
**UNIVERSIDADE FEDERAL DO CEARÁ**

**CAMPUS QUIXADÁ**

**BACHARELADO EM SISTEMAS DE INFORMAÇÃO**

**YAGO ALVES DA SILVA**

# IDENTIFICAÇÃO DE REGIÕES POPULARES A PARTIR DE TRAJETÓRIAS

**QUIXADÁ – CEARÁ**

**2017**

YAGO ALVES DA SILVA

# IDENTIFICAÇÃO DE REGIÕES POPULARES A PARTIR DE TRAJETÓRIAS

Monografia apresentada no curso de Sistemas de Informação da Universidade Federal do Ceará, como requisito parcial à obtenção do título de bacharel em Sistemas de Informação. Área de concentração: Computação.

Orientadora: Ma. Lívia Almada Cruz Rafael

QUIXADÁ – CEARÁ

2017

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)

---

S584i Silva, Yago Alves da.
Identificação de regiões polares a partir de trajetórias / Yago Alves da Silva. – 2017.
40 f. : il. color.

Trabalho de Conclusão de Curso (graduação) – Universidade Federal do Ceará, Campus de Quixadá, Curso de Sistemas de Informação, Quixadá, 2017.
Orientação: Profa. Ma. Lívia Almada Cruz Rafael.

1. Táxis. 2. Região-Interesse. 3. Análise por agrupamento. 4. Trajetória. I. Título.

CDD 005

---

YAGO ALVES DA SILVA

IDENTIFICAÇÃO DE REGIÕES POPULARES A PARTIR DE TRAJETÓRIAS

Monografia apresentada no curso de Sistemas de Informação da Universidade Federal do Ceará, como requisito parcial à obtenção do título de bacharel em Sistemas de Informação. Área de concentração: Computação.

Aprovada em: __/__/____

BANCA EXAMINADORA

---

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

---

Dra. Ticiana Linhares Coelho da Silva
Universidade Federal do Ceará - UFC

---

Me. Regis Pires Magalhães
Universidade Federal do Ceará - UFC

Aos meus pais, Fátima e Antônio,

a minha avó, Maria.

Ao resto da minha família.

A todos os amigos, em especial a Michele Nascimento.

## AGRADECIMENTOS

À CAPES, pelo apoio financeiro com a manutenção da bolsa de auxílio. A UFC, pela estrutura e apoio financeiro da bolsa do PET. A Prof.Lívia Almada, pela excelente orientação, e aos professores participantes da banca examinadora Regis Pires e Ticiana Linhares pelo tempo, pelas valiosas colaborações e sugestões.

Aos Professores Davi Romero e Lucas Ismaily, pela orientação na bolsa, junto a outros que contribuíram para minha formação acadêmica, como Paulo Henrique, Ricardo Reis, Carlos Igor, Diana Braga, Jefferson Carvalho.

Agradeço aos meus pais, Fátima e Antônio, por terem me dado a educação e apoio necessários para minha formação, e ao meu irmão, Mateus. Também as minhas tias e minha Madrinha Auxilia, pelo apoio.

Agradeço aos meus amigos Matheus Pereira, Danrley, Alex, Guilherme, Alexsandro, Wellington, Junior Leonel, Araújo, Alysson, Daniel, Kerley, Amanda, Alan, Camilla, Emiliana, Paula Ana, Hugo, Neto, Flávio, Rodrigo e todos que conviveram comigo ao longo da minha formação. E a todos que direta ou indiretamente fizeram parte da minha formação.

"Saber muito não lhe torna inteligente. A inteligência se traduz na forma que você recolhe, julga, maneja e, sobretudo, onde e como aplica esta informação"

(Carl Sagan)

## RESUMO

Este trabalho descreve um processo de identificação de regiões populares, por meio de trajetórias de taxistas, onde são identificados os pontos onde o carro está parado. Os dados aqui usados pertencem ao aplicativo TaxiSimples e representam 7 dias de oficio de taxistas, do mês de junho de 2016 na cidade de Fortaleza-CE. Aqui, são avaliadas duas abordagens para chegar a esse objetivo, a primeira estrategia é o SMoT, proposto por Alvares et al. (2007a) e é baseada em interseção entre as trajetórias e pontos de interesse previamente mapeados, esses pontos foram mapeados pela Google e extraídos por meio de um Script que usa uma API disponibilizada por ele. Já a segunda, é o CB-SMoT, proposto por Palma et al. (2008) e usa uma variância do algoritmo DBSCAN para perceber onde o carro está com a velocidade baixa para identificar regiões de interesse. Foi usado também o DBSCAN para a clusterização das paradas, com o objetivo de idenficar clusters de regiões populares, medindo-as pela quantidade de paradas existentes perto dos cores dos clusters. Ao fim disso, o Script foi usado novamente para mostrar os pontos de interesse próximos as regiões populares.



**Palavras-chave:** 1. Táxis. 2. Região- Interesse. 3. Análise por agrupamento. 4. Trajetória. I. Título.

## ABSTRACT

This paper describes a process of identification of popular regions, through taxi drivers trajectories, where the points where the car is stopped are identified. The data used here belong to the TaxiSimples application and represent 7 days of official taxi drivers, from June 2016 in the city of Fortaleza-CE. Here, two approaches are evaluated to reach this objective, the first strategy is the SMoT, proposed by Alvares et al. (2007a) and is based on intersection between previously mapped trajectories and points of interest, these points were mapped by Google and extracted by mean of a script that uses an API made available by it. The second one is the CB-SMoT, proposed by Palma et al. (2008) and uses a variance of the DBSCAN algorithm to figure out where the car is at low speed to identify regions of interest. DBSCAN was also used for the clustering of the stops, with the objective of identifying clusters of popular regions, measuring them by the number of stops existing near the colors of the clusters. At the end of this, the Script was used again to show points of interest near the popular regions.



**Keywords:** 1. Taxis. 2. Region-Interest. 3. Analysis by grouping. 4. Trajectory. I. Title.

## LISTA DE FIGURAS

## LISTA DE TABELAS

SUMÁRIO

# 1 INTRODUÇÃO

Hoje, graças à informatização, é possível ver um grande crescimento no uso de dispositivos *mobile* e *smartphones*. Junto a esses instrumentos, tornou-se comum o uso de ferramentas de localização como o *Global Position System* (GPS). O GPS foi criado para fins militares e é um aparelho que recebe sinais enviados por satélite e determina onde a pessoa ou aquele lugar que ela busca está (PARKINSON; ENGE, 1996). O uso do GPS gerou, e ainda gera, uma quantidade enorme de dados que podem ser explorados em diversas áreas, como por exemplo, gestão de tráfego, migração de animais, comportamentos humanos em lugares públicos, etc.

Em geral, os dados disponibilizados pelos aparelhos de GPS contêm como atributos comuns um identificador, latitude, longitude e a data, que é composta também pela hora em que o dado foi coletado. Quando agrupados de maneira cronológica e pertencentes ao mesmo veículo, esses dados são chamados de trajetória. Entretanto, extrair informações de dados de trajetória não é uma tarefa fácil, devido ao grande volume de dados e da complexidade dos mesmos (ALVARES et al., 2007b). Na maioria dos casos também se faz necessária uma etapa de pré-processamento para transformação dos dados em um formato adequado, removendo inconsistências e campos que não têm valor semântico. Além disso, os dados de GPS podem não se referir a uma única viagem, mesmo quando coletados em um único dia.

A análise dos dados de trajetórias é útil para descoberta de padrões, tais como: padrões de locomoção através da descoberta de matrizes Origem/Destino e identificação de regiões mais visitadas, (GIANNOTTI et al., 2009). O resultado dessas análises, no caso específico das regiões mais visitadas é uma informação valiosa para a recomendação, uma área bastante estudada atualmente. Esses resultados podem ser usados para recomendar pontos turísticos populares, como também outros pontos de interesse populares como lojas, restaurantes, *shoppings*, etc. Pode-se levar em consideração por exemplo a cidade de Fortaleza, que é uma das mais visitadas do Brasil por parte da comunidade nacional e internacional. Segundo o portal de notícias G1[1], foram registrados 100 mil desembarques internacionais em 2016. Os principais países de origem são Argentina, Chile, Paraguai, Uruguai, França, Alemanha, Itália, Inglaterra, Portugal e Espanha. Neste trabalho, os dados usados são de trajetórias que representam o percurso de taxistas em ofício na cidade de Fortaleza e o seu objetivo é identificar regiões e/ou

[1] http://g1.globo.com/ceara/noticia/2017/01/fortaleza-bate-recorde-de-turistas-estrangeiros-em-2016-diz-setur.html

pontos de interesse populares, ou seja, os locais mais visitados. Esse tipo de informação pode ser usada por turistas que estão de passagem pela cidade ou até mesmo pessoas que residem em Fortaleza e buscam por opções de lazer.

O primeiro passo aqui, consiste em separar as regiões de interesse mapeadas em sites públicos e os dados das trajetórias. No segundo, uma etapa de pré-processamento deve ser feita identificando dentro dessas trajetórias, onde o carro estava em movimento e o momento em que ele parou. O desafio nessa parte, é identificar onde realmente houve uma parada para embarque ou desembarque de passageiros, visto que essa pausa pode ter sido, por exemplo, em um semáforo ou engarrafamento. Essas paradas, quando presentes em grande número nas trajetórias, dias e horários diferentes, podem significar a descoberta de pontos de interesse frequentemente visitados.

## 1.1 OBJETIVOS

Este seção tem com intuito expor os objetivos gerais e específicos do presente trabalho.

### *1.1.1 Objetivo Geral*

Identificar lugares populares a partir das trajetórias de taxistas.

### *1.1.2 Objetivos específicos*

- Identificar os pontos de interesse;
- Transformar as trajetórias simples em semânticas;
- Comparar o comportamento dos algoritmos usados para transformação das trajetórias;
- Identificar os lugares populares com base nos pontos de parada.

