em diversos contextos, extremamente úteis para o bem comum. Entre tais aplicações podemos citar: a análise de mobilidade sustentável (WANG; MORIARTY, 2018); estratégias de planejamento urbano, monitoramento de comportamentos irregulares de veículos, personalização de serviços baseados em trajetória (KONG *et al.*, 2018), aplicações na área da saúde, auxiliando na identificação de tumores em imagens (CHANDOLA; BANERJEE; KUMAR, 2009) através da detecção de anomalias, no rastreamento de fenômenos naturais e identificação de padrões de migração animal (ZHENG; XIE, 2011; ZHENG *et al.*, 2011).

Diversas pesquisas são realizadas a partir de dados de trajetórias (ZHENG, 2015; KONG *et al.*, 2018; Feng; Zhu, 2016) através do uso de ferramentas que possam auxiliar pesquisadores a extrair informações desses dados. Entre os diversos mecanismos utilizados, a medida de similaridade entre trajetórias demonstra grande eficiência na análise de dados de trajetória, sendo principalmente aplicada em reconhecimento de padrões, análise de dados e aprendizado de máquina (YUAN *et al.*, 2016). Com a análise de similaridade entre trajetórias é possível criar associações, auxiliar na tomada de decisão, detectar padrões, fazer previsões e auxiliar na compressão de dados. Em relação a extração de dados, a similaridade entre trajetórias pode fornecer informações muito importantes como: a previsão de movimento de objetos, monitoramento de tráfego, compreensão de atividades, detecção de anormalidades, reconstrução tridimensional, previsão do tempo e geografia (BIAN *et al.*, 2018).

De acordo com Zheng (2015), nota-se uma grande quantidade de pesquisas sobre dados de trajetórias, porém, se percebe que ainda são poucas as ferramentas capazes de auxiliar pesquisadores a extrairem desses dados todas as informações que eles podem oferecer. A partir dessa necessidade, foi desenvolvido o PyMove (OLIVEIRA, 2019; SANCHES, 2019), biblioteca de manipulação de dados espaço-temporais, com foco em trajetórias, disponível na linguagem de programação Python, desenvolvida pelo Insight Data Science Lab[1], laboratório vinculado a Universidade Federal do Ceará. Por isso, além do objetivo principal deste trabalho que é o

[1] https://www.insightlab.ufc.br/

desenvolvimento de operações de recuperação de trajetórias baseadas em similaridade no tempo e no espaço, as operações desenvolvidas também serão integradas ao PyMove.

A recuperação de trajetórias por similaridade possui diversos desafios, principalmente ao se procurar em bases de dados de trajetórias com grande volume de pontos, devido a complexidade do algoritmos comumente utilizados que calculam a distância entre trajetórias serem de O(n^2) (Magdy *et al.*, 2015). Além disso, algumas trajetórias possuem possuir grande similaridade apenas em um certo conjunto de pontos (similaridade parcial), enquanto nos outros, uma grande diferença, dificultando quando a busca tem como objetivo encontrar trajetórias que em algum momento se cruzaram, ou se tornaram praticamente uma só. As trajetórias também podem ter diferentes taxas de amostragem. Por exemplo, há trajetórias com posições coletadas a cada 5 segundos, enquanto outras podem a cada 2 minutos. Além disso, as trajetórias pode ter diferentes tamanhos e formatos. Há trajetórias longas e curtas; há trajetórias retas e outras bastante circulares, além de diversas outras variações. Cada uma dessas características influencia na busca por similaridade. Neste trabalho, considerou-se a operação que consiste em, dada uma trajetória T e um *dataset* de trajetórias, recuperar as trajetórias do *dataset* que sejam mais similares ou próximas a T, a nível espacial e temporal, considerando ainda que esta proximidade possa ser parcial e que as trajetórias tenham diferentes taxas de amostragem.

1.1 Objetivos

O objetivo principal desse trabalho é o desenvolvimento de operações de recuperação de trajetórias baseada em similaridade no tempo e no espaço.

Os objetivos específicos deste trabalho são:

a) Implementar os algoritmos e métricas de similaridade de trajetórias da literatura;

b) Desenvolver algoritmos de recuperação de trajetórias por similaridade;

c) Integrar as operações desenvolvidas ao PyMove;

d) Avaliar resultados através de estudos de caso;

1.2 Estrutura do trabalho

Este trabalho está organizado da seguinte forma: no Capítulo 2 está descrita a fundamentação teórica, juntamente com os conceitos necessários para o entendimento deste

trabalho; no Capítulo 3 são descritos os trabalhos relacionados a este trabalho e é apresentado uma comparação entre eles; no Capítulo 4 são apresentados os procedimentos e a metodologia executada nesse trabalho; por fim, no Capítulo 5 são apresentados estudos de caso e, no Capítulo 6, as conclusões obtidas com o desenvolvimento deste trabalho.

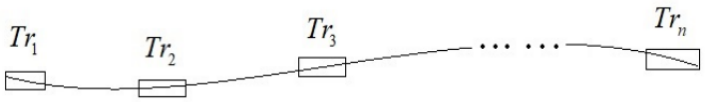
2 FUNDAMENTAÇÃO TEÓRICA

Esta seção tem como fim apresentar os principais conceitos abordados neste trabalho, assim como as ferramentas que serão utilizadas. Primeiramente busca-se obter uma visão geral do estudo de trajetórias, compreendendo primariamente como elas são definidas. Em seguida, o modo como dados de trajetórias podem ser minerados. Por fim, as técnicas mais frequentemente adotadas na literatura para recuperação de trajetórias baseando-se na similaridade.

2.1 Trajetórias

Uma trajetória é definida por Zheng (2015) como um um traço gerado a partir de um objeto em movimento em um espaço geográfico. Como está representada na Figura 1, uma trajetória é composta por um conjunto de pontos ordenados, por exemplo, $Trajectoria = (T_{r_1}, T_{r_2}, ..., T_{r_n})$, onde cada ponto é definido como $T_{r_i} = (x_i, y_i, t_i)$ seja x e y as coordenadas geoespaciais e t um *timestamp* do momento em que o ponto foi gerado. Em algumas circunstâncias específicas, outras propriedades relevantes sobre o movimento do objeto podem ser adicionadas, como velocidade, direção ou aceleração (CAI; LEE; LEE, 2016; WANG; MORIARTY, 2018).

Figura 1 – Trajetória gerada a partir de um dispositivo de GPS.



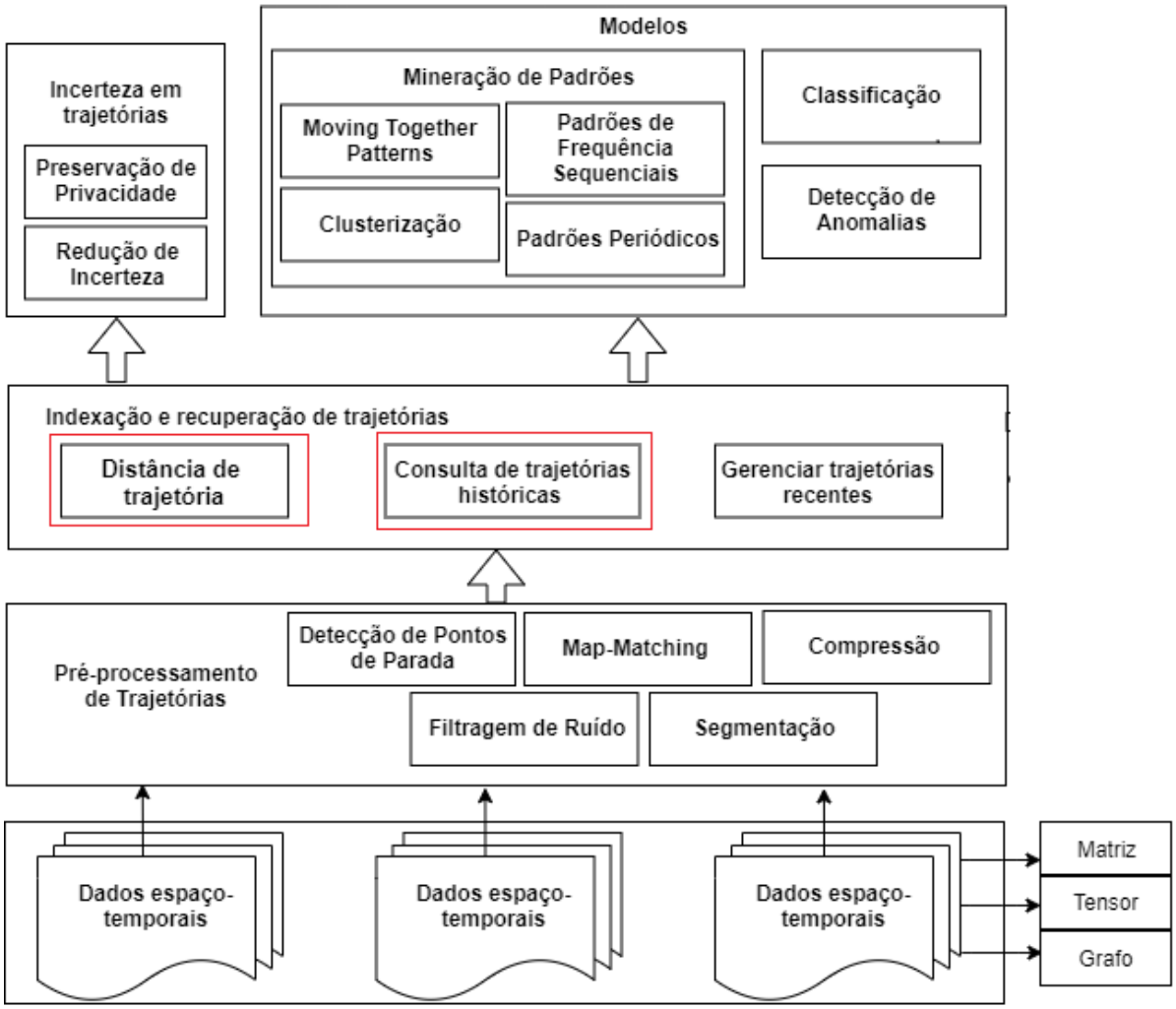
Fonte – (BIAN *et al.*, 2018)

2.2 Mineração de dados de trajetórias

Sabendo então o que são trajetórias, pode-se explorar os diversos modos de trabalhar com elas. O paradigma apresentado por Zheng (2015) que envolve a mineração de dados de trajetórias atualmente é representado na Figura 2, tendo como destaque as etapas correspondentes a similaridade entre trajetórias.

