UNIVERSIDADE DO VALE DO RIO DOS SINOS – UNISINOS
UNIDADE ACADÊMICA DE PESQUISA E PÓS-GRADUAÇÃO
ESPECIALIZAÇÃO EM BIG DATA, DATA SCIENCE E DATA ANALYTICS

Marcelo Lopes Remião

EMPREGANDO CLUSTERING PARA IDENTIFICAÇÃO DE TRAJETÓRIAS DE TRANSPORTE DE ILÍCITOS POR VEÍCULOS

São Leopoldo

2020

UNIVERSIDADE DO VALE DO RIO DOS SINOS – UNISINOS
UNIDADE ACADÊMICA DE EDUCAÇÃO ONLINE
ESPECIALIZAÇÃO EM BIG DATA, DATA SCIENCE E DATA ANALYTICS

Marcelo Lopes Remião

EMPREGANDO CLUSTERING PARA IDENTIFICAÇÃO DE TRAJETÓRIAS DE TRANSPORTE DE ILÍCITOS POR VEÍCULOS

Trabalho de Conclusão de Curso apresentado como requisito parcial para a obtenção do título de Especialista em *Big Data, Data Science e Data Analytics*, pelo Curso de Pós-Graduação Lato Sensu em *Big Data, Data Science e Data Analytics* da Universidade do Vale do Rio dos Sinos – Unisinos
Orientadora: Profa. Dra. Patricia A. Jaques Maillard

São Leopoldo

2020

EMPREGANDO CLUSTERING PARA IDENTIFICAÇÃO DE TRAJETÓRIAS DE TRANSPORTE DE ILÍCITOS POR VEÍCULOS

Marcelo Lopes Remião[1]

[1]Especialização em Big Data, Data Science e Data Analytics

Universidade do Vale do Rio dos Sinos (UNISINOS)

marcelo.remiao@prf.gov.br

Abstract *Spatio-temporal data can be produced in an Intelligent Transport System (ITS) by License Plate Recognition devices, or LPR (License Plate Recognition), and are generated from the registration of vehicles through equipment, fixed or mobile , installed at certain points along routes along highways. These data are useful because they allow to reproduce the trajectories traveled by vehicles. This work seeks to identify trajectories used to transport illicit vehicles by land vehicles, presenting comparative results with significant trajectories, where the DBSCAN algorithm is able to identify the concentration of the most frequent trajectories formed by points of greater density.*

Resumo*. Dados espaço-temporais podem ser produzidos em um Sistema de Transporte Inteligente (ITS) por dispositivos de Reconhecimento de Placas de Veículos, ou LPR (License Plate Recognition), e são gerados a partir do registro de passagem de veículos por equipamentos, fixos ou móveis, instalados em determinados pontos ao longo de trajetos por rodovias. Esses dados são úteis porque permitem reproduzir as trajetórias percorridas por veículos. Esse trabalho busca a identificação de trajetórias utilizadas para o transporte de ilícitos por veículos terrestres, apresentando resultados comparativos com trajetórias significativas, onde o algoritmo DBSCAN mostra-se capaz de identificar a concentração de trajetórias mais frequentes formadas por pontos de maior densidade.*

Keywords: Trajectory Clustering, K-means, DBSCAN

1. Introdução

A mobilidade é fundamental em um mundo globalizado, onde pessoas, bens, dados e até ideias se movem em volumes crescentes a velocidades e distâncias cada vez maiores. Os recentes avanços nas tecnologias de localização (GPS, RFID), comunicação sem fio, computação móvel e detecção ambiental permitem um rastreamento quase onipresente de indivíduos e objetos em movimento no espaço-tempo, resultando em grandes volumes de informações que capturam processos dinâmicos de alta relevância socioeconômica. Seja comunicação móvel e sem fio [Gonzalez *at al*. 2008], transporte [Kellerer *at al*. 2001], videomonitoramento para aplicações de segurança e esportes, ou mesmo pesquisa fundamental sobre ecologia comportamental [Holyoak *at al*. 2008], os dados de movimento são acumulados em volumes e granularidades[1] anteriormente não vistos.

O Brasil é o país que tem a maior concentração rodoviária de transporte de cargas e passageiros entre as principais economias mundiais. Segundo dados da Confederação Nacional do Transporte (CNT), o transporte rodoviário é o segmento de maior participação na matriz de transporte de cargas (61,1%) e o principal modo de deslocamento de passageiros, independentemente da distância [Transporte 2017c], seguido pelo modal ferroviário (20,7%) e aquaviário (13,6%). Os modais dutoviário e aéreo respondem por menos de 5% da matriz de transporte de cargas.

Dados publicados em 2018 pela Central Intelligence Agency (CIA) do governo dos Estados Unidos [CIA 2018], revelam que o Brasil possui cerca de 2 milhões de quilômetros de rodovias, figurando como a quarta maior malha rodoviária do mundo, sendo que apenas 12,3% delas são pavimentadas.

Segundo dados do IBGE [IBGE 2016], o modal rodoviário foi responsável por 55,2% do PIB do setor de transporte em 2014, contribuindo significativamente para a geração de riquezas no país. Contudo, a relevância do transporte rodoviário não é percebida apenas em relação aos demais modais de transporte. Ele foi responsável por 12,7% do PIB do setor de serviços não financeiros, sendo o segundo que mais contribuiu para a geração de valor nesse segmento.

Mas as rodovias também servem como caminho para o transporte e a prática de atividades ilícitas como o tráfico de entorpecentes, o tráfico de armas e munições, o contrabando de mercadorias, os crimes ambientais e o roubo de cargas, dentre outros. Estima-se que a maior parte dos produtos de comercialização ilícita circulam por rodovias devido à sua permeabilidade e diversidade de rotas disponíveis. Alternativas como o transporte aéreo, ferroviário ou aquaviário, além de possuírem maior controle no embarque e desembarque de cargas, apresentam um custo de operação, em geral, mais elevado que o modal rodoviário, apresentando uma capacidade mais limitada para a distribuição de cargas ao longo do território nacional.

Diante deste cenário, coletar informações a respeito da circulação de veículos e mercadorias em rodovias apresenta-se como uma importante ferramenta para o

[1] Quando falamos de menor granularidade, ou granularidade fina, significa maior detalhamento (menor sumarização) dos dados. Maior granularidade, ou granularidade grossa, significa menor detalhamento (maior sumarização). Assim podemos notar que a granularidade e o detalhamento são inversamente proporcionais.

planejamento logístico do país, além de subsidiar a adoção de políticas públicas de redução de acidentes e de enfrentamento aos crimes transfronteiriços que utilizam as rodovias como meio de locomoção. Um sistema integrado de monitoramento pode ser capaz de coletar dados de passagem de veículos, como a geolocalização e o momento de do registro (dados espaço-temporais), gerando informações como trajetórias percorridas por veículos, permitindo ainda que se extraia conhecimento sobre o comportamento de viajantes.

A integração desses sistemas tem potencial para gerar um grande volume de dados sobre circulação de veículos por vias públicas, com informações de origem e destino de rotas, sobre o impacto de circulação de mercadorias nas economias locais, regionais e nacional, além de rotas e veículos utilizados para o cometimento de ações delituosas.

À medida que a tecnologia para rastrear objetos em movimento se torna mais barata e precisa, a quantidade e a disponibilidade de dados de movimento armazenados aumentam rapidamente. A maioria dos dados de movimento é capturada e armazenada na forma de trajetórias, definida como "uma sequência de locais com registro de tempo" [Gudmundsson *at al* 2012]. No entanto, a análise deste crescente volume de dados sobre trajetória pode ser desafiadora, em grande parte, devido à forma como o movimento pode ter inúmeras representações de trajetórias discretizadas diferentes

Sistemas integrados equipados com dispositivos para o reconhecimento de placas de veículos (LPR) geram um grande volume de dados sobre passagem de veículos. Dados como a foto do veículo, sua placa, velocidade, dimensões, localização e momento de sua passagem são constantemente captados e armazenados em bases de dados. Na Figura 1 temos representado um Sistema de Monitoramento de Veículo que capta as informações de passagens de veículos por câmeras instaladas em pontos de rodovias e as armazena em bases de dados.

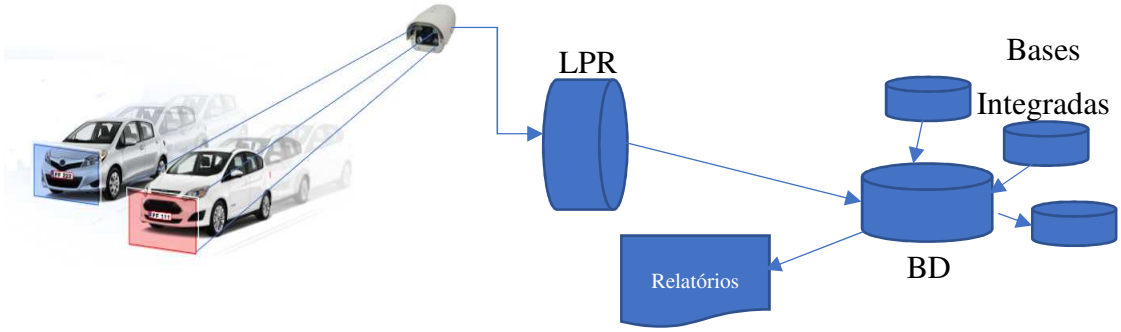


Figura 1: Sistema de Monitoramento de Veículos em Rodovias

Técnicas de mineração de dados podem ser úteis para extrair dessas bases de dados conhecimentos mais complexos, onde se possa identificar padrões de comportamento no deslocamento de veículos em rodovias, mapear as principais rotas e classificar os comportamentos de veículos segundo a atividade que está relacionada.

1.1 Justificativa

O conhecimento extraído a partir dos movimentos de objetos é útil em muitos contextos. Com o aumento da disponibilidade de dados e o número de métodos utilizados para extração e classificação, o estudo sobre padrões de comportamento de objetos em movimento ganha importância em vários domínios, incluindo planejamento urbano, controle de fluxo de tráfego, saúde pública, interações sociais e segurança pública, dentre muitos outros. Encontrar padrões de trajetória é muito útil para se aprender interações entre objetos em movimento. Uma variedade de padrões de trajetória tem sido proposta na literatura de trajetórias [Masciari *at al.* 2013].

Informações do Departamento Nacional de Trânsito (DENATRAN) apontam que em julho de 2020 a frota de veículos do Brasil chegou a 106.289.700 veículos [DENATRAN, 2020]. Sistemas de monitoramento de veículos são capazes de gerar dados sobre a frota de veículos circulante, em tempo real, identificando padrões de deslocamento e parada de veículos, sendo capaz de detectar as variações de rotas utilizadas por viajantes.

No Brasil, os governos têm investido nos últimos anos em tecnologias para o reconhecimento de placas de veículos, ou LPR *(License Plate Recognition)*, para o monitoramento de veículos em rodovias e vias urbanas. O Ministério da Justiça e Segurança Pública apresentou em 2019, dentre seus Projetos Estratégicos para o Combate aos Crimes de Corrupção, Crime Organizado e Crimes Violentos [MJSP 2019], o Projeto Alerta Brasil, um programa de monitoramento de veículos nas rodovias federais por meio de câmeras, gerenciado pela Polícia Rodoviária Federal (PRF). O sistema prevê também a integração com outros sistemas semelhantes, gerenciados por outros órgãos federais, por estados e municípios, que também objetivam o monitoramento de veículos em áreas ou regiões sob sua circunscrição. Nos municípios esses sistemas são conhecidos como *cercamento eletrônico* e visam um maior controle do tráfego de veículos nas cidades.