O restante deste trabalho está organizado da seguinte forma: No Capítulo 2, é feito uma explicação do problema e de alguns conceitos que não são de conhecimento comum e de como eles se fazem necessários para o presente trabalho. O Capítulo 3 mostra um resumo dos conceitos teóricos. No Capítulo 4 são descritos alguns trabalhos relacionados e de como eles se assemelham e se diferenciam deste. O Capítulo 5 mostra os procedimentos metodológicos e detalha as atividades feitas. O Capítulo 6 mostra os resultados alcançados no trabalho. Por fim, o Capítulo 7 descreve as conclusões do trabalho.

# 2 PROBLEMÁTICA

Este Capítulo tem como objetivo mostrar o problema e os conceitos necessários para o bom entendimento deste trabalho. Abaixo, estão as definições formais dos princípios.

**Definição 2.0.1 (Ponto)** *Um* ***ponto*** $p = (id, timestamp, latitude, longitude)$ *representa um dado de rastreamento de GPS, onde:* ***id*** *é o identificador do objeto em movimento;* ***timestamp*** *é a data e hora exata de quando foi reportado aquele dado;* ***latitude*** *é o ângulo entre o plano do equador e a superfície de referência; e* ***longitude*** *é a medida ao longo do Equador, e representa a distância entre um ponto e o Meridiano de Greenwich.*

**Definição 2.0.2 (Trajetória simples)** *Uma* ***trajetória simples*** *é uma sequência de n pontos* $\{p_1 \rightarrow \cdots \rightarrow p_n\}$ *ordenada de maneira cronológica de acordo com o atributo timestamp e que possuem o identificador em comum.*

**Definição 2.0.3 (Parada)** *Uma* ***parada*** *é uma tupla (RC,* $\Delta C$*), em que RC é um polígono e* $\Delta C$ *é um número real estritamente positivo. O conjunto RC é chamado geometria da parada e representa em forma de polígono o a área onde ocorreu a parada.* $\Delta C$ *é chamada de sua duração mínima de tempo (ALVARES et al., 2007b).*

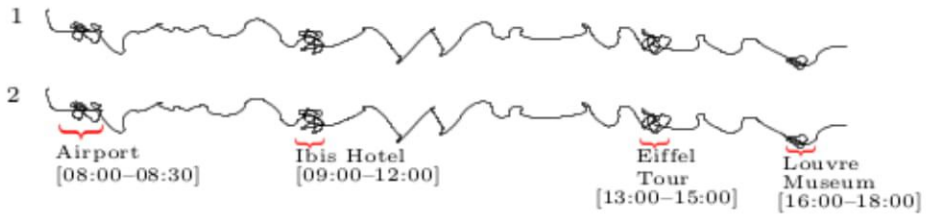
**Definição 2.0.4 (Movimento)** *Um* ***movimento*** *em uma trajetória T é : (i) uma sub-trajetória de T delimitada por duas paradas temporariamente consecutivas de T; ou (ii) a sub-trajetória de T entre o ponto de partida de T e a primeira parada de T; ou (iii) a subtração de T entre a última parada de T e o ponto final de T; ou (iv) a trajetória T em si, se T não tem paradas (ALVARES et al., 2007b).*

**Definição 2.0.5 (Trajetória semântica)** *Uma* ***trajetória semântica*** *é uma trajetória simples que possui uma integração com objetos geográficos (ALVARES et al., 2007a).*

A Figura 1 exemplifica a adição da semântica, onde a trajetória 1 representa uma trajetória simples como um aglomerado de pontos e trajetória 2 representa a integração com objetos geográficos por meio de regiões de interesse: *Airport*, *Ibis* Hotel, etc.

**Definição 2.0.6 (Região de interesse)** *Uma* ***região de interesse*** *é um objeto geográfico que é interessante para uma aplicação específica, geralmente associada a uma atividade humana (BRILHANTE et al., 2012).*

Figura 1 – A Trajetória 1 é uma trajetória bruta e a 2 é uma trajetória semântica.




Fonte : (ALVARES et al., 2007a)

O problema de que trata este trabalho, consiste em identificar regiões de interesse populares, a partir de um conjunto de trajetórias. Para alcançar este objetivo, as trajetórias simples precisam ser transformadas em trajetórias semânticas, para que se possa verificar que regiões de interesse possuem interseções com os pontos de parada.

# 3 FUNDAMENTAÇÃO TEÓRICA

Neste Capítulo, serão abordados os principais conceitos que baseiam esse trabalho e como cada um vai influenciar no desenvolvimento do mesmo.

## 3.1 Mineração de dados

Tendo em vista o aumento avassalador da taxa de geração de dados nos vários campos do conhecimento humano, é fácil ver que também há, consequentemente, uma crescente necessidade de extração de informações úteis a partir destes dados, tanto do ponto de vista científico quanto do ponto de vista prático. (BRAMER, 2007) exemplifica essa eclosão com aplicações atuais, como os satélites de observação da NASA, que diariamente geram cerca de um *terabyte* de dados, ou mesmo instituições que guardam repositórios com milhares de transações dos seus clientes. É normal ver situações onde esses dados estão apenas sendo armazenados diariamente sem ser processados, (**??**) refere-se a essa situação como "rico em dados, pobre em informação".

A Mineração de dados tem sido uma grande aliada da ciência nos últimos anos, como ferramenta de descoberta automática de padrões, mudanças, associações, sequências e anomalias nessas grandes massas de dados. Por ser interdisciplinar, possui várias definições. No presente trabalho, são apresentadas suas principais áreas de atuação, sendo elas: Estatística, Aprendizagem de Máquina e Banco de Dados.

- (HAND; MANNILA; SMYTH, 2001), no ponto de vista estatístico : "Mineração de Dados é a análise de grandes conjuntos de dados a fim de encontrar relacionamentos inesperados e de resumir os dados de uma forma que eles sejam tanto úteis quanto compreensíveis ao dono dos dados".
- (CABENA et al., 1998), no caso de banco de dados : "Mineração de Dados é um campo interdisciplinar que junta técnicas de máquinas de conhecimentos, reconhecimento de padrões, estatísticas, banco de dados e visualização, para conseguir extrair informações de grandes bases de dados".
- (FAYYAD; PIATETSKY-SHAPIRO; SMYTH, 1996), quando a ênfase está aprendizagem de máquina : "Mineração de Dados é um passo no processo de Descoberta de Conhecimento que consiste na realização da análise dos dados e na aplicação de algoritmos de descoberta que, sob certas limitações computacionais, produzem um

conjunto de padrões de certos dados".

Em termos de dimensões técnicas, os principais conceitos e algoritmos desenvolvidos na mineração de dados se enquadram nas seguintes categorias: associações, classificações, predições numéricas e clusterização.

Associação é uma das técnicas mais conhecidas de mineração de dados, devido ao seu forte uso em áreas empresarias, um exemplo disso é o problema da análise da Cesta de Compras, que consiste em identificar o relacionamento dos itens mais frequentes em um determinado conjunto de dados, e permite obter resultados do tipo: SE compra leite e pão TAMBÉM compra manteiga, (MANCHANDA; ANSARI; GUPTA, 1999).

As técnicas de classificação são usadas para prever valores de variáveis do tipo categóricas. Pode-se, por exemplo, criar um modelo que classifica os clientes de um comércio como especiais e customizar produtos e/ou promoções em busca de maior satisfação do mesmo, o que influencia na sua volta e na indicação para possíveis novos usuários (CHEN; CHUANG, 2008).

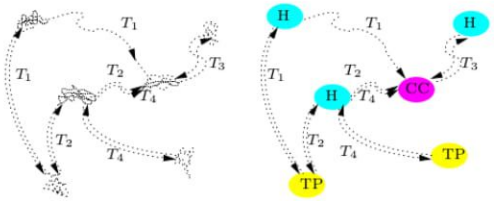
Os métodos de predição visam descobrir um possível valor futuro de uma variável. As predições numéricas visam prever valores para variáveis contínuas. As técnicas de regressão modelam, por exemplo, a previsão de falência adequados para condições econômicas normais e de crise (SUNG; CHANG; LEE, 1999).

As técnicas de clusterização funcionam da seguinte forma: dado um conjunto de registros, são gerados agrupamentos, contendo os registros mais semelhantes. Em geral, as medidas de similaridade usadas são as medidas de distâncias tradicionais, como a distância Euclidiana. Os elementos de um cluster são considerados similares aos elementos no mesmo cluster e dissimilares aos elementos nos outros clusters (ESTER et al., 1996).

A mineração de dados em trajetória pode ser considerada uma das extensões da mineração de dados comum, aprimorando suas técnicas de acordo com o problema a ser resolvido. Segundo (BRAKATSOULAS; PFOSER; TRYFONA, 2004), a análise dos dados de trajetória consiste na integração de dados espaciais, dados não-espaciais e de trajetória. A integração de dados de trajetória com a informação geográfica de base, pode levar à descoberta de padrões de trajetória semântica que muitas técnicas de mineração de dados que consideram trajetórias como amostra de pontos, podem não ser capazes de descobrir. Um exemplo pode ser visto na Figura 2. No lado esquerdo, um conjunto de trajetórias é representado na forma de pontos de amostragem, sem semântica. No lado direito, a informação geográfica é integrada às trajetórias. Levando

em consideração o exemplo fictício de uma pessoa que vai a uma cidade para um congresso, podemos dizer que os pontos nas trajetórias representam um hotel (H-Hotel), um local turístico (*TP-Turistic place*) e um centro de convenções ($CC - ConventionCenter$).

Figura 2 – Trajetória sem semântica e trajetória com semântica.




Fonte : (ALVARES et al., 2007a)

Essa abordagem com semânticas será usada no presente trabalho, com o objetivo de dar semântica às trajetórias simples dos taxistas. Isso pode ser feito a partir da integração do banco de dados de trajetórias com os dados de região de interesse.