Para trabalhar com dados de trajetórias e analisá-los, podemos aplicar uma série de etapas, que são agrupadas em categorias e áreas de estudo. Na Figura 3, (ZHENG, 2015) descreve o paradigma atual da mineração de dados de trajetórias, que é composto por seis módulos, sendo

Figura 2 – Paradigma da mineração de dados de trajetórias



Fonte – Adaptado de (ZHENG, 2015)

um deles a indexação e recuperação de trajetórias, que será a etapa aprofundada neste trabalho, por conter a recuperação de trajetórias baseada na similaridade como um de seus tópicos. Os seis módulos são definidos a seguir:

- **Pré-processamento:** melhoria da qualidade dos dados recebidos através de quatro técnicas principais: filtragem de ruído, detecção de pontos de parada, compressão e segmentação de trajetória.

- **Indexação e Recuperação de Trajetórias:** Indexar e recuperar trajetórias através de buscas. Entre os principais tipos de buscas se encontram o *K-Nearest Neighbor* (KNN) e a *Range*. A recuperação de trajetórias é o foco deste trabalho.

- **Incerteza em Trajetórias:** modelar e reduzir incertezas de trajetórias, e promover a privacidade de usuários que fornecem seus dados espaço-temporais.
- **Classificação:** classificar o tipo da trajetória, ou seja, se a trajetória pertence a uma bicicleta, uma pessoa que caminha, um carro etc.
- **Detecção de Anomalias:** etapa de percepção de itens espaço-temporais que não seguem um comportamento padrão.
- **Mineração de Padrões:** trabalha com a percepção de padrões de localização, de movimentação individual ou de um grupo de trajetórias.

2.3 Similaridade de trajetórias

Consiste em funções que comparam o quanto duas ou mais trajetórias são semelhantes, de acordo com os parâmetros fornecidos. Esse tipo de função é muito útil para recuperação, classificação, agrupamento em clusters e outras tarefas de consulta e mineração de dados de trajetórias. Por exemplo, podemos analisar trajetórias de clientes em um supermercado para encontrar padrões de movimento semelhantes para um melhor gerenciamento de produtos, encontrar padrões de migração freqüentes de um grupo migratório de aves, encontrar movimentos de objetos suspeitos e padrões de trajetória raros, recomendar rotas de viagem, recomendar pontos de interesse para pessoas baseado nos locais visitados, curtidas, além de prever fenômenos como tempestades, furacões e terremotos (Magdy *et al.*, 2015).

Existem duas abordagens para consultas baseadas em similaridade de trajétórias: a forma total, onde se avalia a similaridade considerando todos os pontos de ambas as trajetórias, ou a forma parcial, onde se busca por pontos entre as trajetórias que apresentaram similaridade.

Existem diversas medidas de similaridade de trajetórias, e elas são divididas em duas categorias principais:

- **Similaridade Espacial:** focada em encontrar trajetórias com formas geométricas semelhantes, ignorando a dimensão temporal, ou seja, os *timestamps*;
- **Similaridade Espaço-Temporal:** leva em consideração também a dimensão temporal ao medir a semelhança entre trajetórias.

A seguir, exploramos algumas das medidas de similaridade de trajetórias que estão sendo utilizadas ao longo dos anos, conforme Quehl et al. (2017).

2.3.1 Distância Euclidiana

São medidas baseadas na distância euclidiana média. Existem duas principais: a MEDT (*Mean Euclidian Distance Trajectory*), é uma similaridade baseada em séries temporais, e a MEDP (*Mean Euclidian Distante Predictive*) é uma similaridade baseada em dados espaciais. A MEDT necessita que sejam definidos pontos respectivos ao tempo de cada trajetória. A equação utilizada para defini-la é a seguinte:

$$m_{MEDT}(T_p, T_g) = \frac{1}{n}\sum_{k=0}^{n-1} ||P_p, t_{p_{1+k}} - P_g, t_{p_{1+k}}||2. \quad (1)$$

Onde P_p e t_{p1} são respectivamente um ponto e um intervalo de tempo correspondente a trajetória T_g e P_g e t_{p2} são respectivamente um ponto e um intervalo de tempo correspondente a trajetória T_g. Um problema que podemos perceber no MEDT é a possibilidade de se obter como resultado um valor maior do que 0, mesmo se as trajetórias sejam idênticas. Isso ocorre devido ao fato de que as trajetórias são subdivididas de acordo com o tempo, podendo assim ter valores divergentes devido a possível variância de velocidade de veículos que percorrem o mesmo trajeto.

Já a equação MEDP mede a diferença entre os caminhos preditos. Para cada ponto P_p, de uma trajetória predita T_p com os intervalos $1...N_p$, MEDP buscará pelo ponto mais próximo da trajetória verdadeira e avaliar a distância euclidiana desse ponto. O algoritmo é definido do modo seguinte:

$$m_{MEDP}(T_p, T_g) = \frac{1}{n}\sum_{k=0}^{n-1} ||(P_p, k), (P_g, x)||2. \quad (2)$$

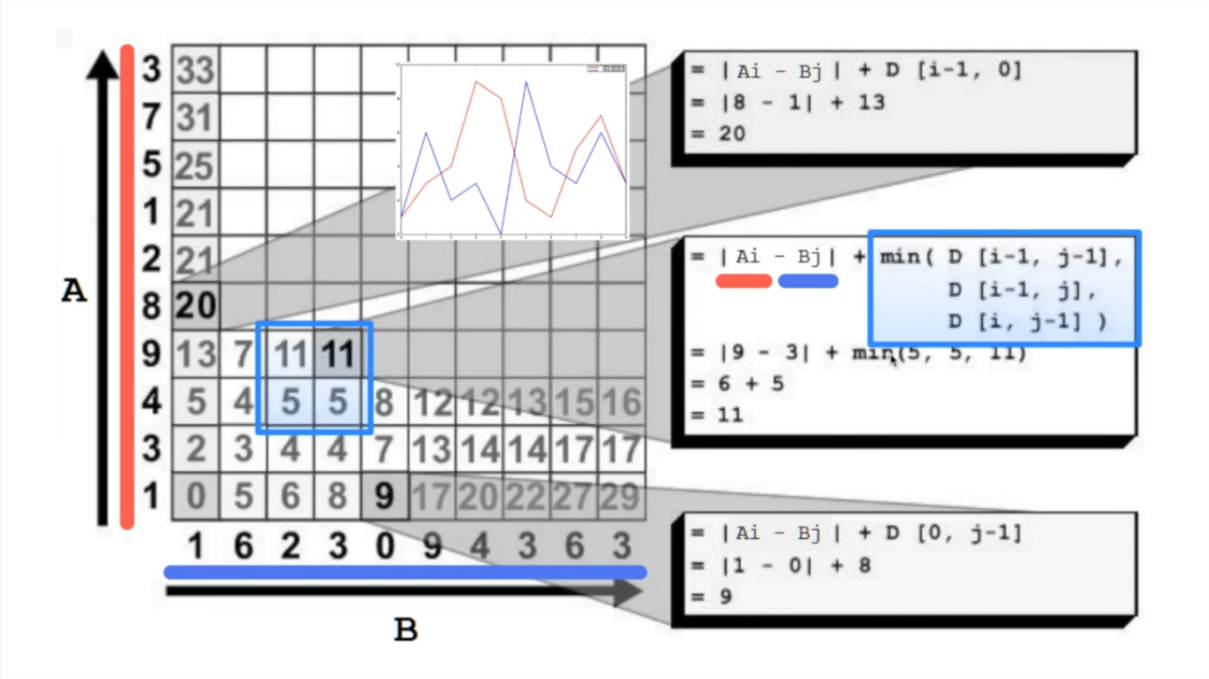
Desse modo, quando duas trajetórias comparadas diferem apenas na velocidade, MEDP terá como retorno uma alta similaridade, enquanto MEDT retornará uma baixa similaridade (Quehl *et al.*, 2017).

2.3.2 *Dynamic Time Warping* (DTW)

Esta medida, baseada em séries temporais, pode ser aplicada mesmo quando as trajetórias possuem tamanhos diferentes. Para fazer isso, o DTW encontra primeiro para cada ponto em T_p um ponto correspondente em T_g. Em seguida, a programação dinâmica é usada para encontrar um time warping que minimiza a distância total entre esses pares de pontos. O DTW avalia o comprimento do warping necessário para distorcer T_p em T_g. Isso ocorre através

da criação de uma tabela onde cada célula é calculada como descreve a Figura 5 (SOUZA; PANTOJA; SOUZA, 2009; Quehl *et al.*, 2017).

Figura 3 – Criação da tabela de Dynamic Time Warping



Fonte – (KöRTING, 2017. Disponível em: https://www.youtube.com/watch?v=_K1OsqCicBY Acesso em: 07 out. 2020)

Em seguida, se escolhe o *warping path*, partindo do valor superior direito, e escolhendo o menor valor entre a célula inferior, a inferior esquerda e a esquerda, sucessivamente, até chegar a célula mais à esquerda. A Figura 6 descreve melhor o funcionamento dessa fase do processo, sendo as células destacadas o *Warping Path* do DTW.