A evolução tecnológica tem propiciado uma maior capacidade de armazenamento e processamento de dados produzidos através de dispositivos que coletam dados e informações e geram um grande volume de dados. Esses dados precisam ser processados, analisados e transformados em *insights*. Para tanto, algumas etapas devem ser consideradas neste processo: preparação dos dados, construção de um modelo analítico e conclusão a partir dos conhecimentos adquiridos, ou seja, transformar dados "brutos" em *insights*.

1.2 Motivação

O uso de técnicas de mineração de dados pode transformar os métodos tradicionais de abordagem policial, partindo de um modo tradicional de abordagem aleatória para um modelo de policiamento orientado por inteligência [OSCE, 2017]. Esse modelo permite que se realizem abordagens mais assertivas e seguras, uma vez que os alvos são selecionados para abordagem através de modelos matemáticos aplicados sobre os dados espaço-temporais gerados a partir dos sistemas de monitoramento de veículos, podendo-se agregar a isso o conhecimento produzido pela inteligência policial. Isso torna as abordagens mais seguras aos policiais e motoristas envolvidos nas fiscalizações, uma vez que é possível prever com antecedência o emprego de recursos materiais e humanos, além de permitir avaliar melhor os riscos envolvidos na ação. Além disso, modelos

baseados em técnicas de aprendizagem de máquina são capazes de detectar alterações nos padrões de comportamento de motoristas infratores, o que permite reprogramar o planejamento das ações de fiscalização e policiamento.

Não encontramos na pesquisa realizada trabalhos que utilizem um conjunto de dados produzidos a partir de dispositivos dotados de câmeras e processados em software de LPR. Grande parte das pesquisas produzidas valem-se de dados produzidos a partir de dispositivos dotados com tecnologia de GPS ou RFID, como se pode ver em 2.2 Movimento e Trajetórias, quando se fala em dados da trajetória. Desenvolver uma pesquisa a partir de bases de dados gerada por Sistemas de Transporte Inteligente (ITS) mostra-se um grande desafio, pois o conjunto de dados apresenta características que inspiram certos cuidados na preparação e tratamento dos dados para um processo de classificação ou clusterização de dados.

1.3 Objetivos

1.3.1 Objetivo Geral

Esse trabalho tem como objetivo geral identificar trajetórias rodoviárias de veículos transportando produtos ilícitos.

1.3.2 Objetivos Específicos

Como objetivos específicos este trabalho propõe-se a realizar alguns estudos sobre padrões de comportamento de viajantes, dentre eles:

- Construir ferramentas para a extração e filtragem de dados de trajetórias de veículos a partir do histórico de passagens de veículos por Sistemas de Transporte Inteligentes (ITS)

- Identificar agrupamento de trajetórias percorridas por veículos envolvidos em atividades em atividades ilícitas

- Explorar padrões de agrupamento de trajetórias relacionadas a determinadas atividades ilícitas utilizando-se técnicas de clusterização (K-*means* e DBSCAN)

1.4 Metodologia

1.4.1 Pesquisa Científica

Este trabalho propõe-se a desenvolver um modelo de pesquisa descritiva quantitativa (estudo de caso) sobre padrões de comportamento de viajantes e rotas de veículos em rodovias.

Iniciou-se a pesquisa através do levantamento de publicações científicas utilizando-se os buscadores *Google Scholar* (Google Acadêmico, em português) e a

ferramenta de busca em periódicos da CAPES/MEC. Palavras chaves como *traveler behavior, trajectory data mining, clustering trajectories e Intelligence Transportation Systems* foram utilizadas para a realização de busca por publicações que demonstraram uma interrelação entre conceitos e técnicas utilizadas para a resolução de problemas similares ao proposto no presente trabalho.

1.4.2 CRISP-DM

> O conhecimento extraído na mineração de dados de trajetória pode ser usado em muitos lugares e domínios de aplicação. Algumas das aplicações possíveis são agrupamento de veículos em movimento em uma cidade, classificação de padrões de compra, localização sequencial padrões na análise da cesta de mercado, compartilhar previsão de tendências de participação de mercado e assim por diante [Yuan *at al.* 2015].

Para o processo de mineração de dados adotou-se a metodologia conhecida como **CRISP-DM (*Cross-Industry Standard Process of Data Mining*)** [Larouse and Larouse 2014]. Ela fornece uma abordagem estruturada para o planejamento de um projeto de mineração de dados e estrutura-se em seis fases, ou etapas:

1 – *Compreensão dos Negócios*: o conhecimento sobre o negócio e sobre os inerentes processos mercadológicos é de fundamental importância para que se definam os objetivos da mineração de dados.

2 – *Entendimento dos Dados*: deve-se descrever os dados de maneira clara e objetiva, sempre explicitando as diversas fontes de obtenção e eventuais comportamentos de interdependência entre variáveis.

3 – *Preparação dos Dados*: análises preliminares dos dados, com eventuais tratamentos sobre *outliers* ou *missing values*, podem ser de grande utilidade para que os métodos de data mining sejam aplicados corretamente. O próprio agrupamento de variáveis ou a categorização por meio de determinado critério pode tornar uma técnica mais adequada do que outra, respeitando os objetivos da análise.

4 – *Modelagem*: diversas técnicas podem ser aplicadas, como a elaboração de técnicas exploratórias, a estimação de modelos confirmatórios ou a implementação de algoritmos, sempre com base nos objetivos propostos.

5 – *Análise dos Resultados*: nesta etapa, é de fundamental importância que participem tanto conhecedores do negócio quanto estatísticos e especialistas nos dados, a fim de que sejam elaboradas avaliações sobre os achados na etapa anterior, a partir da análise de testes e validações.

6 – *Divulgação dos Resultados*: após a modelagem e a análise dos *outputs*, é necessário que todos os envolvidos tomem ciência dos resultados encontrados, a fim de que seja possível a implantação de procedimentos de gestão.

Cada uma dessas fases pode ser observada no presente trabalho, sobretudo no Capítulos 4. A Figura 2 mostra um esquema que descreve o processo de mineração de dados que permite produzir conhecimento a partir dos dados disponíveis.

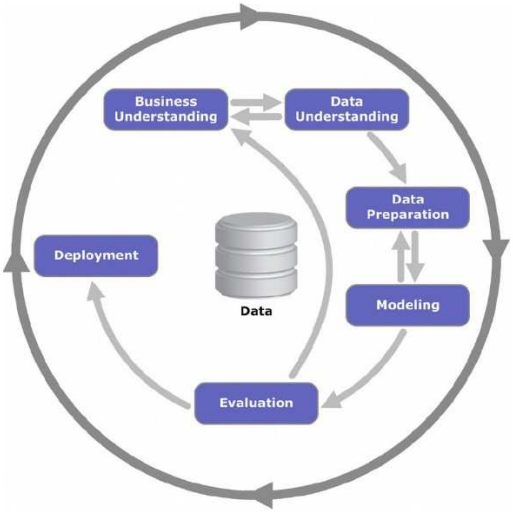


Figura 2: Metodologia CRISP-DM (Cross-Industry Standard Process of Data Mining) Fonte: adaptado de (Larose e Larose, 2014]

Embora algumas técnicas de mineração de dados sejam bastante novas, a mineração de dados em si não é uma nova tecnologia, no sentido de que as pessoas analisam dados em computadores desde que os primeiros computadores foram inventados. E séculos antes disso sem os computadores. A mineração de dados tem assumido muitos nomes, como descoberta de conhecimento, inteligência de negócios, modelagem e análise preditiva. Mas o mais importante é que suas principais tarefas estejam relacionadas à Descrição (ex.: Estatística); Exploração e Visualização de Dados (ex.: Online Analytical Processing – OLAP, Construção de Mapas); Classificação e Predição (ex.: Generalized Linear Models – GLM, Artificial Neural Networks – ANN); Clustering (ex.: Hierarchical Clustering, K-Means Clustering, Self-Organizing Maps – SOM, Árvores de Decisão); Regras de Associação (ex.: Principal Component Analysis – PCA, Análise de Correspondência, Multidimensional Scaling – MDS); Otimização e Simulação (ex.: Programação Linear, Inteira e em Redes, Simulação de Monte Carlo).

Muitas são as ferramentas e softwares desenvolvidos para facilitar a implementação de *Data Mining* por profissionais das mais diversas áreas. Entre os quais, merecem destaque Python, Stata, IBM SPSS Modeler, RStudio, SAS Enterprise Miner, WEKA, KNIME, dentre muitos outros.

Este trabalho está dividido em Capítulo 1 são apresentados os fatores justificam a realização do presente trabalho, suas motivações e objetivos, assim como a metodologia CRISP-DM, empregada para o desenvolvimento da pesquisa. Já no Capítulo 2 alguns conceitos básicos como *Intelligent Transportation Systems* (ITS), movimento, trajetórias e *clustering*, abordados ao longo do trabalho, são definidos e apresentados. O Capítulo 3 consolida alguns trabalhos relacionados ao desenvolvido nesta pesquisa, mostrando um pouco sobre o estado da arte sobre *clustering* de trajetórias. O Capítulo 4 mostra o desenvolvimento do trabalho proposto, descrevendo a característica dos dados explorados, apresentando os principais desafios no desenvolvimento da pesquisa, suas limitações e o processo de preparação e tratamento dos doados, apresentando o modelo desenvolvido e seus experimentos. O Capítulo 5 apresenta os resultados obtidos com as

técnicas de filtragem e clusterização de dados empregadas e avalia as técnicas de clusterização abordadas, avaliando suas performances. Finalmente, o Capítulo 6 propões trabalhos futuros relacionados com a presente pesquisa e tece algumas considerações finais.

2. Conceitos Básicos

2.1 Sistemas de Transporte Inteligentes (ITS)

Os Sistemas de Transporte Inteligente, ou em inglês *Intelligent Transportation Systems* (ITS), são sistemas que utilizam tecnologia da informação como um mecanismo de resolução de problemas relacionados ao transporte em geral [Papageorgiou, 2003]. Automação de autoestradas, sistemas automáticos de coletas de pedágios, sistemas de informação ao usuário e sistemas de controle de tráfego são exemplos de tecnologias utilizadas nos ITS.

Seu objetivo não é apenas melhorar as condições do tráfego de veículos, mas também tornar o setor de transportes mais seguro, mais sustentável e eficiente, evitando os inconvenientes causados pelos congestionamentos dos tráfegos urbanos e efeitos dos problemas climáticos sobre o tráfego. Para isso, o foco é melhorar a gerência dos recursos das cidades e aumentar a comodidade das pessoas através do uso de serviços de informação e alerta [Cunha *et al*. 2017]. Com o advento dos computadores pessoais e a globalização das atividades econômicas, os recursos de ITS passaram a ser assimilados por usuários, operadores e gestores com pouca ou nenhuma familiaridade com níveis básicos de conhecimento tecnológico específico.

Conforme (Sorensen, 2012) a partir dos anos 2000 houve uma evolução nas tecnologias dos ITS. Esta evolução pode ser observada na Tabela 1 que descreve o período em que cada fase se desenvolveu, bem como a tecnologia que foi empregada.