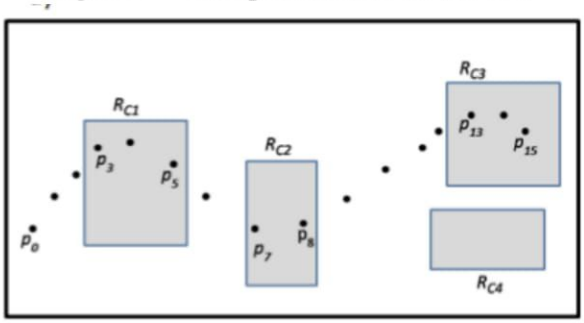
## 3.2 Algoritmos para extração de Paradas e Movimentos

Alvares et al. (2007a) propõem um algoritmo denominado SMoT (*Stops and Moves of Trajectories*), de pré-processamento para integrar trajetórias a uma semântica, a fim de identificar os seus pontos de paradas. A motivação para o seu desenvolvimento consiste no fato de que a maioria dos algoritmos de mineração de dados baseados apenas em geometria não consegue identificar todos os pontos, já que a maioria dos pontos turísticos e hotéis não ficam em locais de densidade grande de pontos, e sim, mais afastados. O problema aqui é que se faz necessário também a criação de um banco de dados com os pontos de interesse em uma determinada cidade.

O algoritmo consiste em uma busca dentro da trajetória, onde os pontos dela se interceptam com uma região de interesse, que é uma das entradas do algoritmo, junto ao banco de dados de trajetórias, um período mínimo de tempo e um número em metros que vai definir o tamanho do $buffer$ criado ao redor da região de interesse. Sendo assim, a intercessão deve durar um período de tempo para ser considerada uma parada. A Figura 3 exemplifica o funcionamento do algoritmo SMoT, onde os $R_c$ representam as regiões de interesse e os $P_s$ os pontos da trajetória.

Os resultados obtidos por eles mostram que essa técnica pode ser usada para descobrir quais pontos de interesse são mais visitados em determinado local, por exemplo. Mas se limita ao conjunto de dados passado como parâmetro, ou seja, só serão encontrados resultados nos

Figura 3 – Comportamento do SMoT.




Fonte : (ALVARES et al., 2010)

lugares de interesse passados como parâmetro, o que dificulta a descoberta de novos pontos visitados, que não se tinha conhecimento prévio.

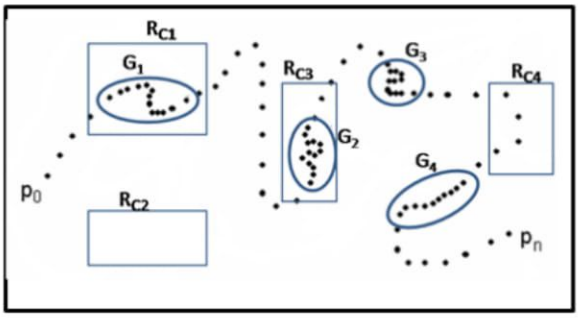
Outra opção usada para identificar as paradas é o CB-SMoT (*Clustering-Based SMoT*). Palma et al. (2008) propõe um algoritmo de duas etapas para extrai-las, onde na primeira etapa, as partes mais lentas de uma trajetória, são identificadas usando a variação do algoritmo DBSCAN. Dentro da segunda etapa, o algoritmo testa se a parte mais lenta teve uma duração maior do que o tempo mínimo, passado por parâmetro no inicio de sua execução. Em caso afirmativo, uma parada é identificada e persistida no banco de dados. O CB-SMoT recebe três parâmetros de entrada, assim como o SMoT, recebe um conjunto de trajetórias e um tempo mínimo, que decide o quanto uma parada deve durar para ser registrada. O terceiro parâmetro é o *eps*, que indica a distância absoluta utilizada para calcular a vizinhança de um ponto. Como é difícil para um usuário comum especificar um bom valor para ele, Palma et al. (2008) usa um parâmetro relativo, relacionado a média e o desvio padrão, em vez do usuário valor *eps* absoluto definido. Esse parâmetro é composto por dois números entre 0 e 1 e a partir dele, o eps é calculado. A Figura 4 exemplifica o comportamento do CB-SMoT, onde os $G_s$ representam a parte da trajetória onde o carro estava com a velocidade baixa e os $R_s$ a representam as regiões de interesse desconhecidas que devem ser encontradas.

## 3.3 DBSCAN

Ester et al. (1996) propõe um método de clusterização baseado em densidade, que é significativamente efetivo para identificar clusters de formato arbitrário e de diferentes tamanhos, identificar e separar os ruídos dos dados e detectar clusters “naturais” e seus arranjos dentro do espaço de dados, sem qualquer informação preliminar sobre os grupos.

Ester et al. (1996) escreve que a noção de *clusters* e o algoritmo DBSCAN se aplicam

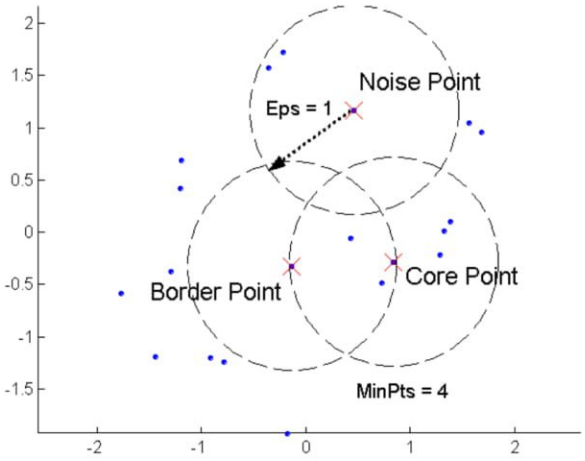
Figura 4 – Comportamento do CB-SMoT.




Fonte : (ALVARES et al., 2010)

para espaços Euclidianos de duas e três dimensões, como para qualquer espaço característico de alta dimensão. O método DBSCAN é aplicável a qualquer base de dados contendo dados de um espaço métrico. Os autores salientam ainda que a abordagem trabalha com qualquer função de distância, de maneira que uma função apropriada pode ser escolhida para alguma dada aplicação. A densidade se dá pela quantidade de pontos dentro de um raio (Eps). Esse número, junto a quantidade mínima de pontos (MinPts) que precisam estar dentro do cluster são as entradas para a implementação desse algoritmo. Um ponto é um *corepoint* se tem mais que um especificado número de pontos (MinPts) dentro do Eps. Caso ele tenha menos que MinPts dentro de Eps, mas esteja na vizinhança de um *corepoint*, é chamado de *borderpoint*. Por fim, um ponto *noisepoint* é qualquer ponto que não é *corepoint* nem *borderpoint*. A Figura 5 mostra graficamente o que foi explicado anteriormente.

Figura 5 – DBSCAN.




Fonte : Introduction to Data Mining by Tan, Steinbach, Kumar

A medida de similaridade escolhida para essa etapa do trabalho foi a Distância Euclidiana, que dá como resultado a menor distância entre dois pontos. No presente trabalho,

os pontos usados serão a latitude e longitude das Paradas obtidas por meio dos algoritmos SMoT e CB-SMoT. Após a aplicação desses algoritmos, teremos os lugares populares, que serão representados pelos *clusters* obtidos como saída do DBSCAN.

# 4 TRABALHOS RELACIONADOS

O trabalho aqui proposto se baseia em conceitos, informações, técnicas e experiências vistas em artigos e trabalhos científicos selecionados na área de mineração de dados e análise de trajetórias semânticas.

Apesar dos trabalhos se encontrarem na mesma área e possuírem objetivos parecidos, apresentam algumas diferenças em relação ao presente trabalho, as quais se referem a adaptações de algoritmos para um propósito intrínseco.

## 4.1 Mining Time-dependent Attractive Areas and Movement Patterns from Taxi Trajectory Data

(YUE et al., 2009) utiliza dados de trajetórias de táxi para explorar locais de interesse e padrões de movimento, que podem ser representados por tráfego elevado em áreas de demanda e movimento de passageiros entre eles, ou seja, o nível de atratividade do local se dá pelo número de visitas recebidas.

Nesse trabalho, a coleta de dados para a análise foi feita com 480 táxis em um domingo, das 8:00 as 24:00, na cidade Chinesa de Wuhan. A estratégia usada por (YUE et al., 2009) se baseia na Clusterização dos pontos de $Pick-up$ e $Drop-off$, que equivalem respectivamente a entrada dos passageiros no ponto de origem e a saída dos mesmos no destino. Antes da aplicação do algoritmo foi feito uma partição dos dados, separando-os em 5 grupos, onde cada um significa um período do dia. O grupo 1 contina os dados das 8:00 as 11:00, o grupo 2 das 11:00 as 13:00, o grupo 3 das 13:00 as 16:30, o grupo 4 das 16:30 as 20:00, e por fim, o grupo 5 das 20:00 as 24:00. Após a análise dos $Clusters$, os resultados encontrados foram padrões de movimentos como, por exemplo, que por tradição comercial, o centro da cidade, atrai uma grande porcentagem de viagens. Sendo assim, as pessoas que moram no centro tendem a se movimentar dentro dele. Já em áreas mais afastadas da cidade, que é separada por um rio, preferem não o atravessar, por causa da distância e dos congestionamentos que acontecem frequentemente nas pontes.

A diferença deste artigo para o presente trabalho, é que o seu foco se mantém em encontrar origens e destinos populares por meio de uma matriz de origem e destino. Já a semelhança está em usar dados de taxistas e da utilização de técnicas de clusterização.

## 4.2 Mining Mobility Behavior from Trajectory Data

Nesse trabalho, (GIANNOTTI et al., 2009) e o grupo The GeoPKDD Project, propõem uma abordagem para extrair padrões de movimentos em trajetórias de automóveis. Um exemplo de resultado esperado é identificar pessoas que têm um mesmo padrão de deslocamento de casa para o trabalho.