O equação pode ser definida do seguinte modo:

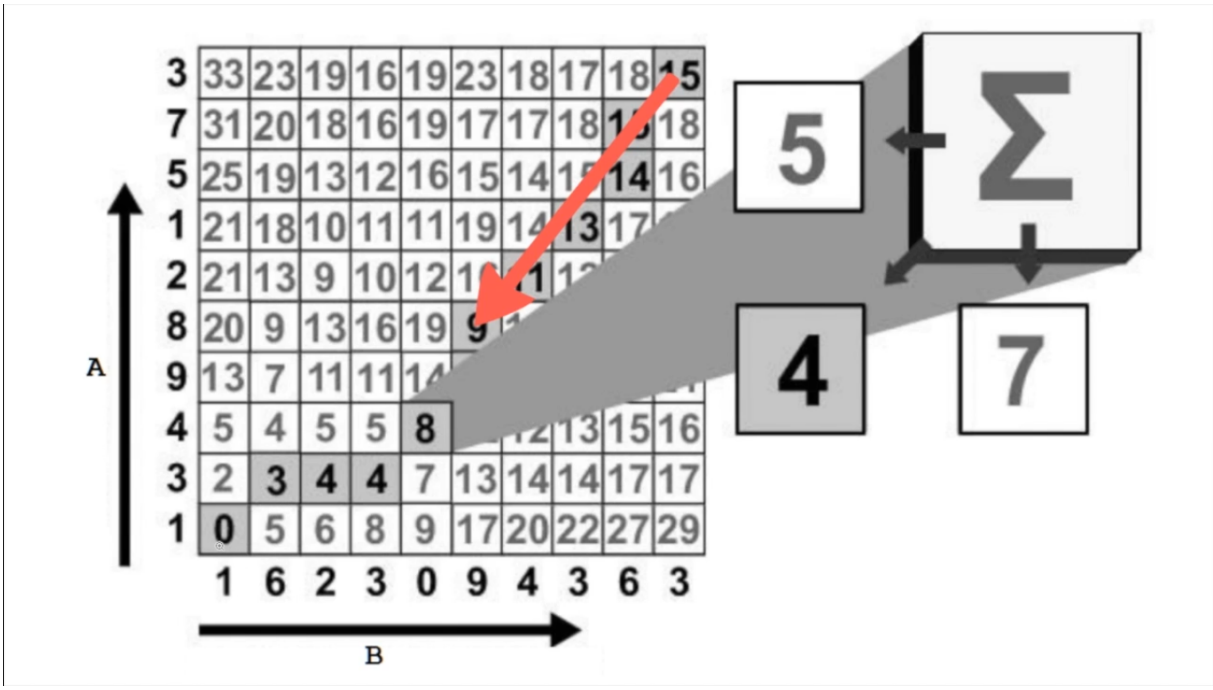
$$DTW(S,T) = min_w[\sum_{k=1}^{p} \delta(w_k)] \tag{3}$$

onde:

$$\delta(i,j) = |s_i - t_j|$$
$$\delta(i,j) = (s_i - t_j)^2 \tag{4}$$

Sendo S e T trajetórias distintas, onde s e t são seus respectivos pontos e i e j seus respectivos índices.

Figura 4 – Escolha do Warping Path do algoritmo Dynamic Time Warping



Fonte – (KöRTING, 2017. Disponível em: https://www.youtube.com/watch?v=_K1OsqCicBY Acesso em: 07 out. 2020)

2.3.3 Maior Subsequência Comum (LCSS - *Longest Common Subsequence*)

Similaridade baseada em séries temporais, que calcula, como o nome sugere, a maior subsequência comum entre duas trajetórias. O LCSS opera a partir do incremento de um valor de retorno, que é iniciado como 0, cada vez que a avaliação do tamanho do trecho dos trajetos em análise estiver abaixo da variável limite de similaridade de tamanho, fornecida como parâmetro da função, e quando a distância entre cada ponto equivalente do trajeto analisado for menor do que o limite de similaridade de ponto fornecida como outro parâmetro na função.

É descrito na Figura **??** descreve o cálculo da medida, onde as variáveis δ e ε regulam o quão próximo as trajetórias/pontos, respectivamente, devem ser similares para incrementarem o possível resultado (Quehl *et al.*, 2017).

2.3.4 *Hausdorff Distance*

Essa medida, baseada em formas geométricas, consiste em encontrar, para cada ponto de T_g o ponto mais próximo de T_p, somar seus valores, e em seguida encontrar, para cada ponto de T_p o ponto mais próximo de T_g, escolhendo como resultado a menor soma entre eles. O

Figura 5 – Cálculo de Maior Subsequência Comum

$$LCSS_{\delta,\varepsilon}(T_A, T_B) = \begin{cases} 0, & \text{if } T_A \text{ or } T_B \text{ is empty} \\ 1 + LCSS_{\delta,\varepsilon}(Head(T_A), Head(T_B)), & \text{if } |m-k| \leq \delta \text{ and } |a_{m,1} - b_{k,1}| \leq \varepsilon \\ & \text{and } \ldots \text{ and } |a_{m,n} - b_{k,n}| \leq \varepsilon \\ max(LCSS_{\delta,\varepsilon}(Head(T_A), T_B), & \\ \quad LCSS_{\delta,\varepsilon}(T_A, Head(T_B)), & otherwise, \end{cases}$$

Fonte – (Quehl *et al.*, 2017)

cálculo da medida se descreve do seguinte modo:

$$m_{HAU}(T_p, T_g) = max\{d_h(T_p, T_g),\ d_h(T_g, T_p)\}$$
$$d_h(T_p, T_g) = \max_{p_p \in T_p} \min_{p_g \in T_g} ||p_p - p_g||2$$

Pode ser utilizada para trajetórias com tamanhos diferentes, porém, sofre com os mesmos defeitos de outras medidas de similaridade: Trajetórias que possuem uma grande proximidade espacial podem sempre serem classificadas como muito similares mesmo se o movimento realizado for diferente, inclusive quando possuem direções opostas (Quehl *et al.*, 2017).

2.3.5 CLEAR - *Multi Object Tracking Accuracy (MOTA)*

Medida de similaridade de dados espaciais. Assim como no LCSS se atribui valores de correspondência para cada ponto predito similar. O cálculo dessa medida é definido do seguinte modo:

$$m_{MOTA}(T_p, T_g) = 1 - \frac{\sum_t (m_t + fp_t + mme_t)}{n_p} \tag{5}$$

Sendo m_t, fp_t e mme_t, respectivamente, número de erros, falso positivos e desencontros quanto se compara T_p e T_g. Um erro nesse contexto ocorre para cada ponto em Tg que não combina com T_p. Um falso positivo é definido como uma combinação para um ponto que está muito distante (definido por um limite) que constitui uma correspondência. Um desencontro ocorre quando mais de um ponto de T_p corresponde ao mesmo ponto em T_g.

A diferença de velocidade das trajetórias também influencia no resultado da avaliação. Ao contrário da LCSS, nessa medida, a fração total de correspondências é considerada em vez da sequência mais longa de correspondências (BERNARDIN; STIEFELHAGEN, 2008; Quehl *et al.*, 2017).

3 TRABALHOS RELACIONADOS

Nesta seção apresentamos alguns estudos relacionados a este trabalho. Eles serão divididos em dois tipos: trabalhos que implementam clusterização de trajetórias baseada em similaridade e trabalhos que implementam consultas baseadas em similaridade de trajetórias.

3.1 Clusterização de trajetórias baseada em similaridade

Quehl *et al.* (2017) apresentam diversos algoritmos para medida de similaridade de trajetórias, com eficácia definida pelo contexto do conjunto de dados analisado. A finalidade do trabalho é calcular quão satisfatória é uma predição de trajetória, para facilitar decisões de rotas de carros autônomos. Atualmente não existe uma medida de similaridade que seja classificada como a melhor para todas as tarefas de predição de trajetória. O artigo propõe uma abordagem que sintetiza uma medida híbrida, a partir de um conjunto de medidas de similaridade de trajetória diversas fornecendo uma heurística para determinar os parâmetros da abordagem proposta. Neste trabalho, precisamos medir também a similaridade entre as trajetórias podendo assim recuperar aquelas que estiverem dentro do nível de similaridade desejada.

TAD... (2020) propõem um novo algoritmo de clusterização de trajetórias, capaz de extrair pontos de permanência baseando-se na análise de densidade espaço-temporal dos dados. Utiliza-se duas novas métricas: função de densidade NMAST (capacidade de movimento de vizinhança e tempo de permanência) e fator NT (tolerância ao ruído). Em primeiro lugar, o NMAST integra as características da vizinhança Move Ability (NMA, estendida do conceito de Move Ability MA), tempo de permanência (ST) e fator de avaliação Eμ para medir a densidade espaço-temporal dos dados.

Em segundo lugar, o NT utiliza os recursos de ruído para avaliar e reduzir dinamicamente a influência do ruído. Os resultados experimentais no conjunto de dados Geolife[2] mostram que as distribuições ocultas nos dados são melhor extraídas, especialmente para várias trajetórias complexas ou especiais com lacunas de longa duração. Nesse trabalho, para que a clusterização de trajetórias ocorra, é levado em consideração o tempo em que cada ponto da trajetória foi gerado. Do mesmo modo, no trabalho aqui proposto, a dimensão temporal poderá também ser levada em conta, devido ao fato de que o momento em que cada ponto de

[2] O conjunto de dados Geolife foi coletado de abril de 2007 a agosto de 2012 pela Microsoft Research Asia. No total, esse conjunto de dados contém 1,7621 trajetórias de 182 usuários e é representado por uma série de pontos com timestamps registrados a cada 5 - 10m ou 1 - 5s. A frequência de amostragem do ponto de dados de trajetória neste artigo é 5s.

uma trajetória é registrado pode ser considerado para a construção de uma clusterização. (BIAN *et al.*, 2018).