Tabela 1: Gerações de Sistemas de Transportes Inteligentes

	Geração	Período	Tecnologia
1	Primeira Geração (ITS 1.0)	2000	Infraestrutura unidirecional
2	Segunda Geração (ITS 2.0)	2000 – 2003	Tecnologia de comunicação bidirecional
3	Terceira Geração (ITS 3.0)	2004 – 2005	Operações automatizadas do veículo e operações automatizadas e interativas do sistema e gerenciamento do sistema
4	ITS (ITS 4.0)	2006 – 2011	Multimodal incorporando dispositivos móveis pessoais, veículos, infraestrutura e redes de informações para operações do sistema, bem como soluções de mobilidade contextual pessoal

Fonte: (Sorensen, 2012)

Os ITS utilizam diferentes tecnologias nos vários setores dos transportes. Segundo Sussman (2000), os ITS podem ser categorizados em:

a) **Sistemas Avançados de Transporte Público:** os sistemas avançados de transporte público, ou em inglês *Advanced Public Transportation Systems* (APTS), têm como objetivo melhorar a segurança e a efetividade dos sistemas de transporte público. Os benefícios para os usuários incluem a redução dos tempos de espera, a segurança e a facilidade para o pagamento da tarifa, bem como informações precisas e atualizadas sobre itinerários e horários das linhas de ônibus;

b) **Sistemas Avançados de Gerenciamento de Tráfego:** os sistemas avançados de gerenciamento de tráfego, ou em inglês *Advanced Transportation Management Systems* (ATMS), têm como objetivo reduzir congestionamentos nas vias urbanas e/ou rurais, garantindo segurança por meio de controle e monitoramento de semáforos e utilização de câmeras de vídeo;

c) **Sistemas Avançados de Informação ao Viajante:** os sistemas avançados de gerenciamento de tráfego, ou em inglês *Advanced Traveler Information Systems* (ATIS), têm como objetivo prover informações ao viajante sobre a via, condições ambientais e trânsito. Fazem uso de sistemas de navegação e informação para garantir segurança ao motorista e para minimizar os congestionamentos;

d) **Operação de Veículos Comerciais:** os sistemas de operação de veículos comerciais, ou em inglês *Commercial Vehicle Operations* (CVO), têm como objetivo principal gerenciar a operação de veículos comerciais. Utilizam tecnologias para melhorar a gerência e o serviço de transporte de cargas com intuito de minimizar as interferências com relação às rotas e tempos perdidos, procurando garantir um alto nível de segurança.

e) **Sistemas Avançados de Controle Veicular:** os sistemas avançados de controle veicular, ou em inglês *Advanced Vehicle Control Systems* (AVCS), têm como objetivo melhorar a segurança viária, permitindo que os veículos auxiliem os motoristas. Para isto, os veículos são equipados com tecnologias que possibilitam ao condutor monitorar as condições de dirigibilidade e tomar medidas necessárias para evitar acidentes.

f) **Sistemas de Transporte Rural Avançados:** os sistemas de transporte rural avançados, ou em inglês *Advanced Rural Transportation Systems* (ARTS), tem como objetivo melhorar a segurança viária em estradas com baixo fluxo de veículos. Acidentes com saída de pista e dispositivos de comunicação para emergência são de interesse particular nesses ambientes.

g) **Gestão Eletrônica de Tráfego e Pedágio:** os sistemas de gestão eletrônica de tráfego e pedágio, ou em inglês *Eletronic Toll and Traffic Management* (ETTM), têm como objetivo prover métodos eficientes para cobrança de pedágio com a finalidade de minimizar o tempo perdido no processo de cobrança e com isto reduzir congestionamentos.

A Tabela 2 apresenta as diversas categorias de Sistemas de Transportes Inteligentes, ressaltando as principais características de cada um deles.

Tabela 2: Características dos principais subsistemas de ITS

		Características
ATMS	Sistemas Avançados de Gerenciamento de Transporte	Gerenciamento de rede, incluindo gerenciamento de incidentes, controle de semáforos, cobrança eletrônica de pedágio, previsão de congestionamento e estratégias de melhoria de congestionamento.
ATIS	Sistemas Avançados de Informação ao Viajante	Informações fornecidas aos viajantes antes e durante a viagem no veículo. O ATMS ajuda a fornecer informações de rede em tempo real
AVCS	Sistemas Avançados de Controle Veicular	Um conjunto de tecnologias projetadas para aprimorar o controle do motorista e a segurança do veículo. Isso varia de acordo com os Sistemas de Rodovia Automatizada (AHS), onde os motoristas cedem todo o controle ao sistema.
CVO	Operação de Veículos Comerciais	Tecnologias para aprimorar a produtividade da frota comercial, incluindo pesagem em movimento (WIM), procedimentos de pré-liberação, registros eletrônicos, coordenação interestadual.
APTS	Sistemas Avançados de Transporte Público	Informações e tecnologias de passageiros para aprimorar as operações do sistema, incluindo cobrança de tarifas, transferências intramodais e intermodais, programação e controle de progresso.
ARTS	Sistemas Avançados de Tranporte Rural	Principalmente tecnologias de segurança e proteção (por exemplo, Socorro) para viagens em áreas pouco povoadas.

Fonte: (Sussman, 2000)

Em uma abordagem mais recente, [Qureshi and Abdullah 2013] afirmam que os Sistemas Avançados de Gerenciamento de Tráfego (ATMS) desempenham um papel central nos Sistemas de Transporte Inteligentes (ITS), aumenta a eficiência geral do transporte, gerando maior fluidez, aumento de segurança, maior mobilidade e incremento na produtividade econômica, fator fundamental para um mercado de ITS. Um *Advanced Traffic Management System* coleta as informações em tempo real de diferentes componentes de hardware, como câmeras, sensores de velocidade, etc., e transmite para o *Transportation Management Center* (TMC), onde são processadas e analisadas. Um sistema de gerenciamento de tráfego pode ser utilizado no tráfego ferroviário, no tráfego rodoviário ou no gerenciamento de tráfego aéreo [Davery 2012].

O Sistema de Transporte Inteligente integra tecnologias de comunicação atuais e crescentes. Devido ao surgimento de muitas tecnologias, o sistema de transporte é capaz de melhorar as condições, a segurança e os serviços de transporte. Como exemplo, pode-se citar aqui duas tecnologias, das mais utilizadas, de detecção de veículos:

- Detecção por laço indutivo

- Detecção de veículo por vídeo

Detecção por laço indutivo: um sistema de detecção de veículo por laço indutivo é um sistema de detecção que usa um laço indutivo com um ímã para induzir uma corrente elétrica em um fio. Este componente é instalado sob o chão e é conectado ao detector veicular. Esses sistemas são capazes de detectar a presença ou a passagem de um veículo, permitindo realizar-se a contagem volumétrica e, até mesmo, a velocidade de cada veículo.

O sistema é composto por três componentes:

1. Laço Indutivo
2. Cabo de Extensão do Laço Indutivo
3. Detector Veicular

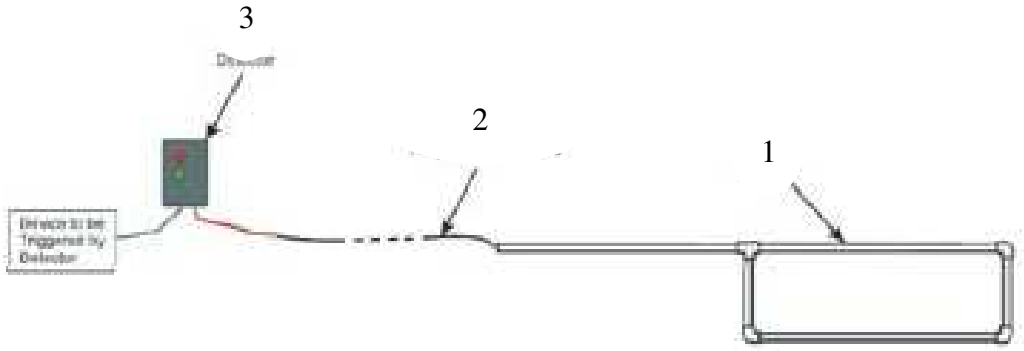


Figura 3: Sistema de Detecção de Veículos por Laço Indutivo [MARSH 2010]

Sistema de detecção de veículos por vídeo (VVDS): a detecção de veículo por vídeo tem se mostrado como uma forma cada vez mais usual de detecção em sistemas de transporte inteligente.

O sistema VVD tem se popularizado como uma alternativa aos sistemas de detecção de veículo por laço indutivo. Basicamente, o sistema é composto por um dispositivo de aquisição de imagens de vídeo (câmera), cabeamento apropriado, unidade de processamento de imagem e software de processamento de imagens (LPR). São sistemas que reduzem custos de instalação e manutenção, além de valer-se dos avanços tecnológicos que permitem uma maior capacidade de processamento de imagens e maior velocidade no envio de dados. [[Yiyan, W. *et al*. 2009].

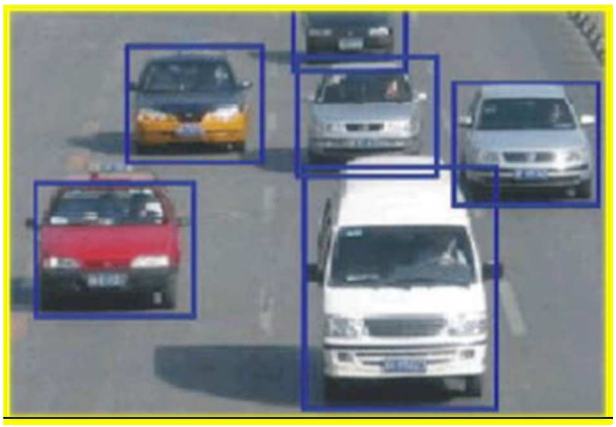


Figura 4: Sistema de Detecção de Veículos por Vídeo [Qureshi and Abdullah 2013]

2.2 Movimento e Trajetórias

Objeto em movimento

Também chamado de objeto móvel ou objeto dinâmico, é definido como uma entidade cuja posição muda com o tempo. Neste trabalho objetos em movimento são conceituados por pontos em movimento (ou seja, como entidades geométricas 0-D). A localização é geralmente indicada usando-se coordenadas geográficas (φ, λ) ou coordenadas cartesianas (x, y, (z)). Assim, no presente trabalho, a localização de um objeto com identificação *id* em movimento no tempo t é especificada por uma tupla[2] de dimensões (*id*, x, y, t).

Parâmetros de Movimento

São quantidades mensuráveis (isto é, primitivas) e seus derivados, como posição, tempo, velocidade, distância, direção. Os parâmetros de movimento são categorizados em dimensões espaciais, temporais, espaço-temporais, conforme descrito na Tabela 3.

[2] Tupla é uma sequência ordenada e finita de elementos. Na terminologia de BD Relacional, uma linha é chamada tupla, um nome de coluna é chamado de atributo e cada tabela é chamada de relação.

Tabela 3: Parâmetros de movimento

Dimensões	Parâmetros		
	Primitivos	**Derivados primários**	**Derivados secundários**
Espacial	Posição (x, y)	Distância $f(posição)$	Distribuição espacial $f(distância)$
		Direção $f(posição)$	Mudança de direção $f(direção)$
		Extensão espacial	Sinuosidade $f(distância)$
Temporal	Instância (t)	Duração $f(t)$	Distribuição temporal
	Intervalo (t)	Tempo de viagem $f(t)$	Mudança de duração $f(duração)$
Espaço-temporal	-	Rapidez $f(x, y, t)$	Aceleração $f(rapidez)$
		Velocidade $f(x, y, t)$	Taxa de aproximação

Modelagem do Movimento e Espaços de Movimento

Modelar movimento significa modelar as entidades em movimento, além de modelar o espaço onde elas se movimentam [Gottfried and Aghajan 2009]. O modelo de dados conceitual escolhido que incorpora o movimento - por exemplo, um espaço euclidiano 2D ou 3D, um espaço de rede ou uma partição espacial - indica como as entidades podem se mover e, consequentemente, afeta as ferramentas de análise necessárias para entender esse movimento. A Figura 5 ilustra os espaços básicos de movimento e como o movimento pode ser modelado nele.