O primeiro passo nesse caso foi coletar os dados usando aparelhos de GPS, na cidade de Milão, que fica na Itália. Após a aplicação dos métodos de mineração e análise dos dados, foram encontrados resultados que foram validados com a ajuda da Agência de Mobilidade de Milão. Dentre os resultados obtidos, estavam uma matriz de origem e destino mais precisa do que a usada pela Agência anteriormente, padrões de movimento focados na ida e na vinda de pessoas ao trabalho e os principais itinerários para um destino, com, por exemplo, estacionamentos. A abordagem usada aqui baseia-se em clusterização.

## 4.3 Spatiotemporal Structure of Taxi Services in Shanghai:Using Exploratory Spatial Data Analysis

Nesse trabalho, (DENG; JI, 2011) procura analisar a estrutura espaço temporal de serviços de táxi na cidade de Xangai, que fica na China, através de um grupo de métodos conhecidos como análise exploratória de dados espaciais (ESDA), que detectam se existe um padrão espacial e como descrever ou modelar o mesmo. Seu objetivo é encontrar os locais onde o uso de Táxi é extremamente solicitado.

No experimento, cerca de 27.000 táxis foram equipados com receptores GPS, o que gerou 18.976 dados de trajetória coletados em um dia de trabalho, cobrindo toda a área da cidade. Os resultados obtidos foram representados em mapas, onde cores separam as áreas onde o serviço foi mais utilizado. Um exemplo de resultado é que onde as atividades empresariais, sociais e culturais convergiam dia e noite, as exigências do uso do serviço foram de proporção superior as das zonas de subúrbio, devido à falta de atividades de grande escala e sustentáveis.

Apesar de buscar resultados parecidos, a diferença entre este e o presente trabalho está na forma de chegar aos resultados. Apesar de usar clusterização, nesse trabalho a medida de similaridade usada é diferente da Euclidiana.

O Quadro 1 mostra uma comparação entre as características dos trabalhos relacionados :

Quadro 1 - Quadro de comparação

| | Região de interesse | Dados de taxista | Matriz O/D | Algoritmo |
|---|---|---|---|---|
| (Yue el al., 2009) | Não | Sim | Sim | CB-SMoT |
| (GIANNOTTI et al.,2009) | Não | Sim | Sim | T-Patterns |
| (DENG;JI,2001) | Não | Sim | Não | ESDA |
| Presente Trabalho | Sim | Sim | Não | SMoT, CB-SMoT, DBSCAN |

Fonte : Elaborada pelo autor

# 5 PROCEDIMENTOS METODOLÓGICOS

Este capítulo tem o intuito de mostrar todos os passos que foram seguidos para a realização deste trabalho.

## 5.1 Processamento dos dados

O primeiro passo deste trabalho consiste em alimentar os bancos de dados que foram usados como entradas no SMoT. O parâmetro inicial são as trajetórias dos taxistas, disponibilizadas pela empresa TaxiSimples[1]. Vale ressaltar que no presente trabalho foram usadas apenas trajetórias de taxistas.

O segundo passo foi criar o banco de dados de pontos de interesse presentes em sites públicos. Esse é o segundo parâmetro do algoritmo SMoT, que foi usado para intersecção com o primeiro para identificar dentro das trajetórias, a parada que houve naquele local. Essas informações foram extraídas da API Google Places [2] por meio de um *script*[3] desenvolvido na linguagem JavaScript. Dentro desse *script*, foi necessária a criação de uma chave pedida pela API, que pode ser registrada no Gerenciador de APIs [4] disponibilizado pela Google. O *script* funciona a partir de alguns argumentos de entrada que devem ser passados dentro do código, sendo eles: um ponto, composto por latitude e longitude; um raio, que representa o tamanho do raio que fica em volta do ponto; e por último, um tipo que vai decidir quais pontos de interesse vão ser coletados e baixados em forma de um arquivo em formato JSON. A API permite a pesquisar vários tipos de locais, por exemplo, caso seja passado como argumento o nome Café, será retornado um JSON com as informações de estabelecimentos que foram mapeados com essa característica. No caso do presente trabalho, separamos doze tipos que jugamos úteis, onde os mesmos são locais que atendem às seguintes categorias: Aeroportos, Café, Clubes noturnos, Estádios, Lojas, Lojas de eletrônicos, Museu, Parques, Parques de diversão, Pontos de Táxi, Restaurante e Shoppings.

Após o *download* dos dados citados anteriormente, um pré-processamento foi feito neles para a extração de campos que não possuem valor semântico e, por isso, não são importantes para a análise. Para essa atividade foram usandos a ferramenta Pentaho [5] e um

[1] http://taxisimples.com.br/
[2] https://developers.google.com/places/javascript/?hl=pt-br
[3] https://github.com/YagoAlves/Script—Locais-de-interesse-
[4] https://console.developers.google.com/apis/library?project=testeapimaps-168522hl=pt-br
[5] http://www.pentaho.com/

*Parser*[6] implementado pelo autor, com o objetivo de transformar os arquivos que foram baixados em tabelas do banco. A Tabela 1 mostra a como as categorias foram persistidas no banco, usando os dados obtidos nos arquivos JSON, sendo uma tabela para cada categoria. A Tabela 2 exemplifica os atributos da trajetória salvos no banco.

Tabela 1 – Estrutura das tabelas de locais de interesse

| Atributo | Significado |
|---|---|
| Gid | Identificador do local de interesse |
| Nome | Nome do local de interesse |
| Latitude | Latitude referente ao local de interesse |
| Longitude | Longitude referente ao local de interesse |
| Geometria | O ponto formado pela junção da latitude e da longitude |
| Avaliação | A avaliação atribuida pelo Google para esse local (1 a 5) |

Fonte : Elaborada pelo autor

Tabela 2 – Estrutura da tabela de trajetórias

| Atributo | Significado |
|---|---|
| Tid | Identificador da trajetória |
| Gid | Identificador da parada (caso seja identificada) |
| Latitude | Latitude referente ao local do ponto |
| Longitude | Longitude referente ao local do ponto |
| Geometria | O ponto formado pela junção da latitude e da longitude |
| Tempo | O timestamp do ponto |

Fonte : Elaborada pelo autor

## 5.2 Adição de semântica às trajetórias

O próximo passo do trabalho, consiste em executar o algoritmo SMoT e CB-SMoT. Esses algoritmos estão presentes na ferramenta Weka-STMP [7]. A primeira parte dessa etapa consistiu em diminuir o tamanho do banco de dados de trajetórias, pois o mesmo se encontrava com o mês inteiro de Junho de 2016, o que estava dificultando a execução por possuir muitos pontos. Uma tabela alternativa foi criada, onde foi armazenada nela apenas a primeira semana do mês. Vale ressaltar que durante todo o trabalho foi usada apenas uma semana. Após isso, foi necessária a configuração do arquivo config.properties dentro da pasta do Weka, para que o mesmo pudesse ter acesso ao banco de dados, e com a execução da classe TrajectoryFrame.java, presente no pacote weka.gui.stpm, foi possível realizar a execução dos algoritmos.

[6] https://github.com/YagoAlves/TCC/tree/master/ParsingJson
[7] https://github.com/yipeng/WEKA-STPM

Para a identificação de regiões populares utilizando o SMoT, foi realizada apenas a escolha dos parâmetros e a execução do algoritmo. Como o SMoT já tem como parâmetro os pontos de interesse, esses foram ordenados de acordo com a quantidade de paradas encontradas, sendo os mais populares aqueles com uma maior quantidade de paradas. Porém, foi identificada a necessidade de um pós-processamento do resultado do CB-SMoT.

## 5.3 Clusterização das paradas do CB-SMoT

Após esse período de testes, foi percebido que o CB-SMoT encontrou uma quantidade de Paradas muito maior do que as do SMoT, identificando regiões de interesse maiores e que juntas, cobriam praticamente toda a cidade, como mostra a Figura 6. Isso já era esperado visto que o CB-SMoT identifica também regiões desconhecidas e como ele se baseia na velocidade do carro, presume-se que seus resultados envolvem não só parada para embarque e desembarque, mas também zonas de engarrafamentos. Logo, sabendo que essas áreas maiores influenciariam no resultado da clusterização, foi utilizada a função utmzone [8] para a criação de um filtro, que encontra áreas com o mesmo tamanho que o SMoT cria a partir de uma região de interesse, ou seja, depois do filtro restaram apenas as regiões com até 50 metros de tamanho. As demais funções já implementadas pelo Postgis [9] são todas baseadas em graus, o que dificultou a medição do tamanho das áreas, porém, a utmzone recebe o centroide de uma região em graus e retorna em metros, o que foi um fator de suma importância para sua escolha.

Após esse processamento, foi obtido o centro das áreas por meio da função STCentroid [10], para que fosse possível usar o DBSCAN disponibilizado pelo *Scikit - learn* [11]. Também foi testada a clusterização com a área toda, por meio de uma matriz de distancias, mas devido ao grande número de dados gerados, não foi possível usar essa abordagem.