Li *et al.* (2018) realizam um trabalho a partir da observação dos sistemas de identificação automática (AISs), que servem como um complemento aos sistemas de radar e que são instalados e amplamente utilizados a bordo de navios para identificar alvos e melhorar a segurança da navegação com base em um esquema de comunicação de dados de alta frequência. Neste trabalho foi visto que a mineração de dados de trajetória é uma importante direção de pesquisa para extrair informações úteis com alta precisão e baixos custos computacionais, mapear trajetórias e promover métodos de clusterização. Foi utilizado a *Merge Distance* (MD) para medir as semelhanças entre diferentes trajetórias, e o Multidimensional Scaling (MDS) para construir uma expressão espacial de baixa dimensão adequada das semelhanças entre trajetórias. Um DBSCAN também é utilizado para agrupar pontos espaciais para adquirir o cluster ideal. Uma fusão dos algoritmos MD, MDS e DBSCAN aprimorados identifica o curso das trajetórias para obter um melhor desempenho de clusterização. Os experimentos também mostram que o método implementado apresenta uma precisão mais alta que os clássicos, como clusterização espectral e clusterização por propagação por afinidade. O modelo implementado por Li *et al.* (2018) possui métricas de distância implementadas semelhantes as propostas neste trabalho para calcular a similaridade total entre as trajetórias.

3.2 Consultas baseadas em similaridade de trajetórias

Baldus e Bringmann (2018. Disponível em: http://arxiv.org/abs/1803.00806 Acesso em: 03 fev. 2021) descrevem uma implementação de consultas rápidas por vizinhos próximos do tipo *Range*, utilizando como medida de distância a de Fréchet. O algoritmo é projetado para ser eficiente em trajetórias com bastante continuidade, como as de GPS. Foi utilizada uma estrutura de dados *quadtree* para enumerar todas as curvas no banco de dados que têm pontos de início e extremidade semelhantes à curva de consulta. Nessas curvas, foram executados filtros positivos e negativos para limitar o conjunto de possiveis resultados. Apenas para aquelas trajetórias onde essas heurísticas falham, foi calculada a distância de Fréchet, executando uma nova variante recursiva de um algoritmo de diagrama de espaço livre clássico. Neste trabalho também é utilizado consultas do tipo de Range e nos estudo de caso é utilizado um *dataset* com trajetórias de GPS.

Xu, Lu e Güting (2017) estudaram trajetórias onde, cada uma delas, continha uma

sequência de locais com seus respectivos *timestamps* e um conjunto de atributos característicos. Eles realizaram consultas de intervalo, assim como neste trabalho, que retornam trajetórias contendo valores de atributos particulares e passando por uma determinada área durante o tempo de consulta. Também integraram trajetórias e atributos padrão em uma estrutura unificada e propuseram uma estrutura de índice, bem como o algoritmo de consulta. A estrutura é geral e flexível em termos de tratamento de trajetórias de multi-atributos e trajetórias padrão, respondendo a uma variedade de consultas e oferecendo suporte a aplicativos de atualização intensiva. A avaliação da consulta é conduzida em um sistema de banco de dados de protótipo e resultados experimentais, demonstrando que o método desenvolvido supera métodos alternativos por um fator de 3-10 em um conjunto de dados de um milhão de trajetórias reais e valores de atributos sintéticos.

Shi *et al.* (2019) propõem um *framework* de processamento de consulta probabilística do tipo *Range* de objetos em movimento na rede viária. Uma estrutura de índice espaço-temporal é proposta com base nesse *framework*. Este índice pode representar efetivamente pesos de tempo e a relação entre seções de estradas. O algoritmo de consulta *Range* de intervalo probabilístico de objetos em movimento na rede rodoviária com uma trajetória incerta causada pela frequência de amostragem é projetado e implementado. Os experimentos verificam que o método proposto por Shi *et al.* (2019) pode melhorar a eficiência das consultas e garantir a precisão das mesmas. Neste trabalho, consultas do tipo *Range* também são realizadas.

O Quadro 1 apresenta a comparação deste trabalho e os trabalhos desenvolvidos em Quehl *et al.* (2017), TAD... (2020), Li *et al.* (2018), Baldus e Bringmann (2018. Disponível em: http://arxiv.org/abs/1803.00806 Acesso em: 03 fev. 2021), Xu, Lu e Güting (2017) e Shi *et al.* (2019). Neste são avaliados os seguintes pontos:

- **Taxas de amostragem:** Tamanho das trajetórias que podem ser aceitas pelas operações desenvolvidas pelo trabalho;
- **Disponibilidade:** Disponibilidade do acesso as operações implementadas pelo trabalho para download e/ou uso;
- **Tipos de operação:** Tipos de operações implementadas pelo trabalho;
- **Densidade aceita:** Densidade de trajetórias aceita pelas operações implementadas pelo trabalho;

Quadro 1 – Comparação entre os trabalhos relacionados e o proposto.

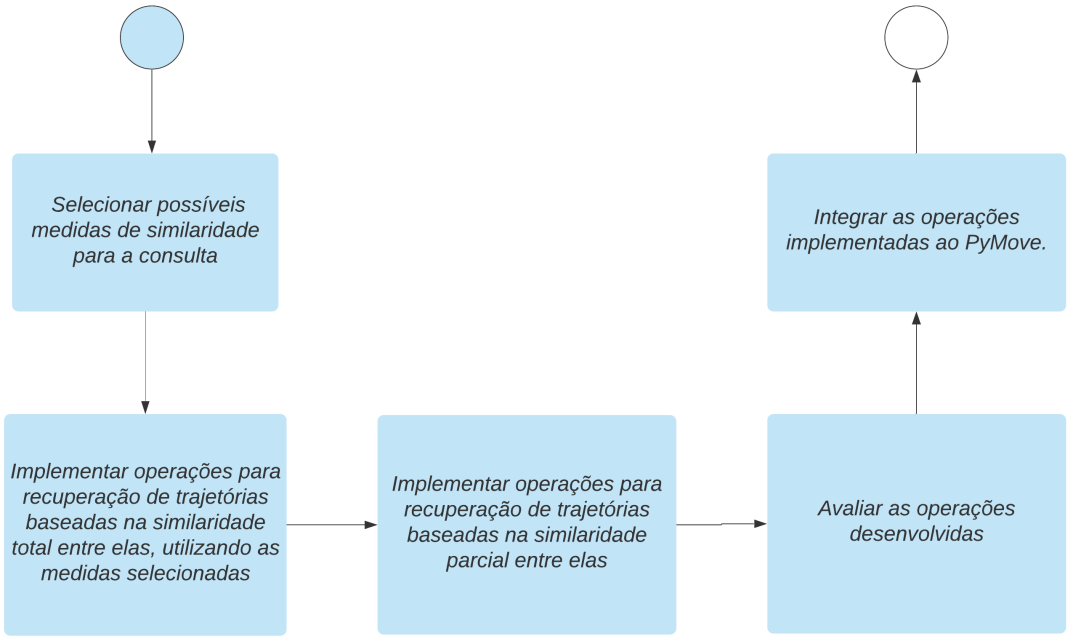
Trabalhos	**Taxas de amostragem**	**Disponibilidade**	**Tipos de operação**	**Densidade aceita**
Trabalho proposto	Qualquer taxa	Disponível através do PyMove	*Range* e KNN com similaridade total e *Range* com similaridade parcial	Qualquer densidade
(Quehl *et al.*, 2017)	Qualquer taxa	Não é informado se está disponível	Medida de similaridade GPE (Análise de predição generalizada)	Qualquer densidade
(TAD. . . , 2020)	Não foi testado com trajetórias de baixa amostragem	Não é informado se está disponível	Algoritmo de clusterização de trajetórias com função de densidade NMAST e fator NT	Não foi testado com trajetórias de baixa densidade
(Li *et al.*, 2018)	Qualquer taxa	Não é informado se está disponível	Uma fusão dos algoritmos MD, MDS e DBSCAN de forma aprimorada	Testado apenas com trajetórias com alta densidade
(BALDUS; BRINGMANN, 2018. Disponível em: http://arxiv.org/abs/1803.00806 Acesso em: 03 fev. 2021)	Qualquer taxa	Disponível para download	Busca por similaridade usando *Range* com medida de distância de Frechét	Desenvolvido para trajetórias de alta densidade e testado apenas com trajetórias de alta densidade
(XU; LU; GüTING, 2017)	É necessário que as trajetórias possuam uma grande quantidade de atributos	Não é informado se está disponível	Operação *Range* para trajetórias com uma grande quantidade de atributos	Qualquer densidade
(SHI *et al.*, 2019)	Qualquer taxa	Não é informado se está disponível	*Framework* para operação do tipo *Range*	Qualquer densidade

Fonte: Elaborado pelo autor.

4 PROCEDIMENTOS METODOLÓGICOS

Nesta seção são apresentados os procedimentos metodológicos para a execução deste trabalho. Na Figura 6 são apresentados os seguintes passos para a execução do trabalho: i) Selecionar possíveis medidas de similaridade para a consulta; ii) Implementar operações para recuperação de trajetórias baseadas na similaridade total entre elas, utilizando as medidas selecionadas; iii) Implementar operações para recuperação de trajetórias baseadas na similaridade parcial entre elas; iv) Avaliar as operações desenvolvidas; v) Integrar as operações implementadas ao PyMove.