No caso mais simples, o espaço é modelado como um espaço euclidiano 2D (Figura 5 (a)). Entidades são livres para se movimentar e são limitadas apenas por obstáculos em potencial, como, por exemplo, os edifícios b_1 e b_2. O espaço pode ser modelado usando um modelo de dados conceitual de entidade ou campo. O movimento de uma entidade é modelado como um conjunto de locais (x, y) que a entidade visitou ao longo do tempo (em 3D, um local é representado por (x, y, z)). Embora o movimento real possa ser uma curva suave, por motivos de simplicidade, o movimento entre as correções conhecidas é tipicamente modelado como conectores de linha reta (veja e_1). Esse rastreamento de movimento é chamado de trajetória. Formalmente, uma trajetória é definida como uma sequência de locais com registro de data e hora $(x, y)_{T_1}, \dots, (x, y)_{T_t}$, onde $T_1, \dots, T_t$ são t etapas consecutivas. Conectar os locais de uma trajetória em ordem temporal produz uma linha poligonal que pode se auto interceptar [Gudmundsson *at al.* 2007].

Uma maneira visualmente elegante de incorporar o tempo na representação do movimento 2D é o uso de um espaço 3D. Neste caso, as duas dimensões espaciais x e y são combinadas com um eixo temporal ortogonal t (Figura 5 (b)). Esse cubo espaço-tempo (às vezes chamado de aquário espaço-temporal) sustenta a chamada geografia do tempo [Hägerstrand 1970]. Além disso, esse conceito permite a modelagem de mobilidade potencial, que é onde uma entidade poderia estar entre duas correções conhecidas. O prisma de espaço-tempo ilustrado por e_3 indica como um volume em que e_3 poderia estar entre t_1 e t_2, dada a restrição de uma velocidade máxima [Miller 1991].

A maioria dos movimentos humanos é restrita a alguma forma de rede. Por exemplo, quando rastreando o movimento de um viajante em uma área urbana, o

movimento é mais adequadamente modelado como a progressão das bordas e nós visitados da rede de transporte público [Schmid *at al*, 2009]. A Figura 5 (c) mostra o movimento das entidades e_{11} e e_{12}, usando uma rede de trânsito com ruas e linhas de trem. Mesmo quando o movimento real é capturado com um dispositivo de rastreamento GPS e se desvia da rede, é semanticamente dado que o movimento está vinculado à rede. A correspondência de mapa é a etapa de pré-processamento que combina as correções imprecisas com as respectivas arestas e nós [Bernstein and Kornhauser 1996].

Por fim, ao rastrear um telefone celular, o rastreamento de movimento é capturado apenas nas ERBs das torres de telefonia celular às quais o telefone estava conectado [Du Mouza and Rigaux 2005]. Por exemplo, na Figura 5 (d), a entidade e_{12} foi conectada às torres de telefone celular A, E, F, C e B. Mesmo que essa forma de informação de movimento não ofereça localizações precisas de pontos, em muitos contextos de aplicação, como rastreamentos vagos é suficiente para a tarefa em questão (por exemplo, uma consulta de proximidade em um serviço baseado em localização).

Todos os quatro espaços de movimento podem acomodar o movimento de indivíduos, mas eles são tão diferentes que precisam de diferentes ferramentas e técnicas de análise de movimento.

No caso deste trabalho, o modelo que se apresenta como o mais apropriado para modelar o movimento de veículos por rodovias e é o apresentado na Figura 5 (c).

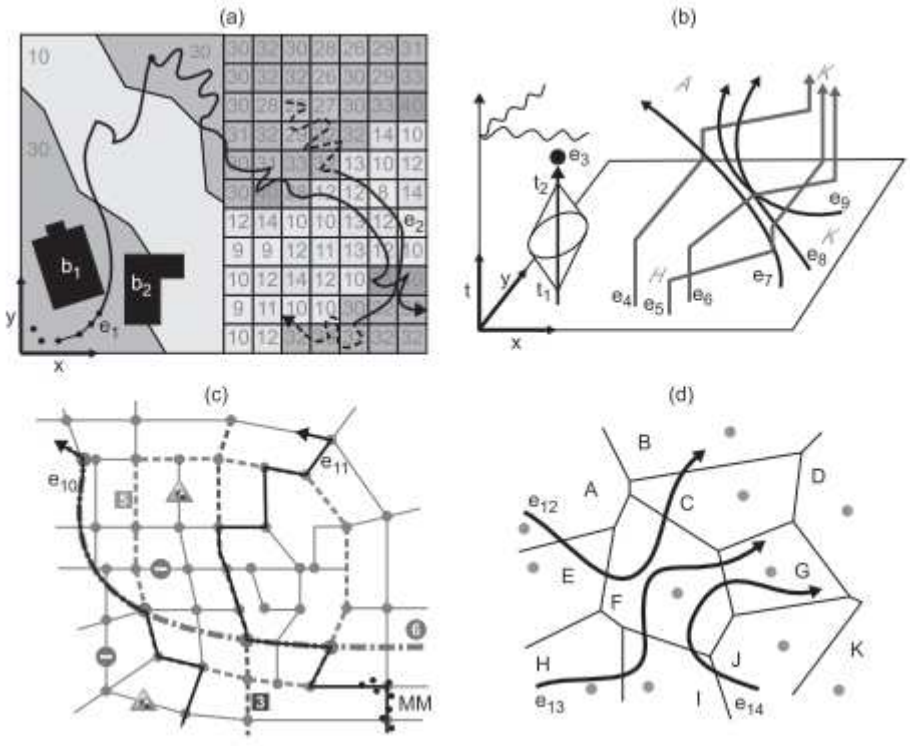


Figura 5: Quatro espaços básicos de movimento: (a) Espaço Euclidiano 2D, (b) Cubo Espacial 3D, (c) espaço na rede, (d) pavimentação irregular, por exemplo, células da torre telefônica Fonte: GUDMUNDSSON *at al.*, 2012

Trajetória

É o percurso realizado por um dado objeto através do espaço em função do tempo. Na matemática discreta, trajetória é uma sequência $(f^k(x))_{k\in N}$ de valores calculados pela iterada aplicação do mapeamento de f para um elemento x. [Santos, 2011].

O movimento de um objeto pode ser representado por sua trajetória como um conjunto de posições ordenadas pelo tempo [Laube *et al.* 2007]; [Spaccapietra *et al.* 2008], como mostra a Figura 6. Consiste em uma série de observações discretas no espaço-tempo.

Assim, considerando-se um espaço bidimensional, a componente posicional do objeto móvel na trajetória pode ser representada pelas coordenadas (x, y). O tempo, o qual pode ser representado por t, indica o momento de cada posição do objeto na trajetória. Tomando id como a identificação do objeto que está em movimento, podemos definir trajetória como uma sequência de pontos, cada um dos quais representados pela tupla (id, x, y, t).

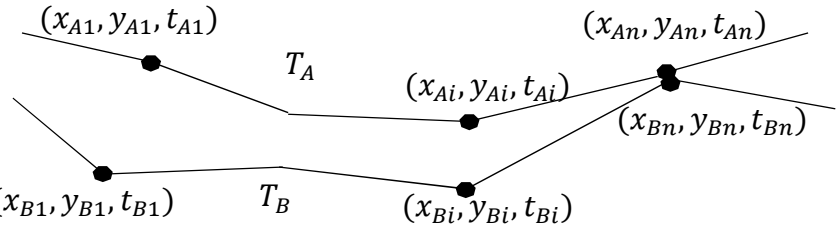


Figura 6: Representação simplificada das trajetórias de veículos em rodovias Fonte: Autor

Dados da trajetória

Com base na tecnologia empregada para a sua captação e registro, os dados de mobilidade podem estar disponíveis em diferentes formas. [Spinsanti *et al.* 2013] diferenciaram os dados de trajetórias a partir de tecnologias como GPS (Global Position Systems), GSM (Global System for Mobile Communications) e os dados geosociais. Já Pelekise and Theodoris (2014) citaram outras formas de obtenção de dados de trajetórias como RFID (Radio Frequency Identification) e dados baseados em Redes Wi-Fi.

Os dados baseados em GPS são compostos de sequências ordenadas temporalmente de coordenadas geográficas registradas por um dispositivo com o GPS habilitado, transportado pelo objeto em movimento.

Os dados baseados na tecnologia GSM são compostos por sequências ordenadas temporalmente de identificadores das ERBs[3] (Estações Rádio Base) pelas quais o objeto em movimento passa. Dados baseados em redes geo-sociais são conteúdos encontrados

[3] Estações Rádio Base ou ERBs são equipamentos que fazem a conexão entre os telefones celulares e a companhia telefônica, ou mais precisamente a Central de Comutação e Controle.

nas redes sociais da Internet e às quais coordenadas geográficas foram anexadas (fotos, check-in).

Os dados baseados em RFID contêm uma sequência de identificadores de leitores RFID através dos quais o objeto em movimento passou, ao passo que os dados baseados em Wi-Fi contêm uma sequência de identificadores de pontos de acesso que se comunicaram como objeto em movimento.

Embora as propriedades dessas formas de dados, por exemplo, sua precisão, sejam muito diferentes [Spinsanti *et al*. 2013] [Pelekise and Theodoris 2014], eles foram usados para abordar assuntos similares ou relacionados, usando métodos de mineração semelhantes ou relacionados.

Dados espaço-temporais

Quaisquer informações relacionadas a espaço e tempo. Podem ser representados por um conjunto de pontos localizados no tempo e no espaço. Esses dados são provenientes de dispositivos móveis, equipados ou não com sistemas GPS (*Global Positioning System*) e, atualmente, são gerados em grande volume [Miranda & Abreu 2018]. Esta abordagem considera especificamente os dados que envolvem objetos pontuais que se movem ao longo do tempo. Os termos *entidade* e *trajetória* se referem a tal objeto pontual e à representação de seu movimento, respectivamente. Os padrões de movimento em tais dados referem-se a eventos e episódios expressos por um conjunto de entidades [Gudmundsson, J. *at al* 2008] (Figura 7).