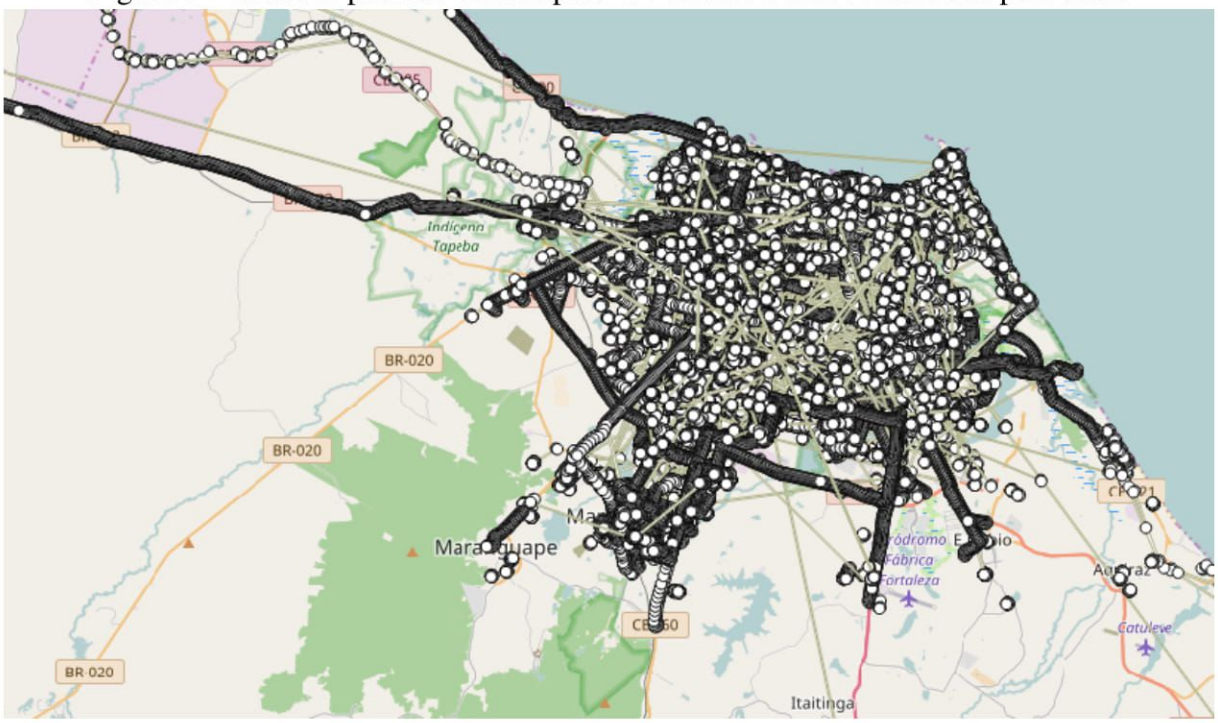
## 5.4 Análise dos resultados

Nessa etapa, os clusters foram salvos no banco para que pudesse ser feita a visualização dos resultados por meio do QGIS [12], que permite recuperar os dados do mesmo e exibir em mapas e figuras.

---

[8] https://trac.osgeo.org/postgis/wiki/UsersWikiplpgsqlfunctionsDistance
[9] http://postgis.net/
[10] https://postgis.net/docs/ST$_{C}entroid.html$
[11] http://scikit-learn.org/stable/modules/generated/sklearn.cluster.DBSCAN.html
[12] http://www.qgis.org/pt$_{B}R/site/$

Figura 6 – Todas as paradas obtidas pelo CB-SMoT. Fonte : Elaborada pelo autor




Fonte : Elaborada pelo autor

Após o Download do mapa da cidade de Fortaleza por meio da extenção do QGIS *OpenLayers* [13], os dados dos *clusters* foram exibidos com cores diferentes, onde cada cor representa um *cluster*, para melhor entendimento.

[13] https://plugins.qgis.org/plugins/openlayers$_p$*lugin*/

# 6 RESULTADOS

Esse capítulo tem como objetivo expor os resultados encontrados pelas atividades explicadas nos procedimentos metodológicos.

## 6.1 Resultados do algoritmo SMoT

Nessa etapa, foi feito um processo de experimentação. Sendo assim, os algoritmos foram executados várias vezes e com parâmetros de tempo diferentes, com o intuito de comparar a variância entre os resultados e ter certeza de qual deles escolher. A Tabela 3 mostra os valores que foram mudados a cada teste, sendo que o $buffer$ diz respeito apenas ao SMoT, enquanto os outros dois são comuns aos dois.

Tabela 3 – Testes de parâmetros

| Tempo mínimo | Trajetórias | $Buffer$ |
|---|---|---|
| 1 minuto | 1 mês | 30 metros |
| 3 minutos | 1 dia | 50 metros |
| 5 minutos | 1 semana | 10 metros |

Fonte : Elaborada pelo autor

Após esses testes, as paradas de 1 minuto foram descartadas por serem muito curtas e as que possuíam 3 minutos ou mais de duração foram usadas para análise e o $buffer$ padrão da ferramenta, no caso 50 metros. No caso das trajetórias foi escolhido uma semana, por ser uma quantidade boa de dados e não demorar muito na execução.

Foram encontradas 536 paradas ao todo, distribuídas em 9 das 12 categorias de regiões de interesse, como mostra a Tabela 1. Os parâmetros passados para obtê-las foram: 3 minutos para o tempo mínimo de duração da parada; 50 metros de $buffer$; trajetórias do dia 1 ao dia 7 de junho de 2017; 237 regiões de interesse que representam os estádios, aeroportos, restaurantes, shoppings, café, clubes noturnos, lojas, lojas eletrônicas, museus, parques, parques de diversão e pontos de táxi.

As Tabelas de 4 a 14 seguem uma mesma estrutura, onde possuem 4 atributos, sendo eles Nome, Quantidade, Tempo e Avaliação, que representam respectivamente o nome da região de interesse, a quantidade de paradas encontradas lá, a média do tempo das parada encontradas e a avaliação dada pelo *Google* a mesma, que reflete a avaliação dos usuários que já visitaram aquele ponto de interesse. As Tabelas estão ordenadas por ordem de popularidade, medida pela

Tabela 4 – Resultados do SMoT por região de interesse

| Categorias | Quantidade de paradas |
|---|---|
| Café | 173 |
| Shopping | 139 |
| Ponto de táxi | 95 |
| Restaurante | 26 |
| Clube Noturno | 25 |
| Loja eletrônica | 19 |
| Parque | 19 |
| Loja | 17 |
| Museu | 17 |
| Parque de diversão | 6 |
| Estádio | 0 |
| Aeroporto | 0 |

Fonte : Elaborada pelo autor

quantidade de paradas encontradas em ponto. Vale ressaltar que não foi obtido resultados em duas categorias, sendo elas Estádio e Aeroporto.

Tabela 5 – Resultados com restaurantes

| Nome | Quantidade | Tempo | Avaliação |
|---|---|---|---|
| Carneiro do Ordones | 4 | 01:27:44 | 4.4 |
| Arre Égua Bar | 4 | 00:27:54 | 4.1 |
| Cangaceiro Sanduíches | 3 | 00:52:53 | 4.1 |
| Ibis Iracema | 2 | 00:19:40 | 4.2 |
| At Home Pub | 2 | 00:18:23 | 4.1 |
| McDonald s | 2 | 00:39:06 | 3.5 |
| Coco Bambu | 2 | 00:12:37 | 4.6 |
| Florence L escale | 1 | 00:05:53 | 4.3 |
| Sal e Brasa | 1 | 00:03:44 | 4.4 |
| Boteco do Arlindo | 1 | 00:03:03 | 4.1 |
| Casa do Frango Sushibar | 1 | 00:28:30 | 4.2 |
| Mercure Meireles | 1 | 00:10:29 | 4.1 |
| Pirata Bar | 1 | 00:03:19 | 4.3 |
| Restaurante Ideal Clube | 1 | 00:09:18 | 4.3 |

Fonte : Elaborada pelo autor

Tabela 6 – Resultados com Shoppings

| Nome | Quantidade | Tempo | Avaliação |
|---|---|---|---|
| Shopping Aldeota | 112 | 19:12:50 | 4 |
| Shopping Parangaba | 10 | 03:08:13 | 4.3 |
| S. dos Fabricantes | 4 | 01:17:03 | 4 |
| Monsenhor Tabosa | 3 | 00:25:43 | 3.8 |
| Revista Vitrine | 2 | 00:12:16 | |
| Shopping Fortaleza Sul | 2 | 00:13:31 | 4.1 |
| Shopping Metrô | 2 | 00:11:15 | 3.9 |
| Shopping Jardins | 2 | 00:40:45 | 4.5 |
| Shopping Benfica | 1 | 00:03:45 | 4.1 |
| Salinas Casa Shopping | 1 | 00:07:25 | 4.3 |

Fonte : Elaborada pelo autor

Tabela 7 – Resultados com Café

| Nome | Quantidade | Tempo | Avaliação |
|---|---|---|---|
| Café Pagliuca | 153 | 26:32:38 | 4.3 |
| Vinyle Café | 10 | 08:11:20 | 4.1 |
| Moykano s tattoo shop | 3 | 01:31:10 | 4.7 |
| Navona Espresso | 2 | 00:10:11 | |
| Padaria Vovó Joana | 2 | 00:06:56 | 4.3 |
| Hard Rock Café | 1 | 00:10:49 | 3.8 |
| Casa Glacê Alimentos | 1 | 00:07:46 | 3 |
| Confeitaria Sublime | 1 | 00:13:41 | 4.5 |

Fonte : Elaborada pelo autor

Tabela 8 – Resultados com Clubes noturnos

| Nome | Quantidade | Tempo | Avaliação |
|---|---|---|---|
| Austin Pub | 12 | 05:07:13 | 4.4 |
| Level Club | 5 | 00:34:26 | 4.2 |
| Pink Elephant Fortaleza | 2 | 01:39:30 | 3.4 |
| Brom s Partyhouse | 1 | 00:03:51 | 3.3 |
| Forró do Damasio | 1 | 00:07:36 | 4.8 |
| Reator 51 | 1 | 00:17:18 | 4.5 |
| Salão belíssima | 1 | 00:03:25 | 4.8 |
| Obará Danças | 1 | 00:04:39 | |
| Bar Boteco Original | 1 | 00:03:06 | 4.3 |