Figura 6 – Procedimentos Metodológicos



Fonte – O próprio autor

4.1 Selecionar possíveis medidas de similaridade para a consulta

Foi realizada uma busca literária para identificar os principais algoritmos de medida de similaridade, onde o resumo das principais medidas encontradas seja neste trabalho, na seção de fundamentação teórica. Diante do tipo de dado de trajetória utilizado na ferramenta PyMove, composto apenas de informações de pontos no espaço, e pontos espaço-temporais,

foram implementadas as medidas de similaridade exploradas neste trabalho que são baseadas na distância euclidiana (MEDP e MEDT).

4.2 Implementar operações para recuperação de trajetórias baseadas na similaridade total entre elas, utilizando as medidas selecionadas

Foi realizada uma análise utilizando *datasets* de trajetórias com características bem distintas. Um *dataset* com trajetórias que possuem movimento livre, ou seja, trajetórias que não são limitadas a percursos de movimentação limitada, como estradas, ruas, rodovias, etc, enquanto o outro *dataset* possui trajetórias com percursos de movimentação limitada, inserido dentro de uma rede de ruas.

Duas operações foram implementadas para cada tipo de medida de similaridade selecionada: *Range* e KNN. A operação Range consiste em selecionar uma trajetória, pertencente a um conjunto de trajetórias, informando uma taxa minima de similaridade, e a partir dessa taxa, selecionar as trajetórias pertencentes ao conjunto que possuirem similaridade igual ou inferior a taxa escolhida. A operação KNN (*K Nearest Neighbor*) consiste também em selecionar uma trajetória pertencente a um conjunto de trajetórias, mas dessa vez, se informa um valor k. O algoritmo KNN irá retornar as k trajetórias do conjunto mais próximas a trajetória inicialmente selecionada.

Foi implementada a operação *Range*, com sua assinatura descrita na Figura 7, utilizando MEDT como medida de similaridade e outra operação *Range* utilizando MEDP como medida de similaridade. A função recebe como parâmetros uma trajetória $traj$, um $DataFrame$ com diversas trajetórias distintas, além do tipo de distância a ser utilizada, o valor mínimo de similaridade, e o nome das *labels* correspondentes ao identificador, a latitude, a longitude e o $timestamp$ de cada ponto.

Figura 7 – Assinatura da operação *range* para similaridade completa entre trajetórias

```
def range_query(
    traj: DataFrame,
    move_df: DataFrame,
    _id: Optional[Text] = TRAJ_ID,
    min_dist: Optional[float] = 1000,
    distance: Optional[Text] = MEDP,
    latitude: Optional[Text] = LATITUDE,
    longitude: Optional[Text] = LONGITUDE,
    datetime: Optional[Text] = DATETIME
)
```

Fonte – O próprio autor

Também foi implementada a operação KNN, com sua assinatura descrita na Figura 8, tendo uma operação para cada tipo de medida de similaridade selecionada no primeiro procedimento metodológico. Assim como a operação *range*, a operação recebe como parâmetros uma trajetória $traj$, um *DataFrame* com diversas trajetórias distintas, além do tipo de distância a ser utilizada, a quantidade de k trajetórias mais similares a $traj$ e o nome das *labels* correspondentes ao identificador, a latitude, a longitude e o *timestamp* de cada ponto.

Figura 8 – Assinatura da operação KNN para similaridade completa entre trajetórias

```
def knn_query(
    traj: DataFrame,
    move_df: DataFrame,
    k: Optional[int] = 5,
    id_: Optional[Text] = TRAJ_ID,
    distance: Optional[Text] = MEDP,
    latitude: Optional[Text] = LATITUDE,
    longitude: Optional[Text] = LONGITUDE,
    datetime: Optional[Text] = DATETIME
)
```

Fonte – O próprio autor

4.3 Implementar operações para recuperação de trajetórias baseadas na similaridade parcial entre elas

Foi desenvolvida uma operação do tipo *Range* que não busca uma similaridade total, mas procura por qualquer similaridade parcial que houver entre as trajetórias, tendo como

parâmetros para filtrar os pontos considerados similares, a distância minima espacial em metros e a distância mínima temporal.

A operação, recebe como entrada uma trajetória T, uma lista de trajetórias L, a distancia minima espacial, e a distancia minima temporal. A partir desses parâmetros, é realizada uma busca em L por todos os pontos que possuírem uma proximidade, espacial ou temporal, menor ou igual aos parâmetros de distância fornecidos. Cada ponto é então adicionado ao *Move Dataframe* de retorno, que irá conter a distância espacial entre o ponto retornado,e o ponto da trajetória T espacialmente mais próximo, e os respectivos dados desse ponto (latitude, longitude e *timestamp*). A seguir, o pseudo-código da operação:

```
sim_parcial(T, L, distancia_min_espacial, distancia_min_temporal) {
    resultado = MoveDataFrame([]);
    para cada ponto_T da trajetória T {
        pontos_retornados = _meters_filter(ponto_T, L,
            distancia_min_espacial);
        pontos_retornados = _datetime_filter(ponto_T, pontos_retornados,
            distanca_min_temporal);
        resultado.append(pontos_retornados);
    }
    retorne resultado;
}
```

4.4 Avaliar as operações desenvolvidas

Para avaliar as medidas de similaridade e as operações de recuperação de trajetórias baseando-se na similaridade total, foi realizada uma análise utilizando *datasets* de trajetórias com características bem distintas. Um *dataset* com trajetórias que possuem movimento livre, ou seja, trajetórias que não são limitadas a percursos de movimentação limitada, como estradas, ruas, rodovias, etc, enquanto o outro *dataset* possui trajetórias com percursos de movimentação limitada, inserido dentro de uma rede de ruas.

Para a avaliação das operações para recuperação de trajetórias baseadas na similaridade parcial, foi realizada uma análise utilizando *datasets* de trajetórias com características distintas. Nesse caso, um *dataset* com trajetórias com uma continuidade maior

(dados de GPS), enquanto o outro *dataset* possui trajetórias com uma continuidade menor, pois é gerada a partir de diversos sensores distribuidos em uma cidade.

O uso de diferentes tipos de *datasets* ocorreu para facilitar a percepção da precisão que o tipo de similaridade escolhido possui no processo de recuperação.

4.5 Integrar as medidas de similaridade e as operações implementadas ao PyMove

As medidas implementadas, MEDP e MEDT, foram devidamente integradas a biblioteca PyMove, no pacote *utils.distances*, sendo devidamente adaptadas aos padrões exigidos para integração de novas funcionalidades à biblioteca.

Para as operações consulta de similaridade entre trajetórias implementadas, um novo pacote foi criado no PyMove: o pacote *query*. Nele as operações de consulta baseadas em similaridade total foram integradas assim como as consultas baseadas em similaridade parcial também serão.

5 RESULTADOS

Neste capítulo serão apresentados os resultados das análises das medidas de similaridade e operações de recuperação de trajetórias desenvolvidas.

5.1 Avaliação de medidas existentes e recuperação baseada em similaridade total

Neste estudo de caso foram utilizadas as medidas de similaridade MEDP e MEDT e, para que houvesse uma compreensão melhor da qualidade dos algoritmos, duas operações de recuperação de trajetórias baseadas na similaridade total foram então desenvolvidas, sendo uma do tipo *Range* e outra *K*NN. As duas operações podem utilizar similaridade a MEDP ou a MEDT, ambas as operações analisam a similaridade total entre as trajetórias.

Para este estudo de caso e visualização das operações implementadas o *dataset* selecionado é composto por um conjunto de trajetórias de Furacões e Tufões fornecido pela *National Oceanic and Atmospheric Administration*[3] (NOAA). Eles realizam uma análise pós-tempestade de cada ciclone tropical na bacia do Atlântico (ou seja, Oceano Atlântico Norte, Golfo do México e Mar do Caribe) e do Oceano Pacífico Norte para determinar a avaliação oficial do histórico de ciclones. Essa análise faz uso de todas as observações disponíveis, incluindo aquelas que podem não estar disponíveis em tempo real. Além disso, conduzem revisões contínuas de quaisquer análises retrospectivas de ciclones tropicais trazidas à sua atenção e atualizam regularmente o registro histórico para refletir as mudanças introduzidas. Eles publicam o histórico do banco de dados de ciclones tropicais em um formato conhecido como HURDAT, abreviação de *Hurricane Database*. Esses bancos de dados contêm informações de seis horas sobre a localização, ventos máximos, pressão central e (começando em 2004) o tamanho de todos os ciclones tropicais e subtropicais conhecidos.

Para o estudo de caso realizado em específico, apenas trajetórias de furacões e tufões, entre os anos 2012 e 2015, do Oceano Pacífico foram selecionados para compor o *dataset*. A trajetória do furacão Gonzalo[4] foi selecionada como base para que as operações *Range* e KNN retornem como resultado as trajetórias mais próximas que estiverem dentro dos requisitos de similaridade de cada tipo de operação. Para cada operação, dois diferentes graus de similaridade

[3] Agência científica americana do Departamento de Comércio dos Estados Unidos que se concentra nas condições dos oceanos, dos principais cursos de água e da atmosfera. Site disponível em: https://www.noaa.gov/

[4] Um poderoso ciclone tropical do Atlântico, que atingiu a escala Saffir-Simpson de categoria 4, provocando transtorno em parte do Canadá, Porto Rico, e destruição nas Pequenas Antilhas, Bermuda, Territórios britânicos ultramarinos, Ilhas britânicas, e norte da Europa, em outubro de 2014.

foram utilizados. Para a visualização do resultado das operações, o módulo de visualização da biblioteca PyMove foi utilizado.

A trajetória do furacão Gonzalo, representada na Figura 9, foi selecionado para ser a *query* dos testes das medidas de similaridade.