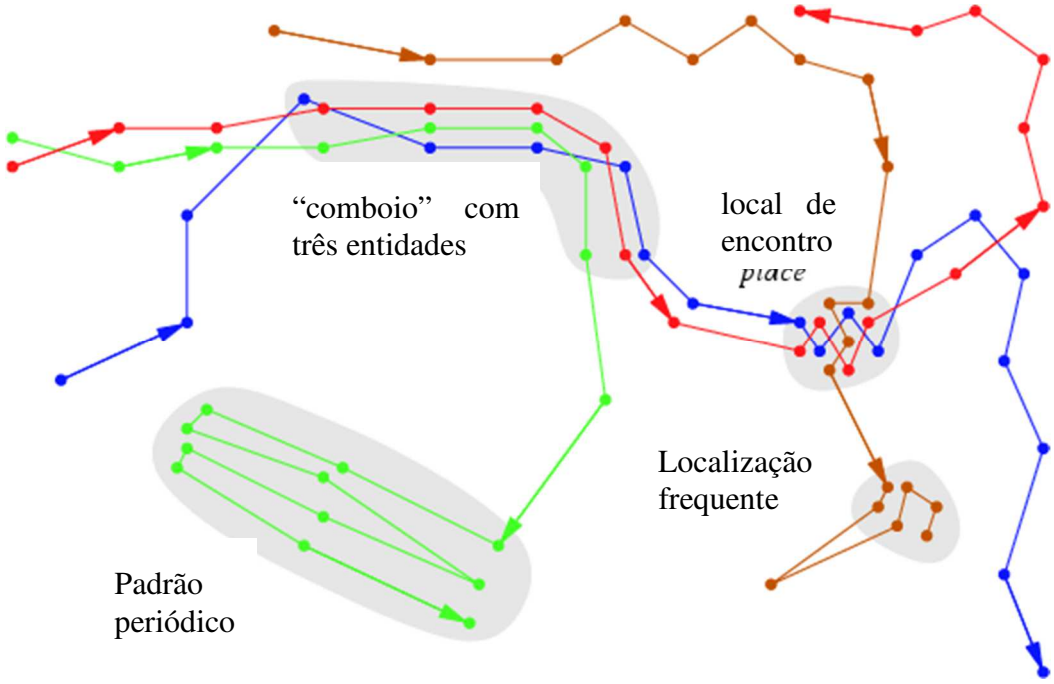


Figura 7: Ilustrando as trajetórias de quatro entidades em movimento em mais de 20 etapas temporais Fonte: Adaptado de GUDMUNDSSON *at al*., 2012

Análise de dados de trajetórias

Existem diversos trabalhos na literatura que analisam dados brutos de trajetórias, buscando extrair padrões de comportamento de grupos de trajetórias. Um padrão é um comportamento que se repete dentro da mesma trajetória ou entre trajetórias diferentes. [Bogorny e Baz, 2012]

O trabalho de Laube (2002) foi um dos pioneiros na área de análise de comportamento de dados espaço-temporais gerados por dispositivos móveis e tem sido a base de muitas outras pesquisas na área. Nele define-se um padrão para trajetórias que possuem comportamento similar, analisando a mudança de direção. Um padrão deve conter um número mínimo de trajetórias que se movimentam na mesma direção. Nesse estudo os aspectos temporais não são considerados.

Em 2005, Laube *at al.* (2005) propuseram um novo tipo de padrão de trajetórias, analisando, além da direção do movimento, a região/localização onde o determinado movimento ocorreu. Definiram-se os quatro tipos mais conhecidos de padrões de trajetórias, os quais são chamados de padrões geométricos: *flock, leadership, convergence e encounter.* Na Figura 8 há um exemplo de cada um desses padrões. O primeiro refere-se a um grupo de objetos que se movem na mesma direção e suas trajetórias estão próximas umas das outras (exemplo: um bando de aves). O padrão *leadership* caracteriza um conjunto de objetos que se movem na mesma direção, suas trajetórias estão próximas umas das outras e um dos objetos lidera o grupo, estando à frente dos demais. O padrão *convergence* (Figura 8, direita) diz respeito a um grupo de objetos que se movem em direção ao mesmo local (exemplo: estudantes chegando ao campus universitário ou usuários acessando o mesmo site). O padrão *encounter* caracteriza um grupo de objetos móveis que se deslocam para o mesmo local e ao mesmo tempo.

Diversos outros padrões podem ser encontrados na literatura.

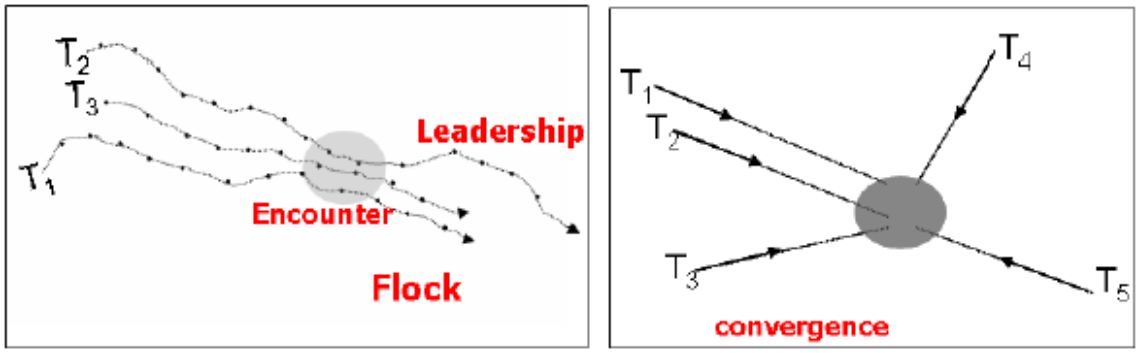


Figura 8 – Exemplos de padrões de comportamento em grupos de trajetórias Fonte: Bogorny e Braz, 2012

Modelagem de dados de trajetória

Muitos dos modelos espaço-temporais existentes podem ser usados para modelar trajetórias. Para se modelar e gerenciar incerteza da trajetória adequadamente, é possível utilizar-se diferentes métodos de interpolação. O traçado de um objeto em movimento no espaço geográfico é contínuo enquanto uma trajetória é apenas uma amostra dos pontos de localização pelos quais o objeto em movimento passa [Feng and Zhu 2016]

As tarefas de mineração de dados mais importantes e frequentemente usadas [Feng and Zhu 2016] são: classificação, *clustering*, padrão mineração, descoberta de conhecimento, mineração frequente de padrões e assim por diante.

Trajetória Bruta

É a representação mais simples da trajetória. Uma trajetória bruta pode ser representada por meio de um conjunto de pontos (T_{id}, x, y, t), onde T_{id} é o identificador da trajetória e (x, y), uma coordenada geográfica que corresponde a um lugar no espaço num instante de tempo t. A Figura 9 (esquerda) traz um exemplo de trajetória bruta, em que um enorme número de pontos está associado a uma única trajetória.

Trajetória Semântica

Trajetórias brutas têm sido enriquecidas com informações através de aplicativos de redes sociais, como o nome do local visitado por um objeto, denominado Ponto de Interesse (POI), e o tempo que o indivíduo permaneceu em cada POI. Quando trajetórias são associadas a informações contextuais, elas são chamadas trajetórias semânticas.

Parent *et al*. (2013) propuseram novas ideias e técnicas relacionadas à elaboração e análise de trajetórias semânticas. Sugerem que dados brutos da trajetória devem ser combinados com dados contextuais.

Em outras palavras, o movimento é uma sequência de paradas (locais visitados) e movimentos (pontos espaço-temporais entre as paradas) [Spaccapietra *et al*. 2008]. Um exemplo é mostrado na Figura 9 (direita). Com a nova representação de trajetórias, o movimento se torna um tipo de dados mais complexo, tendo agora mais dimensões a serem consideradas: espaço, tempo e semântica.

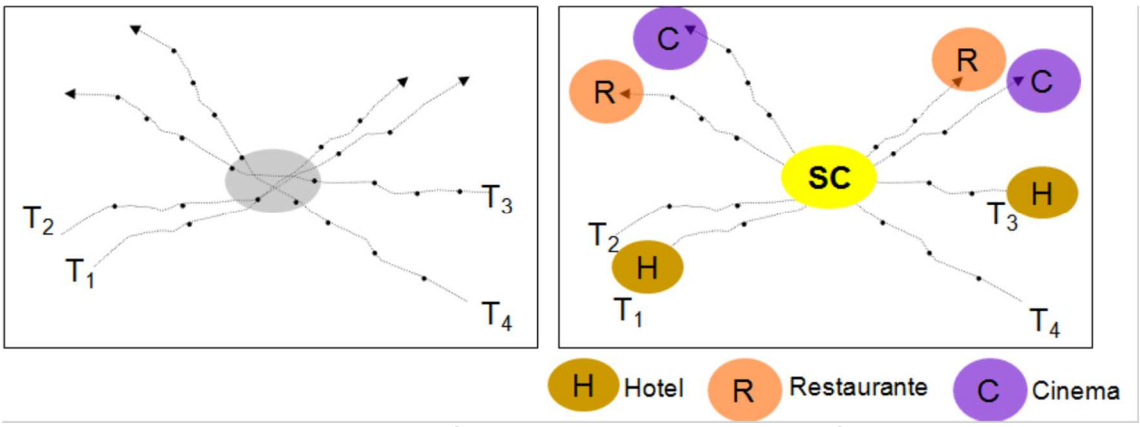


Figura 9 – Representações de trajetórias Brutas (esquerda) e Semânticas (direita) Fonte: Bogorny e Braz, 2012

2.3 Clustering de Trajetórias

Cluster

Um cluster de dados de trajetórias é útil para se agrupar trajetórias baseadas em padrões de movimento semelhantes. Os objetos em movimento são agrupados com base em medidas de movimento de similaridade, como velocidade de movimento, direção do movimento, distância, dispersão espacial, duração temporal e sintaxe e significado semântico das localizações. Encontrar apenas locais importantes de uma trajetória é um exemplo de agrupamento. *Clustering* baseado em mobilidade é menos sensível do que um clustering baseado em densidade ou tamanho de conjunto de dados de trajetória [Feng and Zhu 2016].

Clustering

É uma análise multivariada usada para agrupar objetos semelhantes (próximos em termos de distância) e no mesmo grupo (cluster). Ao contrário dos métodos de aprendizado supervisionado (por exemplo, classificação e regressão), uma análise de agrupamento não usa nenhuma informação de rótulo, mas simplesmente usa a similaridade entre recursos de dados para agrupá-los em clusters. Agrupar o conteúdo de um conjunto de dados é uma tarefa realmente útil para classificar e rotular os dados de acordo com certos parâmetros semelhantes entre si.

O *clustering* não se refere a algoritmos específicos, mas é um processo para criar grupos com base na medida de similaridade. A análise de agrupamento usa algoritmo de aprendizagem não supervisionado para criar clusters.

Os algoritmos de agrupamento geralmente funcionam com base no princípio simples de maximização de semelhanças *intracluster* e minimização de semelhanças *intercluster*. A medida de similaridade determina como os clusters precisam ser formados.

Similaridade é uma caracterização da razão do número de atributos que dois objetos compartilham em comum em comparação com a lista total de atributos entre eles. Os objetos que têm tudo em comum são idênticos e têm uma semelhança de 1,0. Objetos que não têm nada em comum têm uma similaridade de 0,0.

O modelo de clustering é uma noção usada para significar que tipo de clusters estamos tentando identificar. Os quatro modelos mais comuns de métodos de *clustering* são *clustering* hierárquico, *clustering k-means, clustering* baseado em modelo e *clustering* baseado em densidade:

- ***Clustering* hierárquico**: Ele cria uma hierarquia de clusters e apresenta a hierarquia em um dendrograma. Este método não requer que o número de clusters seja especificado no início. Conectividade de distância entre observações é a medida.

- **Clustering *K-means***: Ele também é conhecido como *clustering* simples. Ao contrário do *clustering* hierárquico, ele não cria uma hierarquia de clusters e requer o número de clusters como entrada. No entanto, seu desempenho é mais

rápido do que o *clustering* hierárquico. Distância do valor médio de cada observação / cluster é a medida.

- ***Clustering* baseado em modelo** (ou *modelos de distribuição*): Tanto o *clustering* hierárquico quanto o *clustering* k-means usam uma abordagem heurística para construir clusters e não dependem de um modelo formal. O agrupamento baseado em modelo assume um modelo de dados e aplica um algoritmo EM para localizar os componentes do modelo mais prováveis e o número de clusters. A significância da distribuição estatística das variáveis no conjunto de dados é a medida.

- ***Clustering* baseado em densidade**: Ele constrói clusters em relação à medição de densidade. Os clusters neste método têm uma densidade mais alta do que o restante do conjunto de dados. A densidade no espaço de dados é a medida.