Fonte : Elaborada pelo autor

Tabela 9 – Resultados com Lojas

| Nome | Quantidade | Tempo | Avaliação |
|---|---|---|---|
| 6Bocas Autocenter | 6 | 01:01:45 | 3.9 |
| Paulinas Livraria | 4 | 01:20:05 | 4.2 |
| Mercado Central | 3 | 00:39:08 | 4.2 |
| Lojas Marisa | 3 | 00:17:50 | 3.8 |
| Marisa | 1 | 00:24:04 | 3 |

Fonte : Elaborada pelo autor

Tabela 10 – Resultados com Lojas eletrônicas

| Nome | Quantidade | Tempo | Avaliação |
|---|---|---|---|
| Foto Planalto Color | 4 | 00:59:47 | 3.4 |
| Dr Micro | 2 | 00:09:53 | |
| Atacadão dos Eletros | 2 | 00:41:44 | |
| Click Mix | 2 | 00:07:52 | 3.8 |
| Magazine Luiza | 2 | 00:21:49 | 3.6 |
| ‖ Luiza Centro" | 1 | 00:03:31 | 1.5 |
| Super Film Digital | 1 | 00:33:39 | 4.9 |
| CMS Informatica | 1 | 00:03:26 | 3 |
| Atacadão dos E. | 1 | 00:03:27 | |
| Ibyte Rui Barbosa | 1 | 00:11:04 | 3.7 |
| ‖ Washington Soares | 1 | 00:03:07 | 3.4 |
| Saraiva | 1 | 00:03:17 | 3.9 |

Fonte : Elaborada pelo autor

Tabela 11 – Resultados com Museus

| Nome | Quantidade | Tempo | Avaliação |
|---|---|---|---|
| Museu do Automóvel | 9 | 02:02:09 | 3.5 |
| Museu do Ceará | 3 | 00:35:06 | 4.5 |
| Museu da Escrita | 1 | 00:03:55 | 4.8 |
| Ibeu Art Gallery | 1 | 00:06:20 | |
| Museu do Humor | 1 | 00:07:43 | 4.2 |
| Museu de Arte | 1 | 00:06:04 | 4.5 |
| Cultura Cearense | 1 | 00:14:15 | 4.7 |

Fone : Elaborada pelo autor

Tabela 12 – Resultados com parques

| Nome | Quantidade | Tempo | Avaliação |
|---|---|---|---|
| Praça da Igreja Da Glória | 11 | 03:27:30 | 3.4 |
| Chica Zelosa | 3 | 00:40:49 | 3 |
| Praça do Guajeru | 2 | 00:11:04 | 3.9 |
| Praça Engenheiro Pedro Felipe Borges | 1 | 00:08:21 | 3.9 |
| Parque Guararapes - Fortaleza | 1 | 00:04:21 | 4.1 |
| Ecopoint Parque Ambiental | 1 | 00:08:01 | 4.5 |

Fonte : Elaborada pelo autor

Tabela 13 – Resultados com parques de diversão

| Nome | Quantidade | Tempo | Avaliação |
|---|---|---|---|
| Parquinho Infantil | 3 | 00:29:19 | 5 |
| Ytacaranha Hotel De Serra | 1 | 00:04:15 | |
| Parque Aquatico | 1 | 00:04:27 | 5 |
| Giga Play | 1 | 00:03:16 | 4.1 |

Fonte : Elaborada pelo autor

Tabela 14 – Resultados com pontos de Táxi

| Nome | Quantidade | Tempo | Avaliação |
|---|---|---|---|
| Táxi amigo | 80 | 26:35:21 | 5 |
| Táxi Náutico | 7 | 01:50:25 | 4.7 |
| Hiper Bom Preço | 2 | 00:23:19 | |
| Porto das Dunas | 1 | 00:04:07 | 5 |
| Lig Mototáxi | 1 | 00:03:11 | 4.1 |
| GOLDTAXI | 1 | 00:48:25 | |
| Moto Táxi Amigo | 1 | 00:03:47 | 5 |
| Conjutaxi | 1 | 00:03:51 | |
| MOTO-TÁXI IGUATEMI | 1 | 00:03:01 | 5 |

Fonte : Elaborada pelo autor

A Tabela 15 mostra um ranking com os lugares mais populares encontrados pelo SMoT. Vale ressaltar que as regiões de interesse presentes nas tabelas são apenas aquelas que tiveram paradas identificadas, as demais foram desconsideradas na fase de análise.

Tabela 15 – Ranking das regiões mais populares do SMoT

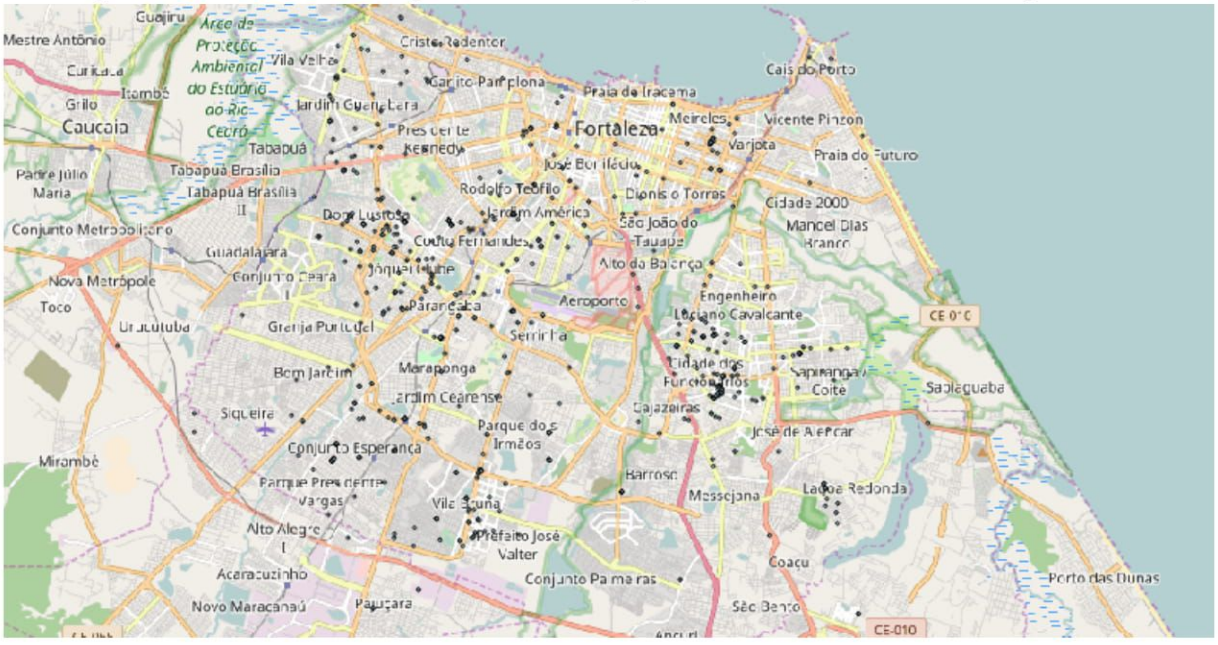
| Nome | Quantidade | Tempo | Avaliação |
|---|---|---|---|
| Café Pagliuca | 153 | 26:32:38 | 4.3 |
| Shopping Aldeota | 112 | 19:12:50 | 4 |
| Táxi amigo | 80 | 51:35:21 | 5 |
| Austin Pub | 12 | 05:07:13 | 4.4 |
| Praça da Igreja Da Glória | 11 | 03:27:30 | 3.4 |
| Vinyle Café | 10 | 08:11:20 | 4.1 |
| Shopping Parangaba | 10 | 03:08:13 | 4.3 |
| Museu do Automóvel | 9 | 02:02:09 | 3.5 |
| Táxi Náutico | 7 | 01:50:25 | 4.7 |
| 6Bocas Autocenter | 6 | 01:01:45 | 3.9 |
| Level Club | 5 | 00:34:26 | 4.2 |

Fonte : Elaborada pelo autor

## 6.2 Resultados do algoritmo CB-SMoT

O primeiro resultado do CB-SMoT retornou 37 mil paradas em toda a cidade, o que pode ser visto na Figura 6. O parâmetro usado como o tempo mínimo foi 3 minutos; O conjunto de trajetórias foram as do dia 1 ao dia 7 de junho de 2017 e para o *eps*, 0.3 e 0.9. Após o pré-processamento feito pela função utmzone do postgis, o número total diminuiu para 1186 paradas, o que pode ser visto na Figura 7.

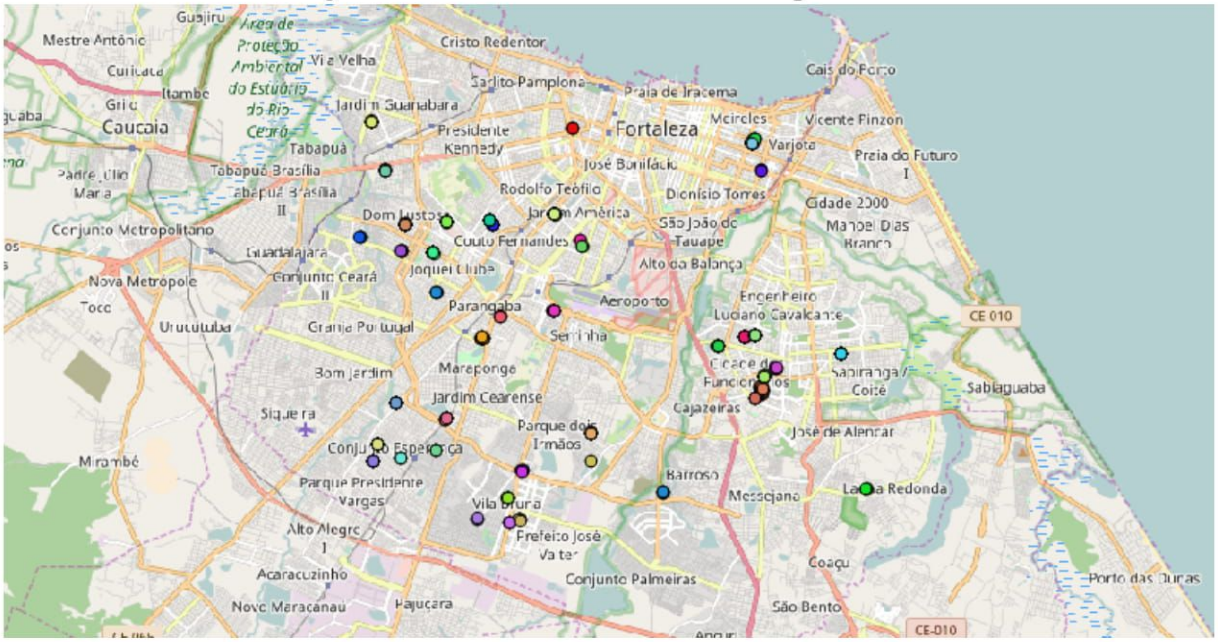
Figura 7 – Paradas obtidas pelo CB-SMoT após o filtro. Fonte : Elaborada pelo autor