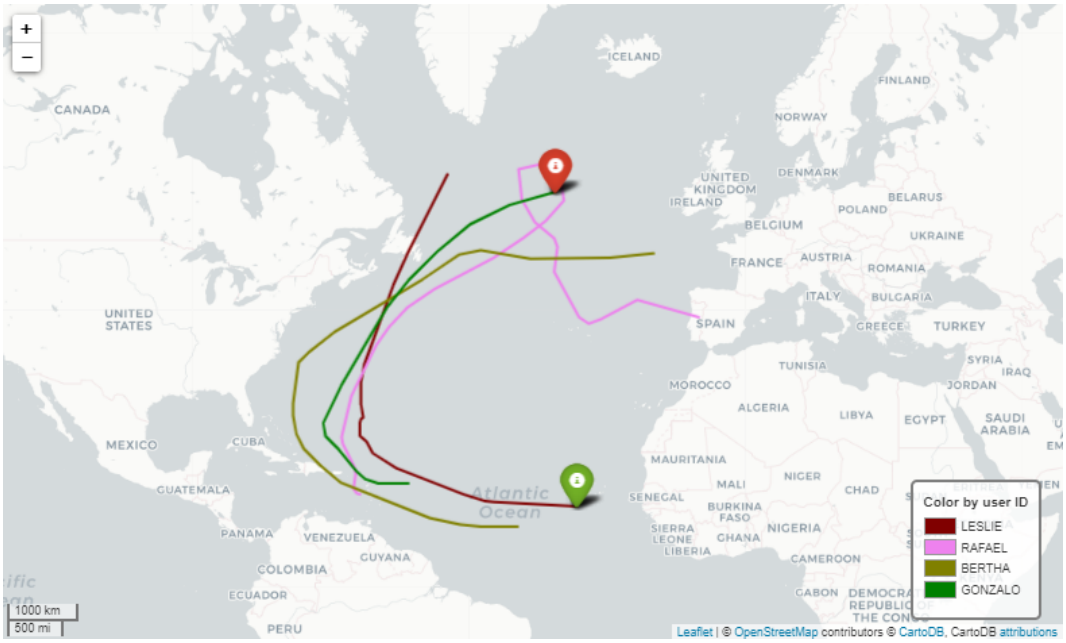
Figura 9 – Trajetória do Furacão Gonzalo

Fonte – O próprio autor

Na Figura 10 é apresentada a visualização das trajetórias recuperadas através da operação *Range*, com valor 200 e medida de similaridade MEDP.

Na Figura 11 é apresentada a visualização das trajetórias recuperadas através da operação *KNN*, sendo $k = 5$, temos as 5 trajetórias mais próximas, espacialmente, do furacão Gonzalo. Nessa operação a medida de distância foi a MEDP

Figura 10 – Operação *Range*, com valor 200 e medida de similaridade MEDP



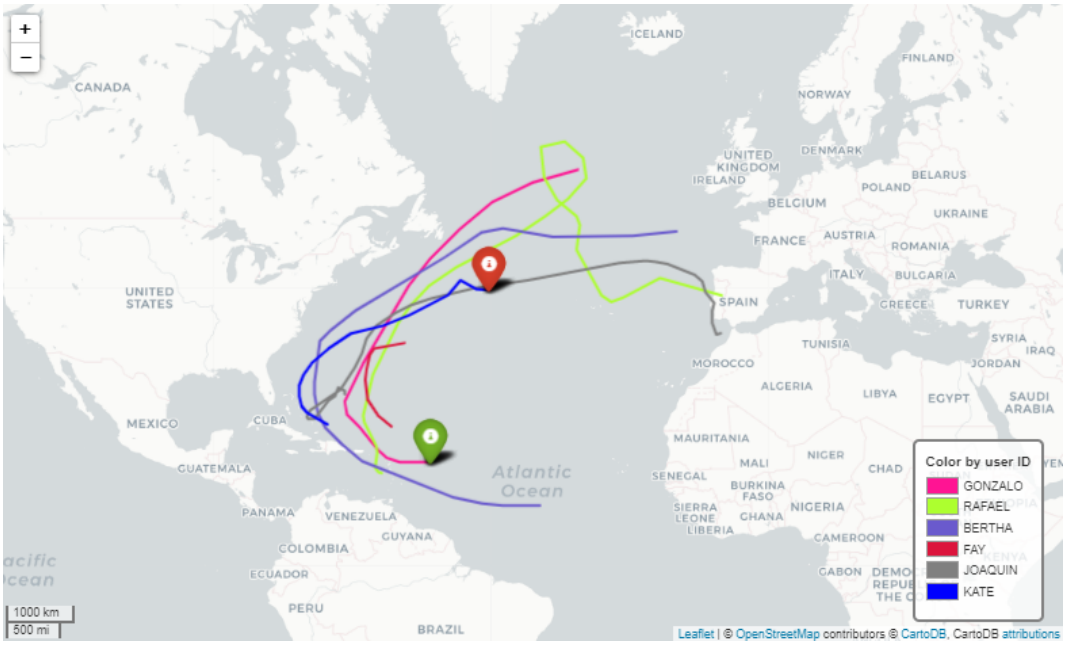
Fonte – O próprio autor

Os testes das medidas usando a operação *Range* retornaram de fato as trajetórias com maior similaridade ao furacão Gonzalo.

O aumento da taxa de *Range* retornou mais trajetórias, sem perder a característica de aparentarem de fato serem as trajetórias mais próximas, espacialmente, ao furacão gonzalo.

As operações kNN também tiveram resultado satifátorio, retornando corretamente as k trajetórias mais próximas ao furacão Gonzalo, espacialmente.

Figura 11 – Operação kNN, com k = 5 e medida de similaridade MEDP.



Fonte – O próprio autor

As operações utilizando MEDT como medida de similaridade também retornaram resultados favoráveis. No caso da operação range com valor 1000, apresentada na Figura 12, foram retornados os furacões Ernesto[5] e Cristobal[6]. Ernesto aparentou uma certa proximidade temporal e diferiu um pouco mais a nivel espacial, enquanto no caso do Cristobal, claramente houve uma grande similaridade a nivel espacial e tambem a nivel temporal, pois ocorreu também no ano de 2014 e realizou uma trajetória similar.

Na Figura 13 temos as trajetórias recuperadas de uma operação KNN, com $k = 5$ e medida de similaridade MEDT.

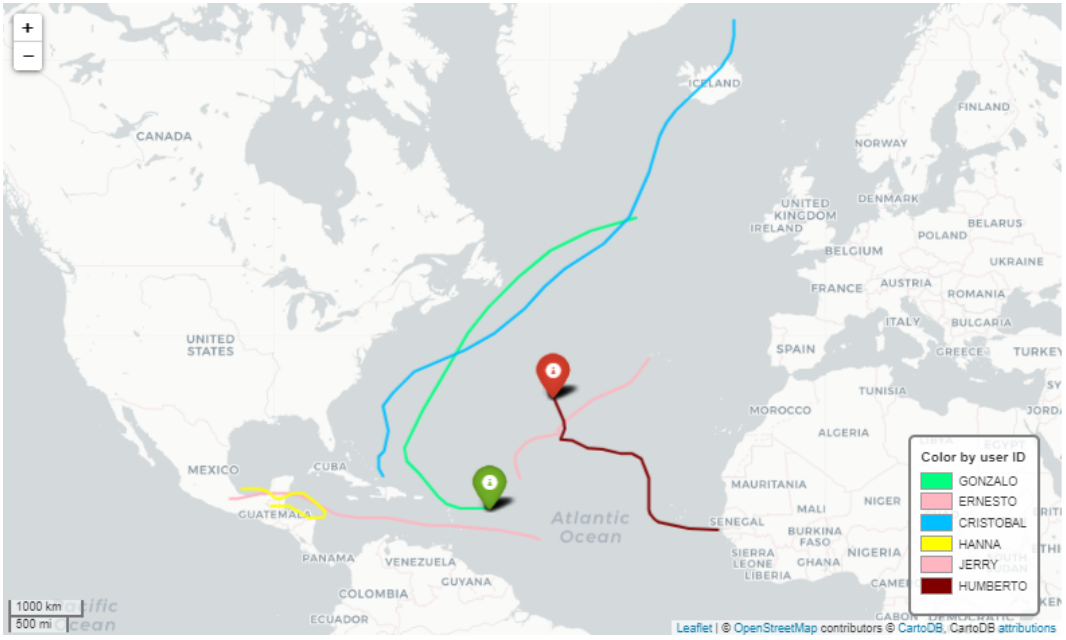
[5] O furacão Ernesto foi um furacão de categoria 2 e um ciclone tropical prejudicial que afetou várias ilhas do Caribe e áreas da América Central em agosto de 2012. Fonte: https://en.wikipedia.org/wiki/Hurricane$_{E}rnesto_{(}$2012)

[6] O furacão Cristobal foi um ciclone tropical moderadamente forte do Atlântico que afetou várias massas de terra de Porto Rico à Islândia no final de agosto e início de setembro de 2014. Fonte: https://en.wikipedia.org/wiki/Hurricane$_{C}ristobal$

Figura 12 – Operação *Range*, com valor 1000 e medida de similaridade MEDT

Fonte – O próprio autor

Figura 13 – Operação kNN, com k = 5 e medida de similaridade MEDT.