Um bom algoritmo de agrupamento pode ser avaliado com base em dois objetivos principais: alta similaridade intra-classe e baixa similaridade entre classes.

O agrupamento baseado em densidade usa a ideia de alcançabilidade de densidade e conectividade de densidade (como uma alternativa para medição de distância), o que o torna muito útil na descoberta de um agrupamento em formas não lineares. Este método encontra uma área com uma densidade maior do que a área restante. Um dos métodos mais famosos é o *Density-based spatial clustering of applications with noise* (*DBSCAN*). Ele usa o conceito de acessibilidade de densidade e conectividade de densidade.

DBSCAN

O algoritmo de agrupamento baseado em densidade DBSCAN é uma técnica de agrupamento de dados fundamental para encontrar clusters de forma arbitrária, bem como para detectar outliers.

Ao contrário do K-*means*, DBSCAN não requer o número de clusters como parâmetro. Em vez disso, infere o número de clusters com base nos dados e pode descobrir clusters de forma arbitrária (para comparação, o K-*means* geralmente descobre clusters esféricos). Os métodos de particionamento (K-*means*, agrupamento PAM) e agrupamento hierárquico são adequados para localizar *clusters* de formato esférico ou *clusters* convexos. Em outras palavras, eles funcionam bem para *clusters* compactos e bem separados. Além disso, eles também são severamente afetados pela presença de ruído e *outliers* nos dados.

As **vantagens** do agrupamento baseado em densidade são:

1. Nenhuma suposição sobre o número de clusters. O número de clusters geralmente é desconhecido *a priori*. Além disso, em um fluxo de dados em evolução, o número de clusters naturais costuma mudar.

2. Descoberta de *clusters* com forma arbitrária. Isso é muito importante para muitos aplicativos de fluxo de dados.

3. Capacidade de lidar com *outliers* (resistente a ruído).

As **desvantagens** do agrupamento baseado em densidade são

1. Se houver variação na densidade, os pontos de ruído não são detectados
2. Sensível aos parâmetros, ou seja, difícil de determinar o conjunto correto de parâmetros.
3. A qualidade do DBSCAN depende da medida da distância.
4. DBSCAN não consegue agrupar bem conjuntos de dados com grandes diferenças em densidades.

Este algoritmo funciona em uma abordagem paramétrica. Os dois parâmetros envolvidos neste algoritmo são:

- *e* (*eps*) é o raio das vizinhanças em torno de um ponto de dados *p* .
- *minPts* é o número mínimo de pontos de dados que queremos em uma vizinhança para definir um *cluster*.

Uma limitação do DBSCAN é que ele é sensível à escolha de *e*, principalmente se os clusters têm densidades diferentes. Se *e* for muito pequeno, os clusters mais esparsos serão definidos como ruído. Se *e* for muito grande, os clusters mais densos podem ser mesclados.

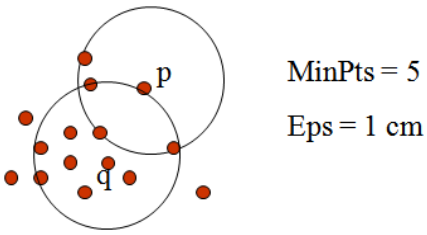


Figura 10 – Parâmetros do Algoritmo DBSCAN

Uma vez que esses parâmetros são definidos, o algoritmo divide os pontos de dados em três pontos:

- *Pontos principais*: Um ponto *p* é um ponto central se, pelo menos, os pontos *minPts* estiverem dentro da distância *l* (*l* é o raio máximo da vizinhança de *p*) dele (incluindo *p*).
- *Pontos de fronteira:* Um ponto *q* é fronteira de *p* se existe um caminho p_1 , ... , p_n com $p_1 = p$ e $p_n = Q$, em que cada p_{i+1} é diretamente acessível a partir de p_i (todos os pontos da caminho deve ser pontos centrais, com a possível exceção de q).
- *Outliers*. Todos os pontos não alcançáveis de qualquer outro ponto são *outliers*.

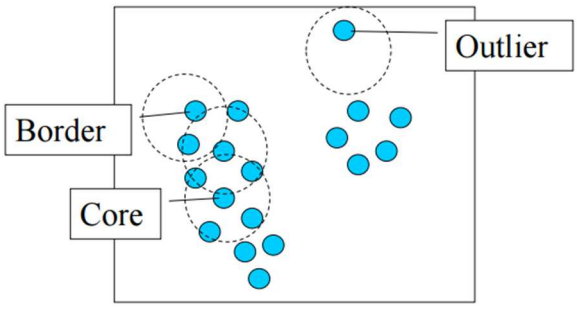


Figura 11 – Núcleo, fronteira e *outlier*

As etapas no *DBSCAN* são simples após definir as etapas anteriores:

a) Escolha aleatoriamente um ponto que não está atribuído a um cluster e calcule sua vizinhança. Se, na vizinhança, este ponto tiver *minPts*, faça um cluster em torno dele; caso contrário, marque-o como outlier.

b) Depois de encontrar todos os pontos principais, comece a expandi-los para incluir os pontos de fronteira.

c) Repita essas etapas até que todos os pontos sejam atribuídos a um cluster ou a um *outlier*.

Análise de silhueta

A análise de silhueta permite que se calcule quão similar cada observação é com o cluster a que está atribuído em relação a outros clusters. Essa métrica (largura da silhueta) varia de -1 a 1 para cada observação em seus dados e pode ser interpretada da seguinte forma:

- Valores próximos de 1 sugerem que a observação é bem combinada com o cluster atribuído

- Valores próximos de 0 sugerem que a observação está no limite de correspondência entre dois clusters

- Valores próximos de -1 sugerem que as observações podem ser atribuídas ao cluster errado

3. Trabalhos Relacionados

A literatura científica que trata sobre classificação de trajetória nos mais diversos domínios de aplicação apresenta, em sua grande maioria, pesquisas utilizando-se bases de dados produzidos a partir de dispositivos embarcados em objetos móveis, como GPS (Global Position System) ou RFID (Radio-Frequency IDentification). Poucos trabalhos têm sido desenvolvidos visando à classificação de trajetórias. A grande maioria deles resume-se à extração de características relevantes para o domínio considerado, especializando-se em resolver um problema específico.

Sistemas Inteligente de Transportes (ITS), mais especificamente os Sistema de detecção de veículos por vídeo (VVDS), geram dados espaço-temporais onde é possível extrair-se trajetórias percorridas por veículos ao longo de rodovias por onde encontram-se instalados seus sensores. Algumas características desses sistemas exigem que seus dados sejam analisados partindo-se de algumas premissas que contemplem suas particularidades, permitindo que os modelos adotados possam extrair o conhecimento em suas bases dados.

Sistema de detecção de veículos por vídeo, que captam imagens de placas de veículos e às transformam em dados de passagem através de softwares de reconhecimento de placas (LPR), podem gerar trajetórias incompletas ou fragmentos de trajetórias. Isso se dá porque os dispositivos de detecção de veículos por vídeo encontram-se instalados de forma não homogênea em pontos de rodovias, dependendo de ações de implementações de órgão governamentais e privados, podendo estar concentrados em determinados regiões em detrimento de outras. Além disso, podem concentrar-se em determinados tipos de vias ou regiões, mostrando-se raros ou ausentes em determinadas localidades. Vale ainda ressaltar que esses sistemas geram informações com ruído ou falha de leitura, o que demanda um processo de limpeza e correção dos dados na fase de preparação, visando minimizar os impactos desses *outliers* ou ruídos nos dados.

Apresentamos aqui alguns trabalhos desenvolvidos fundamentados em técnicas de clusterização espacial, empregadas em alguns domínios de aplicação, destacando-se o escopo de cada uma das publicações.

No artigo *Trajectory Data Mining: An Overview* (ZHENG, Yu., 2015) o autor realiza um levantamento sistemático das principais pesquisas em mineração de dados de trajetórias, fornecendo um panorama da área, além do escopo de seus tópicos de pesquisa. Apresenta ainda um roteiro da derivação de dados de trajetória, pré-processamento de dados de trajetória, gerenciamento de dados de trajetória e uma variedade de tarefas de mineração (como mineração de padrão de trajetória, detecção de outlier e classificação de trajetória). Além disso, a pesquisa explora as conexões, correlações e diferenças entre essas técnicas existentes. Um arcabouço sobre *datamining* de trajetórias, apresentando seus diferentes estágios e explorando suas características e suas diferenças..

No artigo *A Survey on Trajectory Data Mining* (Tanuja V. & Govindarajulu P, 2016) apresentam uma revisão sobre as pesquisas que estão sendo feitas em mineração de dados de trajetória, onde levantou-se a literatura sobre localização e mineração de trajetória baseada na comunidade, bases de dados de trajetória e consulta de trajetória. A revisão ainda explora a estrutura de mineração de dados de trajetória. Os autores apresentam um largo escopo sobre as tendências atuais sobre mineração de dados de trajetórias e suas potenciais aplicações para organizações governamentais e privadas no sentido de reduzir o custo de supervisão e gerenciamento.

O artigo *Survey on Trajectory Clustering Analysis* (BIAN, Jiang *et al*., 2018) analisa de forma abrangente o desenvolvimento do agrupamento de trajetórias. Considerando o papel crítico da mineração de dados de trajetória em sistemas inteligentes modernos para segurança de vigilância, detecção de comportamento anormal, análise de comportamento de multidão e controle de tráfego, o agrupamento de trajetória tem atraído atenção crescente. Citam os autores que apesar de atingir certo nível de desenvolvimento, o sucesso do *clustering* de trajetórias é limitado por condições complexas, como cenários de aplicação e dimensões de dados. Este artigo ainda apresenta

uma análise abrangente de métodos representativos e direções futuras promissoras. O trabalho aborda métodos de agrupamento de trajetórias e os classifica em três categorias: algoritmos não supervisionados, supervisionados e semi supervisionados. Os autores concluem que algoritmos não supervisionados têm as desvantagens de alto custo de computação e alta carga de memória, embora não haja necessidade de dados de treinamento e supervisão de especialistas humanos. Já os algoritmos semi supervisionados combinam as vantagens de algoritmos supervisionados e não supervisionados e podem resultar em métodos mais eficientes.

Em *An Extended K-means Technique for Clustering Moving Objects* (Ossama *et al*., 2011) os autores apresentam um novo algoritmo de agrupamento baseado em padrões que amplia o algoritmo k-means para agrupar dados de trajetória de objetos em movimento. O algoritmo propõe o uso da direção como heurística para determinar o número diferente de clusters para o algoritmo k-means. Também aborda o uso do coeficiente silhueta do método proposto. Este artigo propõe uma extensão do algoritmo K-*means* para dados de trajetórias, o E-KM, ou K-*means* Estendido. Compara com os resultados utilizando-se os algoritmo K-means, mostrando que supera as desvantagens conhecidas desse algoritmo como a dependência do número de clusters (k) e a dependência da escolha inicial dos centróides dos clusters, garantindo nessa abordagem a criação de clusters de alta precisão.

(Khaing, H. & Thein, T., 2014) propõem em seu artigo An Efficient Clustering Algorithm for Moving Object Trajectories um algoritmo de agrupamento baseado em Clustering Espacial Baseado em Densidade de Aplicações com Ruído (DBSCAN), apresentando uma variação deste método capaz de agrupar bem conjuntos de dados com grandes diferenças de densidade. Os resultados da avaliação mostram que o algoritmo de agrupamento proposto pode fornecer melhor desempenho e erro mínimo do que o DBSCAN. O experimento apresenta um estuda sobre o efeito da alteração do tamanho dos dados entre a trajetória no tempo de computação de agrupamento para DBSCAN e algoritmo proposto. Os autores demostram que o algoritmo de agrupamento proposto funciona bem para grandes conjuntos de dados, sendo que o algoritmo DBSCAN leva mais tempo para agrupar todos os objetos.