Fonte : Elaborada pelo autor

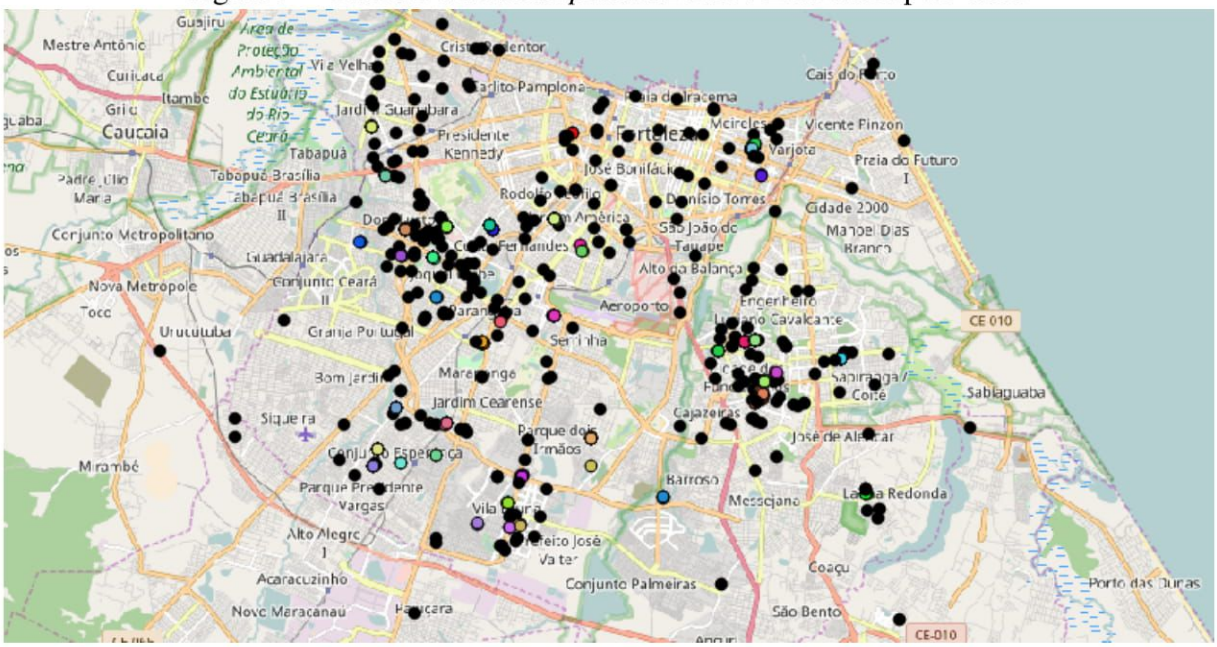
Aqui, também foram feitos testes para chegar em um consenso de que parâmetros usar para obter um resultado satisfatório. Neste caso, um *eps* pequeno (0.05) foi escolhido para manter os pontos dos clusters sobrepostos, e serem caracterizados como regiões de interesse. Posteriormente, ao executar o DBSCAN com o centro das áreas das paradas, foram encontrados 44 clusters, sendo eles as regiões populares encontradas. Apos alguns testes, os argumentos passados para o algoritmo foram 0.05 km por radianos de *eps* e 5 de *minpoints*, pois quando o eps passado era maior, os clusters ficavam muito distantes, e o ideal é que ficassem sobrepostos. A Figura 8 representa os clusters encontrados e a Figura 9 mostra os clusters acrescentados dos *noisepoints*.

Figura 8 – Clusters. Fonte : Elaborada pelo autor




Fonte : Elaborada pelo autor

Figura 9 – Clusters com *noisepoints*. Fonte : Elaborada pelo autor




Fonte : Elaborada pelo autor

A tabela 16 mostra um *raking* com as regiões mais populares, onde a popularidade se dá pela quantidade de pontos encontrados nela.

Tabela 16 – Clusters mais populares

| Identificador do Cluster | Bairro | Popularidade |
|---|---|---|
| 13 | Sapiranga/Coité | 52 |
| 15 | Cidade dos funcionários | 43 |
| 0 | Damas | 39 |
| 5 | Jardim da Oliveiras | 34 |
| 10 | Aldeota | 27 |
| 41 | Maraponga | 23 |
| 18 | Dom Lustosa | 22 |
| 42 | Cidade dos funcionários | 21 |
| 4 | Dom Lustosa | 21 |

Fonte : Elaborada pelo autor

Após o processo de descobrimento das regiões mais populares, foi usado novamente o *script* utilizado para coletar pontos de interesse para que a partir dos centroides do *cluster* encontrado na Aldeota, os possíveis pontos de interesse buscadas nesses lugares fossem identificados. O *script* retornou um conjunto de locais dentro de 50 metros de raio a partir do centroides, que é mostrado na Tabelas 17.

Tabela 17 – Regiões de interesse próximas ao cluster da Aldeota

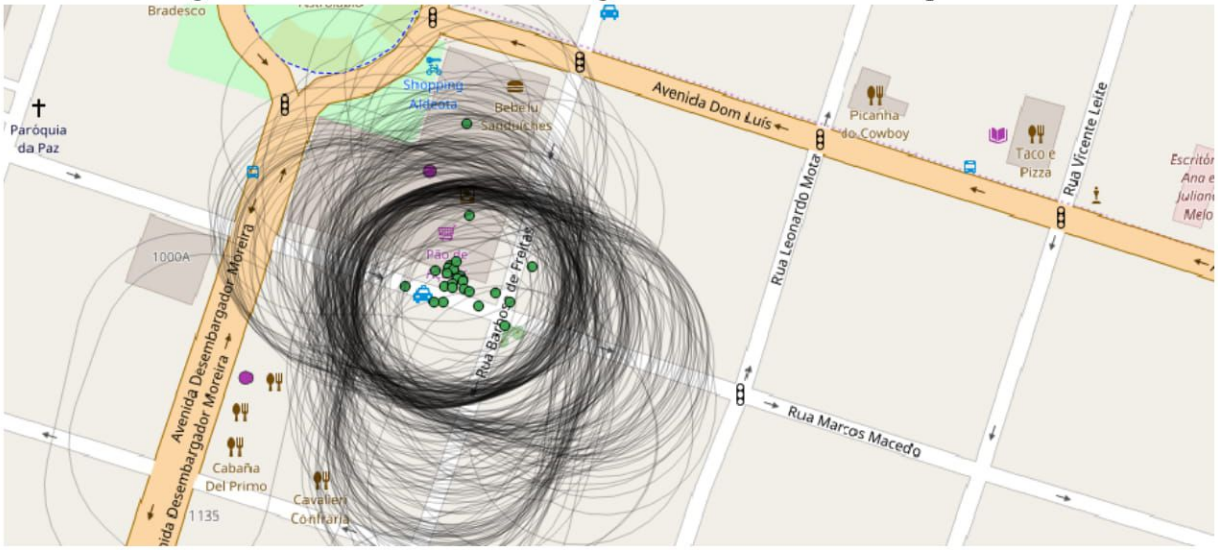
| Nome | Avaliação |
|---|---|
| Zerifelli | 5.0 |
| Shopping Aldeota | 4.1 |
| Lojas Marisa | 5.0 |
| Imobiliária Mauro Sales | 5.0 |
| Arcoíris Cinemas | 4.2 |
| Polishop | 3.8 |
| Planeta Brinquedos | 4.7 |
| Skyler Aldeota | |
| Lojas Tricolaço | 5.0 |
| Doutor Cell | 4.7 |
| Handara | 3.0 |
| WR Engenharia | |
| CVC Shopping Aldeota | 5.0 |
| Meireles Negócios Imobiliários | |
| Sindicato Patronal dos Hotéis Restaurantes Bares Similares | 2.0 |

Fonte : Elaborada pelo autor

## 6.3 Comparação dos Resultados dos algoritmos

Apesar do SMoT e o CB-SMoT usarem abordagens diferentes para chegar nos resultados, ao comparar os obtidos no presente trabalho, foi visto que em alguns pontos houve paradas identificadas em locais iguais ou próximos. A Figura 10 mostra visualmente onde isso ocorre. Os pontos em verde representam o resultado do CB-SMoT e as regiões circuladas em preto os do SMoT.