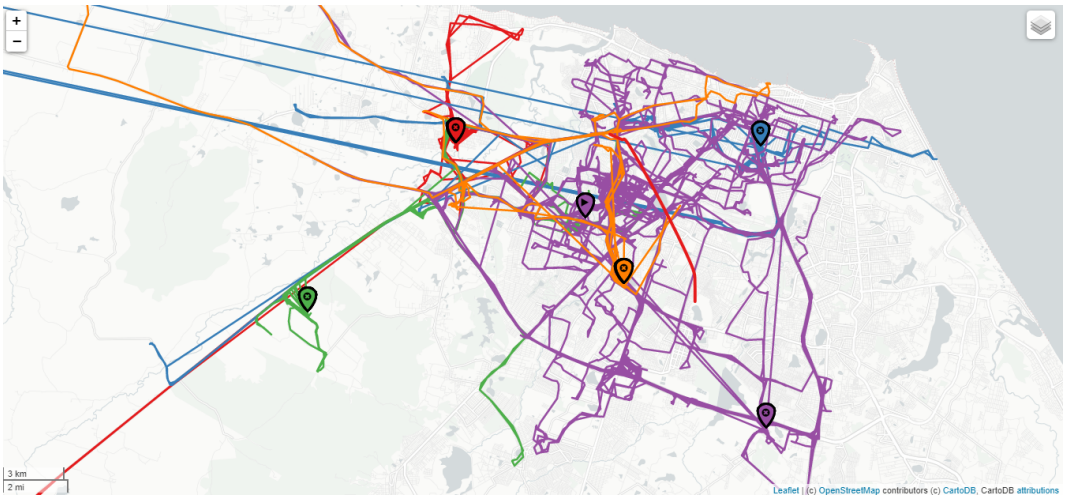


Fonte – O próprio autor

Em seguida, foram realizados testes em trajetórias limitadas a rede de ruas, com alta densidade pelo fato de que as trajetórias foram criadas a partir de pontos gerados a cada 3 segundos. O teste foi uma recuperação das 4 trajetórias mais próximas, através da operação KNN, utilizando como medida de distância a MEDP. Diante da alta complexidade da função,

O($n * m * l$), foi necessário realizar o pré-processamento dos dados, através da compressão das trajetórias e da remoção dos pontos de parada. Ainda assim a função encontrou uma grande demora em sua execução, a tornando um pouco inviável para esse tipo de trajetória. Diante disso, foram geradas *Grids*[7], a partir dos dados das trajetórias, e a similaridade foi analisada a partir das células geradas.

Figura 14 – Recuperação das 4 trajetórias mais próximas à trajetória em vermelho, através da operação KNN, utilizando a medida de distância MEDP.



Fonte – O próprio autor

Foi perceptível, como se pode verificar na Figura 14, que era possivel encontrar alguns casos em que havia similaridade entre trajetórias (A trajetória laranja e a verde cruzam bastante com a vermelha) porém, em outros casos, como na trajetória de cor roxa, não era perceptível uma considerável similaridade dela com a trajetória vermelha. Em muitos casos, pode se desejar buscar apenas por trechos específicos em que se houve similaridade, ou seja, buscar por similaridade parcial, enquanto nesse caso, se procurou por uma similaridade total.

5.2 Avaliação de operações de recuperação baseadas na similaridade parcial

Foi criada uma operação que possui 2 filtros como parâmetros, um espacial, onde se define a distância espacial minima, em metros, para que o ponto da trajetória análisada seja considerado similar, e outro temporal, onde se define a distância temporal mínima para que o ponto da trajetória analisada seja considerado similar.

[7] Divisão do espaço em grades, criando um índice temporal para as trajetórias que caem em cada célula dessa grade. Cada segmento que cai em uma grade é representado por um ponto com as coordenadas iguais ao ponto com horário inicial e o ponto com horário final do segmento. (OLIVEIRA, 2019)

Para esse estudo de caso, dois conjuntos de dados foram utilizados. Um com trajetórias de alta densidade (trajetória de GPS) e outro com trajetórias esparsas, gerada a partir de sensores fixos distribuidos em pontos da cidade. As recuperações tiveram como distância mínima espacial o valor de 300 metros e distância mínima temporal o valor de 3 minutos. Essa distância um tanto grande é devido a baixa continuidade de um dos *datasets*.

Alguns problemas foram encontrados nessa busca, devido a, na maioria dos casos, retornarem apenas pontos de paradas das trajetórias contínuas, o que não era significativo para casos em que se era desejado encontrar a possibilidade das trajetórias pertencerem a objetos que se movimentavam juntos.

Diante disso, um pré-processamento para a remoção dos pontos de paradas foi realizado para esses casos, e, afim de se classificar o nível de similaridade dos pontos definidos como similares em L, se era realizado um ranqueamento dos dados retornados pela operação, gerando um novo *DataFrame* contendo, para cada trajetória retornada, a quantidade de ocorrências (pontos definidos como similares) , a média da distância espacial de todos os pontos, e um *score*, que é definido pela quantidade de ocorrências dividido pela média da distância espacial. Esse *score* serve de peso para classificar-mos o nível de similaridade encontrada para cada trajetória. O novo *DataFrame*, que foi gerado como resultado da consulta, foi ordenado baseando-se no nível do *score*. Foi perceptível, de fato, que as trajetórias com maior score, na maioria dos casos, possuíam maior possibilidade de terem se movimentado em conjunto com a trajetória A.

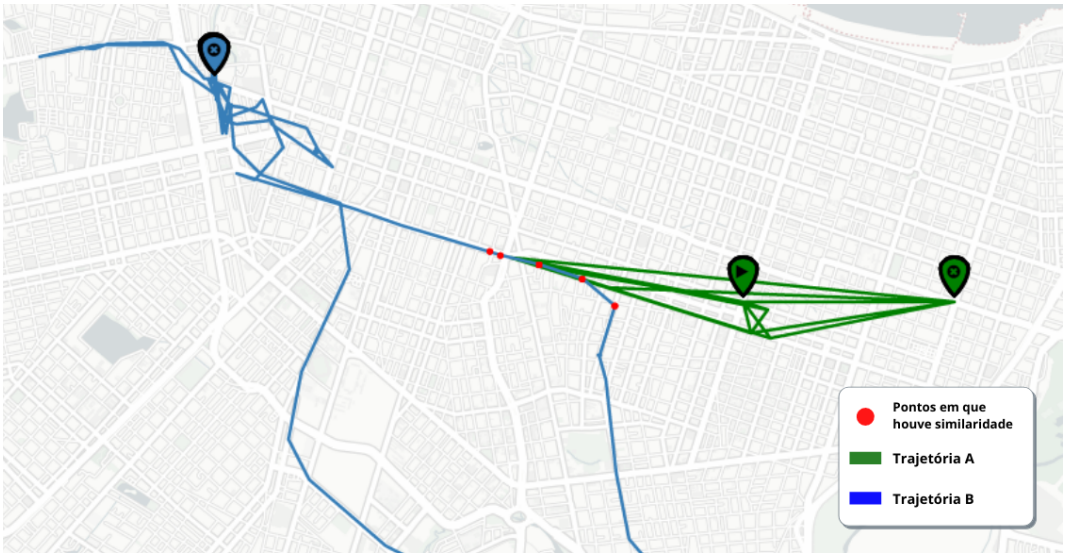
Na Figura 15 temos a visualização de um dos trechos das trajetórias recuperadas em um dos testes. Nesse caso, foram 5 pontos (em vermelho). A trajetória azul apresentou um *score* de 0.238515, o maior entre as trajetórias retornadas pela busca.

Figura 15 – Um dos resultados de busca por similaridade parcial entre trajetórias. A imagem apresenta um dos trechos em que houve similaridade.

Fonte – O próprio autor

Na Figura 16 temos a visualização de outro trecho das trajetórias recuperadas em um dos testes realizados. Nesse caso, foram 5 pontos (em vermelho). A trajetória azul apresentou um *score* de 0.094458, a terceira maior entre as trajetórias retornadas pela busca.

Figura 16 – Um dos resultados de busca por similaridade parcial entre trajetórias. A imagem apresenta um dos trechos em que houve similaridade.

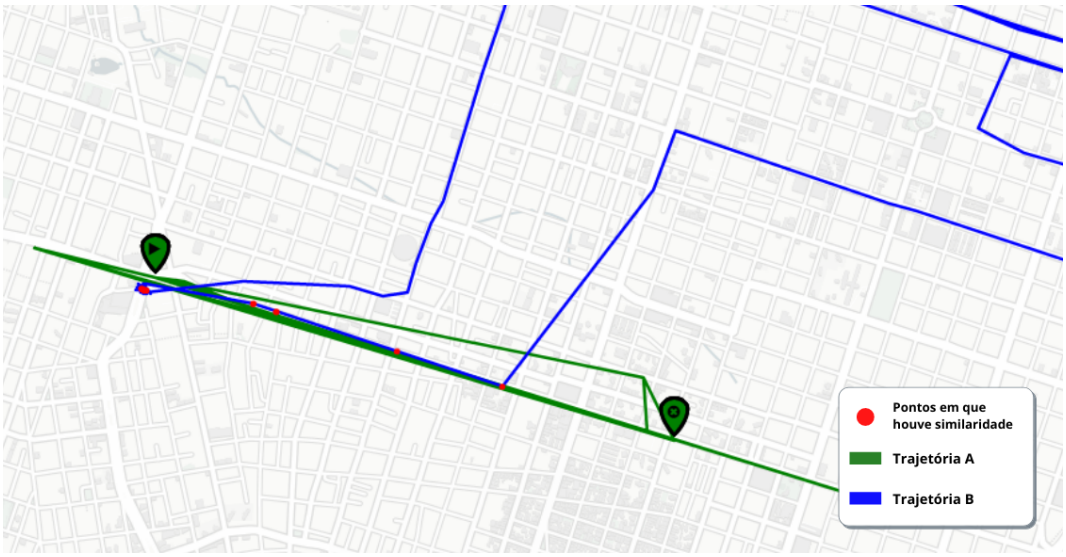


Fonte – O próprio autor

Na Figura 17 temos a visualização de outro trecho das trajetórias recuperadas em um dos testes realizados. Nesse caso, foram 5 pontos (em vermelho). A trajetória azul apresentou

um *score* de 0.238515, o maior entre as trajetórias retornadas pela busca.