O presente trabalho propõe o uso do algoritmo DBSCAN como forma de redução de ruídos e extração de trajetórias mais frequentes percorridas por veículos transportando ilícitos a partir dos pontos onde se apresenta a maior densidade de registros de passagens por veículos envolvidos em ocorrências de apreensão em cada grupo de ilícito estudado.

4. Trabalhos Desenvolvidos

4.1 Contextualização do Problema

Na atividade de policiamento e fiscalização, realizar abordagens a veículos, basicamente, pode ser considerado como um evento aleatório onde policiais abordam veículos por amostragem. Porém, a experiência (ou conhecimento do negócio) do policial envolvido na seleção de veículos para abordagens pode contribuir para o aumento da assertividade em busca de alguma irregularidade, dentro do seu universo de abordagens, baseando-se em sua experiência profissional. Frente ao grande volume de veículos que circulam por rodovias, mesmo com o *feeling* policial, realizar abordagens dessa forma é

algo que praticamente depende de sorte, onde o agente pouco pode influenciar nos resultados.

A Polícia Rodoviária Federal, a partir de 2014, vem desenvolvendo o Projeto Alerta Brasil, um Sistema de Transporte Inteligente, gerido pela PRF, voltado à segurança viária em rodovias federais do país. Consiste em um sistema formado por equipamentos que coletam dados de passagens de veículos em vários pontos de rodovias federais, alimentando uma base de dados estruturada.

Para o registro de ocorrências em rodovias federais há também uma base de dados que armazena os Boletins de Ocorrências Policiais (BOP), onde encontram-se dados sobre as ocorrências policiais registradas pela corporação, que contém informações sobre pessoas e veículos envolvidos em cada ocorrência, além dos crimes que, em tese, os participantes da ocorrência estariam envolvidos.

A partir da base de dados gerada pelo Sistema Alerta Brasil da PRF, que regista a passagens de veículos em rodovias, é possível reconstruir trajetórias percorridas por veículos que circulam por rodovias.

4.1.1 Entendimento dos Dados

O ITS Alerta Brasil é capaz de gerar um conjunto de registros espaço-temporais dado por $T_A = [(x_1, y_1, t_1), (x_2, y_2, t_2), \dots, (x_n, y_n, t_n)]$ e que representa a trajetória percorrida pelo viajante utilizando o veículo com identificação id_A, sendo *n* o número de registros de passagem do veículo que compõe a trajetória percorrida por ele (ver Tabela 4).

Tabela 4: Registros de passagens de veículos pelos pontos (x_i, y_i) cada instante t_i

id veículo	*latitude*	*longitude*	*datetime*
id_A	x_1	y_1	t_1
id_A	x_2	y_2	t_2
id_A	x_3	y_3	t_3
...	...	...	...
id_A	x_n	y_n	t_n

A base de dados BOP (Boletins de Ocorrências Policiais) é formada por um conjunto de dados com informações sobre as ocorrências criminais envolvendo pessoas e veículos. Na Tabela 5 podemos ver um exemplo de representação de um *dataset* extraído da base BOP, onde podemos observar informações relacionadas a cada uma das ocorrências de apreensão registradas.

Tabela 5: Registros ocorrências policiais (BOP)

id veículo	*Tipo de veículo*	*latitude*	*longitude*	*Enquadramento*	*Apreensão*	*datetime*
id_A	automóvel	x_1	y_1	Tráfico de Drogas	Maconha	t_1
id_B	caminhão	x_2	y_2	Contrabando	Cigarros	t_2
id_C	ônibus	x_3	y_3	Tráfico de Armas	Armas	t_3
...	...	...	...			...
id_N	automóvel	x_n	y_n	Tráfico de Drogas	Cocaína	t_n

4.1.2 Principais Desafios

A granularidade pode implicar diretamente no volume de dados armazenados, na velocidade das consultas e no nível de detalhamento das informações do *Data Warehouse*.

Essa dificuldade pode ser observada em um sistema de detecção de veículos por vídeo (VVDS), onde tem-se uma rede em constante evolução e com frequente incremento de novos pontos. Em razão disso, para regiões ou períodos distintos podemos ter uma rede com densidades diferentes.

Ainda sobre a densidade, é relevante destacar que os pontos de registro de passagens de veículos do sistema não estão homogeneamente distribuídos pela malha rodoviária da região estudada. Com isso, algumas trajetórias podem apresentar uma escassez de pontos de registro.

A base de dados contendo as passagens de veículos já possui mais de 8,010 bilhões de registros, o que deve ser considerado ao se implementar algoritmos de maior complexidade.

Uma questão importante é a característica cíclica das rotas rodoviárias, onde os viajantes sempre tendem a retornar ao seu local de origem ou de residência, devendo-se adotar um critério a ser considerado na escolha de pontos de início e fim das trajetórias para que elas possam ser comparadas, com origem e destino em regiões de interesse *(ROIs).*

A maioria dos métodos de mineração de dados aplicados em trajetórias foi desenvolvida na era da escassez de dados e / ou em um domínio diferente. Como os dados de trajetória são inerentemente volumosos e estão cada vez mais disponíveis, trabalhos em mineração de dados de trajetórias devem ser escaláveis para lidar com dados massivos. Isso exige, entre outros, métodos de indexação e agregação mais eficientes que obtenham uma alta redução de dados e minimizem a perda de informações. [Mazimpaka and Timpf 2016]

Embora os pesquisadores dos métodos de mineração de dados de trajetória tenham estudado principalmente a trajetória geométrica sem considerar o contexto da aplicação, esforços recentes de colaboração [Demsar *at al*. 2015] demonstraram que os melhores resultados com relevância na vida real da mineração de dados de trajetória podem ser alcançado por uma colaboração entre especialistas em métodos e especialistas em

domínio de aplicativos. Essa colaboração facilita o acesso a dados relevantes para especialistas em métodos.

Outra questão é que embora trajetórias de animais sejam facilmente disponibilizadas abertamente, o rastreamento humano está associado a sérios problemas de privacidade, tornando os dados de trajetória humana de alta qualidade dificilmente acessíveis aos pesquisadores de métodos de mineração. Apesar de algum trabalho realizado para resolver esses problemas (por exemplo, anonimização [Abula *at al*. 2008] and [Gidofalvi *at al*. 2007]), eles prevalecem. Um grande desafio é que trabalhos sobre mineração de dados de trajetória devam desenvolver estruturas de mineração mais convincentes para equilibrar a extração de conhecimento útil e preservar a privacidade dos indivíduos rastreados.

Para amenizar essa questão é indicado que se trabalhe com dados anonimizados ou pseudo-anonimizados. A Lei Geral de Proteção de Dados Pessoais [LGPD 2018] define o dado anonimizado, como aquele que, originariamente, era relativo a uma pessoa, mas que passou por etapas que garantiram a desvinculação dele a essa pessoa.

Se um dado for anonimizado, então a LGPD não se aplicará a ele. Vale frisar que um dado só é considerado efetivamente anonimizado se não permitir que, via meios técnicos e outros, se reconstrua o caminho para "descobrir" quem era a pessoa titular do dado - se de alguma forma a identificação ocorrer, então ele não é, de fato, um dado anonimizado e sim, apenas, um dado pseudonimizado e estará, então, sujeito à LGPD.

Já a pseudoanonimização consiste num mecanismo de disfarce da identidade, substituindo-se um atributo por outro. Nele dados pessoais são tratados de forma a não poderem mais ser atribuídos ao respectivo titular sem recorrer à outras informações a ele correlatas. Sendo assim, tais informações suplementares são mantidas separadamente e sujeitas a medidas técnicas e organizativas para assegurar a desvinculação do dado pessoal ao seu titular [MACHADO e DONEDA, 2018].

Segundo especialistas, “dados anonimizados são essenciais para o crescimento da inteligência artificial, da internet das coisas, do aprendizado das máquinas, das cidades Inteligentes, da análise de comportamentos, entre outros. Eles indicam ainda que, sempre que possível, uma organização, pública ou privada, realize a anonimização de dados pessoais, pois isso aperfeiçoa a segurança da informação na organização e gera, assim, mais confiança em seus serviços e para seus públicos.” [4]

Feng and Zhu (2016) citam que são necessárias boas técnicas de preservação da privacidade para se fornecer medidas de segurança mais eficazes aos bancos de dados de trajetória. Às vezes, pode ser necessário combinar dados de trajetória com outros métodos de mineração de dados para se obter melhores resultados.

4.2 Preparação dos Dados

Embora construir modelos seja empolgante, os modelos de alto desempenho requerem dados de alta qualidade como entrada. Se alimentarmos um modelo com dados

[4] Fonte: https://www.serpro.gov.br/lgpd/menu/protecao-de-dados/dados-anonimizados-lgpd

de entrada de baixa qualidade, serão obtidos resultados ruins como saída. Mesmo os métodos de aprendizado de máquina mais sofisticados terão baixo desempenho se treinados com os dados errados, porque os modelos aprendem exclusivamente com os dados de treinamento (e não com os dados que gostaríamos que eles tivessem). Portanto, é necessário dispender-se tempo para garantir que os dados de treinamento sejam consistentes e sem erros.

Para se realizar uma análise de cluster, geralmente, os dados devem ser preparados da seguinte forma:

- As linhas são observações (indivíduos) e as colunas são variáveis

- Qualquer valor ausente nos dados deve ser removido ou estimado.

- Os dados devem ser padronizados (ou seja, escalados) para tornar as variáveis comparáveis.

Os dados utilizados para a realização deste estudo são armazenados em um Sistema de Gerenciamento de Banco de Dados Relacional (SGBD) Teradata, onde são modelados de uma forma que sejam percebidos pelo usuário como tabelas (ou relações).

Teradata é um sistema de *Data Warehouse* que armazena e gerencia dados, sendo capaz de processar grandes volumes de dados de diferentes origens e disponibilizá-los para análises estratégicas de negócios.

Através da ferramenta SAS Enterprise Guide as tabelas da base de dados relacional são extraídas e relacionadas, para serem posteriormente processadas. O Enterprise Guide é uma ferramenta OLAP para Windows, orientado por projetos, e que possibilita acesso rápido a uma grande parte da potencialidade analítica do SAS para estatísticos, analistas de negócios e programadores SAS (Figura 12).