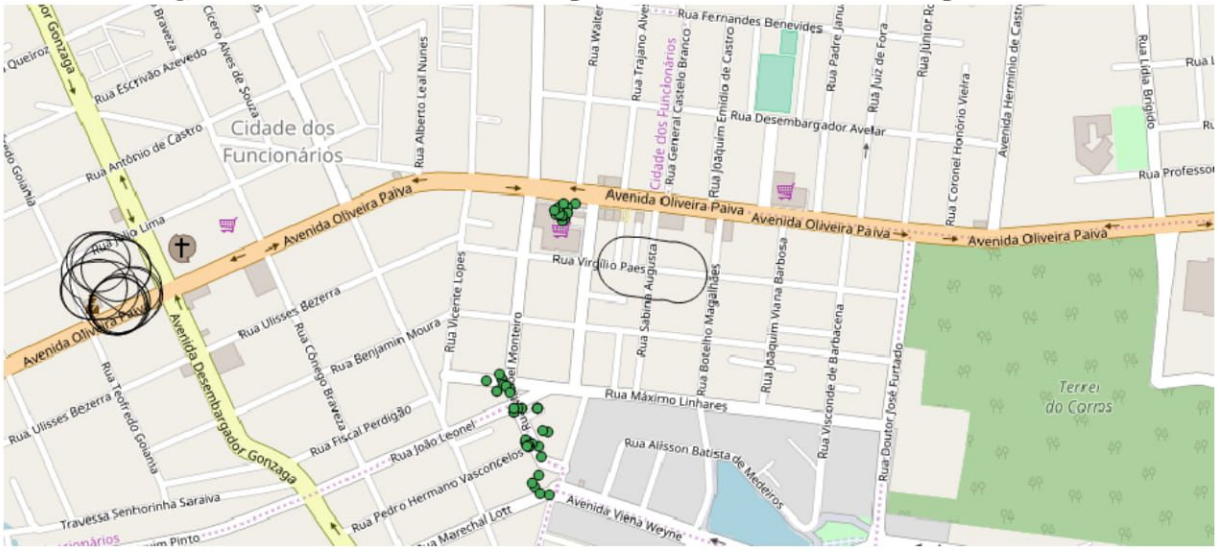
Figura 10 – Resultados em áreas iguais . Fonte : Elaborada pelo autor




Fonte : Elaborada pelo autor

Ao contrario da Figura anterior, a Figura 11 mostra um ponto de interesse identificado pelo CB-SMoT e desconhecido pelo SMoT, que detectou próximo a ele, mas não o mesmo. Esse resultado já era esperado visto que o CB-SMoT encontra regiões de interesse que são desconhecidas e também as conhecidas. Visto que a gestão de trafego na cidade de Fortaleza ainda deixa a desejar, a influencia disso nos resultados do CB-SMoT é grande, o que pode ser considerada uma desvantagem dessa abordagem em outras cidades que sofram do mesmo problema. Porém, graças ao parâmetro de *eps* ser calculado por meio de um intervalo de dois números, a ferramente mostra um nível de abstração que ajuda bastante, pois nem todos os usuários tem uma noção precisa de distancias entre pontos. Já o que dificulta o uso do SMoT é o parâmetro de regiões de interesse, pois no caso de Fortaleza, que é uma cidade enorme, se torna complicado saber onde focar para obter resultados, visto que os táxis podem estar espalhados em qualquer lugar. Já o resultado é bastante completo e já ponto para ser usado no caso de recomendação, o que diminui bastante o trabalho em relação a o CB-SMoT.

Figura 11 – Resultados em áreas próximas. Fonte : Elaborada pelo autor




Fonte : Elaborada pelo autor

# 7 CONSIDERAÇÕES FINAIS

Ao longo do desenvolvimento do presente trabalho, foram usadas abordagens para encontrar paradas em trajetórias. As duas abordagens renderam resultados de forma a se completarem, onde obtivemos regiões de interesse populares conhecidas previamente, e regiões populares desconhecidas. Pode-se usar os resultados aqui encontrados tanto para fim de recomendação para usuários, como para controle dos táxis que interessa, por exemplo, os donos do aplicativo que por meio das regiões populares, sabem onde investir mais para buscar agradar os clientes já existentes, ou em lugares onde sua presença não é tão forte, e buscar novas áreas de atuação.

Foi visto também que o uso do CB-SMoT em lugares que possuem um transito maior pode influenciar bastante no resultado, de forma a deixá-lo difícil de analisar e exigir um pouco mais de tempo e outras técnicas da literatura. Já o SMoT pode facilmente ser usado, mas se limita bastante por se concentrar em pontos já conhecidos, o que dificulta a descoberta de novas regiões populares, fazendo com que os trabalhos que pretendem usa-lo tenham uma ideia previa de onde os táxis se encontram, o que nem sempre é o caso.

Visto que a ferramenta é antiga e o código pode ser considerado legado, uma das maiores dificuldades percebidas foi como encontrar uma forma de deixar os dados de uma maneira que os algoritmos pudessem obter algum resultado. Como não havia documentação, essa situação obrigou-nos a buscar outras formas de usa-la, baixando o projeto, importando em uma IDE e precisando entender parte do código fonte, o que não é a ideia por trás do Weka, visto que o mesmo tem uma interface gráfica e maneiras de usá-lo sem esse nível de entendimento.

Um trabalho futuro pode vir a ser recomendação para usuários dos pontos mapeados pela Google que são mais populares, no itinerário e nas regiões populares aqui identificadas. Isso pode ser melhor fundamentado se o tempo for usado como uma métrica, o que pela experiencia obtida no presente trabalho, obrigaria o autor a usar mais dados, possivelmente todos os dados do mês. Esse é um desafio que engloba também ter um poder computacional maior do que o usado aqui, visto que a quantidade de pontos seria gigante, e poderia fazer com que a maquina usada para isso travasse ou demorasse dias para terminar o processamento, o que poderia influenciar bastante no uso dessas abordagens num âmbito empresarial por exemplo, onde se tem pressa diante dos prazos.

# REFERÊNCIAS

ALVARES, L. O.; BOGORNY, V.; KUIJPERS, B.; MACEDO, J. A. F. de; MOELANS, B.; VAISMAN, A. A model for enriching trajectories with semantic geographical information. In: ACM. **Proceedings of the 15th annual ACM international symposium on Advances in geographic information systems**. [S.l.], 2007. p. 22.

ALVARES, L. O.; BOGORNY, V.; KUIJPERS, B.; MOELANS, B.; FERN, J. A.; MACEDO, E.; PALMA, A. T. Towards semantic trajectory knowledge discovery. **Data Mining and Knowledge Discovery**, 2007.

ALVARES, L. O.; PALMA, A.; OLIVEIRA, G.; BOGORNY, V. Weka-stpm: from trajectory samples to semantic trajectories. In: **Proceedings of the XI workshop de Software Livre, WSL**. [S.l.: s.n.], 2010. v. 10, p. 164–169.

BRAKATSOULAS, S.; PFOSER, D.; TRYFONA, N. Modeling, storing and mining moving object databases. In: IEEE. **Database Engineering and Applications Symposium, 2004. IDEAS'04. Proceedings. International**. [S.l.], 2004. p. 68–77.

BRAMER, M. **Principles of data mining**. [S.l.]: Springer, 2007. v. 180.

BRILHANTE, I. R.; BERLINGERIO, M.; TRASARTI, R.; RENSO, C.; MACEDO, J. A. F. de; CASANOVA, M. A. Cometogether: Discovering communities of places in mobility data. In: IEEE. **Mobile Data Management (MDM), 2012 IEEE 13th International Conference on**. [S.l.], 2012. p. 268–273.

CABENA, P.; HADJINIAN, P.; STADLER, R.; VERHEES, J.; ZANASI, A. **Discovering data mining: from concept to implementation**. [S.l.]: Prentice-Hall, Inc., 1998.

CHEN, C.-C.; CHUANG, M.-C. Integrating the kano model into a robust design approach to enhance customer satisfaction with product design. **International Journal of Production Economics**, Elsevier, v. 114, n. 2, p. 667–681, 2008.

DENG, Z.; JI, M. Spatiotemporal structure of taxi services in shanghai: Using exploratory spatial data analysis. In: IEEE. **Geoinformatics, 2011 19th International Conference on**. [S.l.], 2011. p. 1–5.

ESTER, M.; KRIEGEL, H.-P.; SANDER, J.; XU, X. et al. A density-based algorithm for discovering clusters in large spatial databases with noise. In: **Kdd**. [S.l.: s.n.], 1996. v. 96, n. 34, p. 226–231.

FAYYAD, U.; PIATETSKY-SHAPIRO, G.; SMYTH, P. From data mining to knowledge discovery in databases. **AI magazine**, v. 17, n. 3, p. 37, 1996.

GIANNOTTI, F.; NANNI, M.; PEDRESCHI, D.; RENSO, C.; TRASARTI, R. Mining mobility behavior from trajectory data. In: IEEE. **Computational Science and Engineering, 2009. CSE'09. International Conference on**. [S.l.], 2009. v. 4, p. 948–951.

HAND, D. J.; MANNILA, H.; SMYTH, P. **Principles of data mining**. [S.l.]: MIT press, 2001.

MANCHANDA, P.; ANSARI, A.; GUPTA, S. The "shopping basket": A model for multicategory purchase incidence decisions. **Marketing Science**, INFORMS, v. 18, n. 2, p. 95–114, 1999.

PALMA, A. T.; BOGORNY, V.; KUIJPERS, B.; ALVARES, L. O. A clustering-based approach for discovering interesting places in trajectories. In: ACM. **Proceedings of the 2008 ACM symposium on Applied computing**. [S.l.], 2008. p. 863–868.

PARKINSON, B. W.; ENGE, P. K. Differential gps. **Global Positioning System: Theory and applications.**, v. 2, p. 3–50, 1996.

SUNG, T. K.; CHANG, N.; LEE, G. Dynamics of modeling in data mining: interpretive approach to bankruptcy prediction. **Journal of management information systems**, Taylor & Francis, v. 16, n. 1, p. 63–85, 1999.

YUE, Y.; ZHUANG, Y.; LI, Q.; MAO, Q. Mining time-dependent attractive areas and movement patterns from taxi trajectory data. In: IEEE. **Geoinformatics, 2009 17th International Conference on**. [S.l.], 2009. p. 1–6.