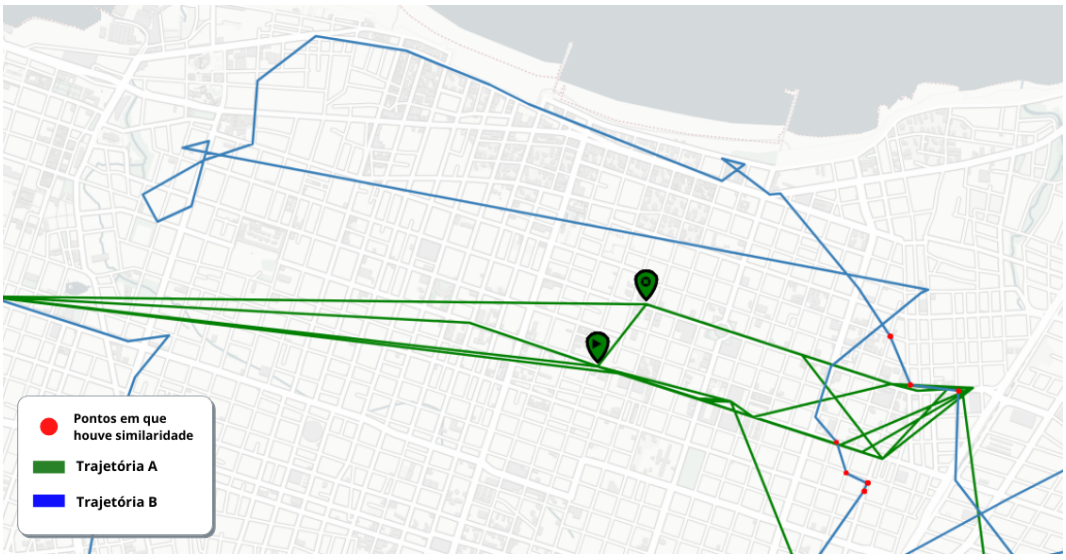
Figura 17 – Um dos resultados de busca por similaridade parcial entre trajetórias. A imagem apresenta um outro trecho em que houve similaridade.



Fonte – O próprio autor

Na Figura 18 temos a visualização de outro trecho das trajetórias recuperadas em um dos testes realizados. A imagem apresenta um outro trecho em que houve similaridade em dois momentos, totalizando 7 pontos (em vermelho). A trajetória azul apresentou um *score* de 0.228674, o segundo maior entre as trajetórias retornadas pela busca.

Figura 18 – Um dos resultados de busca por similaridade parcial entre trajetórias. A imagem apresenta um outro trecho em que houve similaridade em dois momentos.



Fonte – O próprio autor

6 CONCLUSÕES E TRABALHOS FUTUROS

Encontrar similaridade entre trajetórias não foi uma tarefa tão simples. São inúmeras as variáveis nessa busca, e dependendo do tipo de dado, ela pode necessitar de ajustes em seu modo de execução. Existem diversas medidas de similaridade, e é necessário investigar como se comporta a recuperação ao utilizar esses outros tipos, além da possibilidade de se pensar em novas formas de cálcular distâncias. Foi realmente um desafio, como se é previsto por Bian *et al.* (2018), construir operações com uma complexidade que permita o processamento de grandes conjuntos de dados. Se foi também perceptível o quanto o pré-processamento é essencial.

Mesmo diante desses desafios, 3 operações de similaridade foram desenvolvidas: Uma do tipo *Range* para busca por similaridade total, uma do tipo KNN também para busca por similaridade total e uma outra do tipo *range*, porém, para busca por similaridade parcial.

Nas buscas baseadas em similaridade total, não foi definida a métrica para as operações do tipo *range*, o que pode dificultar ao usuário da função a determinar como realizar sua busca. Na medida de similaridade MEDT também houve dificuldade ao se atribuir o quanto a distância espacial e a distância temporal deveriam pesar na consulta. Por isso, na operação baseada em similaridade parcial, houve uma divisão melhor dos atributos de consulta, referentes a distância mínima temporal, e a distância mínima espacial, que foi capaz de contornar esse problema. Um trabalho futuro seria também modificar as funções baseadas em similaridade total, de modo que elas possam carregar parâmetros que facilitem o uso das funções, baseando tais parâmetros em métricas específicas, como distância em metros, distância em segundos, etc.

Apesar das dificuldades, o conjunto de operações desenvolvidas se demonstrou capaz de cumprir uma eficaz recuperação de trajetórias, sejam elas densas, esparsas, presas a rede de ruas ou livres. Um trabalho futuro também seria verificar a escalabilidade dessas operações, quando utilizadas em conjuntos de dados ainda maiores.

Foi também possível fornecer essas operações a qualquer pessoa através do PyMove, facilitando a correta padronização do código e uma continuidade segura, seja por mim, seja por outras pessoas. No mais, esse trabalho traz bases para a construção de operações ainda mais avançadas na busca por similaridade entre trajetórias.

REFERÊNCIAS

BALDUS, J.; BRINGMANN, K. A fast implementation of near neighbors queries for fréchet distance (GIS cup). **CoRR**, abs/1803.00806, 2018. Disponível em: http://arxiv.org/abs/1803.00806 Acesso em: 03 fev. 2021.

BERNARDIN, K.; STIEFELHAGEN, R. Evaluating multiple object tracking performance: The clear mot metrics. **EURASIP Journal on Image and Video Processing**, v. 2008, 01 2008.

BIAN, J.; TIAN, D.; TANG, Y.; TAO, D. A survey on trajectory clustering analysis. **ArXiv**, abs/1802.06971, 2018.

CAI, G.; LEE, K.; LEE, I. A framework for mining semantic-level tourist movement behaviours from geo-tagged photos. In: . [S.l.: s.n.], 2016. v. 9992, p. 519–524. ISBN 978-3-319-50126-0.

CHANDOLA, V.; BANERJEE, A.; KUMAR, V. Anomaly detection: A survey. **ACM Comput. Surv.**, v. 41, 07 2009.

Feng, Z.; Zhu, Y. A survey on trajectory data mining: Techniques and applications. **IEEE Access**, v. 4, p. 2056–2067, 2016.

KONG, X.; LI, M.; MA, K.; TIAN, K.; WANG, M.; XIA, F. Big trajectory data: A survey of applications and services. **IEEE Access**, PP, p. 1–1, 10 2018.

KöRTING, T. S. **How DTW (Dynamic Time Warping) algorithm works.** 2017. Disponível em: https://www.youtube.com/watch?v=_K1OsqCicBY Acesso em: 07 out. 2020.

Li, H.; Liu, J.; Wu, K.; Yang, Z.; Liu, R. W.; Xiong, N. Spatio-temporal vessel trajectory clustering based on data mapping and density. **IEEE Access**, v. 6, p. 58939–58954, 2018.

Magdy, N.; Sakr, M. A.; Mostafa, T.; El-Bahnasy, K. Review on trajectory similarity measures. In: **2015 IEEE Seventh International Conference on Intelligent Computing and Information Systems (ICICIS)**. [S.l.: s.n.], 2015. p. 613–619.

OLIVEIRA, A. **uma arquitetura e implementação do módulo de visualização para biblioteca pymove.** 2019.

Quehl, J.; Hu, H.; Taş, ; Rehder, E.; Lauer, M. How good is my prediction? finding a similarity measure for trajectory prediction evaluation. In: **2017 IEEE 20th International Conference on Intelligent Transportation Systems (ITSC)**. [S.l.: s.n.], 2017. p. 1–6.

SANCHES, A. de J. A. M. **uma arquitetura e implementação do módulo de pré-processamento para biblioteca pymove.** 2019.

SHI, Y.; HUANG, S.; FENG, J.; LU, J. A probabilistic range query of moving objects in road network. **IEEE Access**, v. 7, p. 40165–40174, 01 2019.

SOUZA, C.; PANTOJA, C.; SOUZA, F. C. VerificaÇÃo de assinaturas offline utilizando dynamic time warping. In: . [S.l.: s.n.], 2009. p. 1–5.

TAD: A trajectory clustering algorithm based on spatial-temporal density analysis. **Expert Systems with Applications**, v. 139, p. 112846, 2020. ISSN 0957-4174. Disponível em: http://www.sciencedirect.com/science/article/pii/S0957417419305482 Acesso em: 07 out. 2020.

WANG, S. J.; MORIARTY, P. **Big data for urban sustainability:** a human-centered perspective. 1st. ed. [S.l.]: Springer Publishing Company, Incorporated, 2018. ISBN 3319736086.

XU, J.; LU, H.; GüTING, R. Range queries on multi-attribute trajectories. **IEEE Transactions on Knowledge and Data Engineering**, PP, p. 1–1, 12 2017.

YUAN, G.; SUN, P.; ZHAO, J.; LI, D.; WANG, C. A review of moving object trajectory clustering algorithms. **Artificial Intelligence Review**, v. 47, 03 2016.

ZHENG, Y. Trajectory data mining: An overview. Association for Computing Machinery, New York, NY, USA, v. 6, n. 3, 2015. ISSN 2157-6904. Disponível em: https://doi.org/10.1145/2743025 Acesso em: 06 jun. 2020.

ZHENG, Y.; XIE, X. Learning travel recommendations from user-generated gps traces. **ACM TIST**, v. 2, p. 2, 04 2011.

ZHENG, Y.; ZHANG, L.; MA, Z.; XIE, X.; MA, W.-Y. Recommending friends and locations based on individual location history. **ACM Trans. Web**, Association for Computing Machinery, New York, NY, USA, v. 5, n. 1, fev. 2011. ISSN 1559-1131. Disponível em: https://doi.org/10.1145/1921591.1921596 Acesso em: 27 jun. 2020.