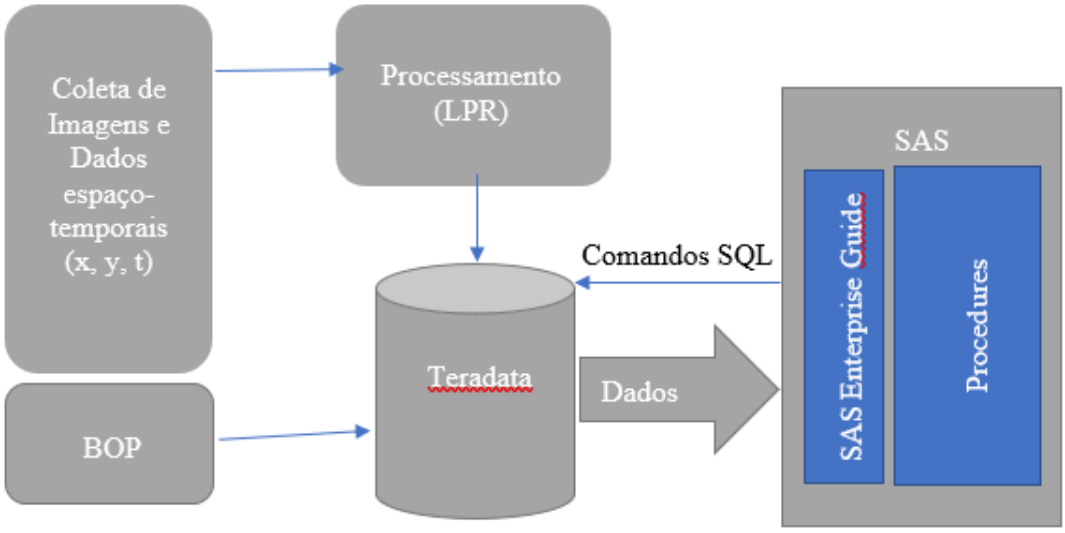


Figura 12: Coleta, processamento e manipulação de dados

Diferentemente dos dados gerados por um GPS onde a trajetória de um objeto é dada a partir de pontos gerados dinamicamente à medida que o objeto se desloca, neste estudo temos uma trajetória discretizada obtida a partir do registro de passagens de um objeto (veículo) por pontos com geolocalização pré-definida. Desta forma, podemos

aplicar a cada uma das trajetórias traçadas uma verificação capaz de identificar se algum ponto foi gerado incorretamente, seja por erro de leitura através do software de LPR, seja por uso indevido de sinal identificador (placa do veículo).

Da mesma forma, previamente é possível verificar a presença de dados faltantes nas tabelas contendo os dados a serem utilizados no processamento.

Não sendo possível recuperar os dados faltantes ou gerados incorretamente, seus registros são desconsiderados do conjunto de dados que compõem as trajetórias de cada objeto.

4.2.1 Mitigando a Inconsistência de Dados

Na base de dados utilizada para o presente estudo, alguns registros de passagens de veículos podem apresentar inconsistência, basicamente, pelos seguintes motivos:

- Leitura incorreta de caracteres da placa: Algoritmos de LPR, que reconhecem padrões de placas de veículos (*license plate*) podem "confundir" alguns caracteres gerados por imperfeições na placa ou caracteres muito parecidos.

- Leitura de caracteres da imagem diversas da placa: Alguns símbolos podem ser confundidos com caracteres de placas, como inscrições ou formas no veículo que podem fazer o algoritmo entender como que seja uma placa de identificação.

- Uso fraudulento da mesma placa por outro veículo (clone): uma fraude comumente utilizada por organizações criminosas é a utilização de placas "clonadas", ou seja, veículos com sinais identificadores alterados, utilizando placa de identificação de outro veículo com as mesmas características.

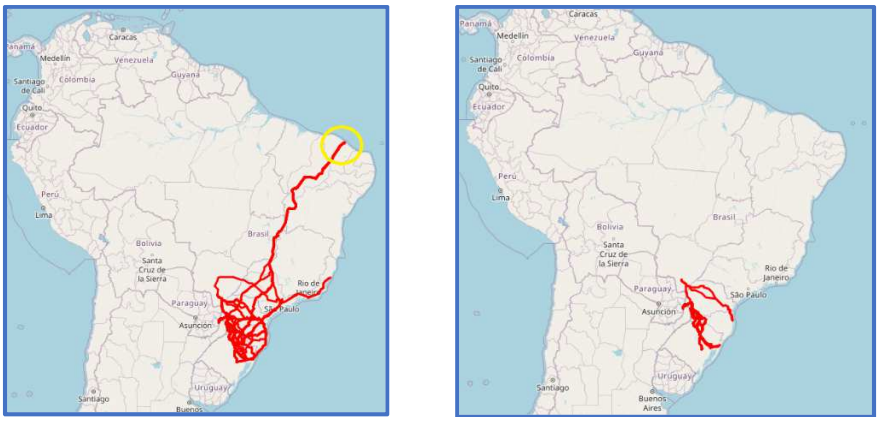
- Inserção falsa de dados manuais (praças de pedágio, etc): alguns locais de passagem de veículos, onde a leitura e o registro são realizados por pessoas, pode gerar registros incorretos.

- Ruído gerado automaticamente pelo sistema (*trash*): bases de dados pode gerar registros incorretos ou nulos, que devem ser tratados ou descartados.

Nesses casos, não sendo possível recuperar esses dados, devemos suprimi-los, considerando-os *outliers.*

Uma forma de identificar preliminarmente alguns desses *outliers* pode ser comparando o tempo de deslocamento entre pontos do sistema com o tempo médio de deslocamento neste trecho. Um deslocamento em um tempo muito menor do que o aceitável, pode ser um bom critério para descartar esse registro como parte de uma trajetória.

Neste trabalho desenvolveu-se um algoritmo em R capaz de traças as rotas rodoviárias estimadas percorridas pelos veículos, refazendo sua trajetória dentro de um determinado espaço de tempo. Na Figura 13 é possível observar as trajetórias construídas a partir de um conjunto de dados de passagens de veículo de um determinado grupo. Na Figura 13 (a) podemos observar trajetórias geradas em razão de dados gerados por erro de leitura de equipamentos. Na Figura 13 (b) temos as trajetórias construídas desconsiderando-se dados *outliers* e trajetórias menos significativas.



(a) (b)

Figura 13 – Trajetórias traçadas a partir de registros de passagens de veículos

No caso das trajetórias estudadas, considera-se *outliers* os registros gerados por falha do sistema e trajetórias menos significativas aquelas que são formadas por pontos com pouca frequência de registros de passagens de veículos (baixa densidade) ou que não atendem às condições descritas em 4.2.2.

Para se construir as trajetórias percorridas por veículos utilizou-se uma rotina em R baseada em uma API do Google Maps, que permite traçar rotas usuais e alternativas entre os pontos de passagem do sistema.

Da mesma forma, os erros de leitura nos pontos de registros de passagens podem ser detectados a partir de uma matriz formada pelo tempo estimado de deslocamento entre os pontos. Padrões equivocadamente entendidos como placas de veículos são adicionados a uma *black list* e suprimidos das amostras. Na figura 14 temos duas imagens em que o algoritmo LPR pode interpretar como sendo o mesmo *id.*

Figura 14 – Leitura incorreta de imagem pelo algoritmo LPR

Inicialmente utilizou-se a ferramenta SAS Enterprise Guide para o processo de extração e construção de tabelas relacionadas de um Banco de Dados Relacional (Teradata), como se pode ver na Figura 15 que mostra o *Process Flow* construído para a extração e filtragem de dados das bases referidas.

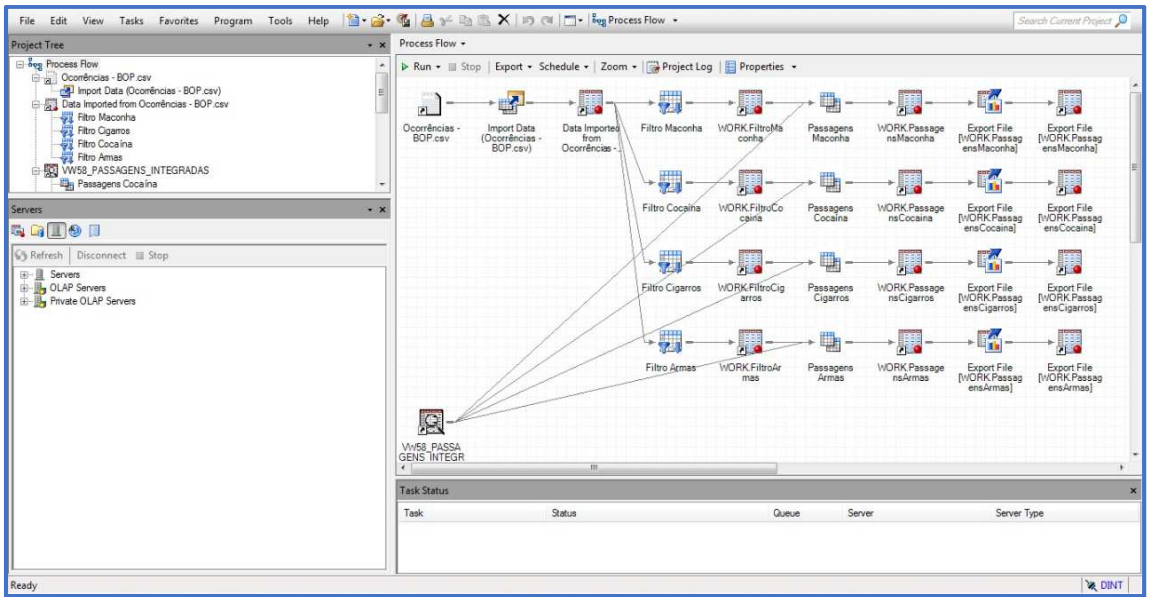


Figura 15 – processo de extração e filtragem de dados através do SAS Enterprise Guide

A partir de uma base de dados de ocorrências criminais (BOP), extraiu-se informações sobre registros de ocorrências relacionadas a atividades criminosas registradas em rodovias federais no estados do Rio Grande do Sul pela PRF, no período de 01/07/2019 a 30/06/2020, construindo-se tabelas com informações como como o *id* do veículo, tipo de veículo, latitude, longitude, enquadramento, tipo de apreensão, data/hora, além de outras informações (Tabela 5)

Alguns filtros e tratamentos de dados foram aplicados nesse *dataset,* como o preenchimento de dados faltantes (quanto possível) e eliminação de linhas contendo dados não recuperáveis. Aqui, também, optou-se por suprimir os registros de passagens de veículos tipo ônibus de linhas regulares, pois a trajetória destes veículos não está condicionada a uma atividade ilícita praticada eventualmente por algum passageiro, pois percorrem rotas pré-definidas e em horários previamente programados. Portanto, merecem um tratamento diferenciado.

De uma base de dados de registro de passagens de veículos do sistema da PRF, extraíram-se os registros de passagens dos veículos selecionados da base de ocorrências criminais, o que permitiu reconstituir as trajetórias percorridas por esses veículos no período selecionado entre 01/01/2019 a 30/06/2020.

Após a extração de dados que permitiram a construção das trajetórias, os dados foram categorizados em 4 tipos de crime: Tráfico de Maconha. Tráfico de Cocaína, Tráfico de Armas e Contrabando de Cigarros.

Durante o período selecionado foram extraídos 273 registros de ocorrências, sendo 5 por Tráfico de Armas, 85 por Contrabando de Cigarros, 93 por Tráfico de Cocaína e 90 por Tráfico de Maconha.

Para o estudo de trajetórias foram selecionadas, dentro do período selecionado, todas as ocorrências por Tráfico Internacional de Armas, as apreensões por Contrabando

de Cigarros com mais de 100 pacotes, por Tráfico de Cocaína com mais de 1 kg da droga e por Tráfico de Maconha com mais de 100 kg.

Após o enriquecimento com os registros de passagens dos veículos que definem suas trajetórias, obteve-se 26.069 registros de passagem de veículos: 482 para Tráfico de Armas, 6.601 para Contrabando de Cigarros, 10.861 para Tráfico de Cocaína e 8.125 para Tráfico de Maconha.

Tabela 5: Número de ocorrências e registros de passagens para cada um dos grupos

	Tráfico de Armas	Contrabando de Cigarros	Tráfico de Cocaína	Tráfico de Maconha
Nº de Ocorrências	5	85	93	90
Nº de Registros de Passagem	482	6.601	10.861	8.125

Na Figura 16 podemos ver uma ilustração que representa o processo de preparação e tratamento dos dados.

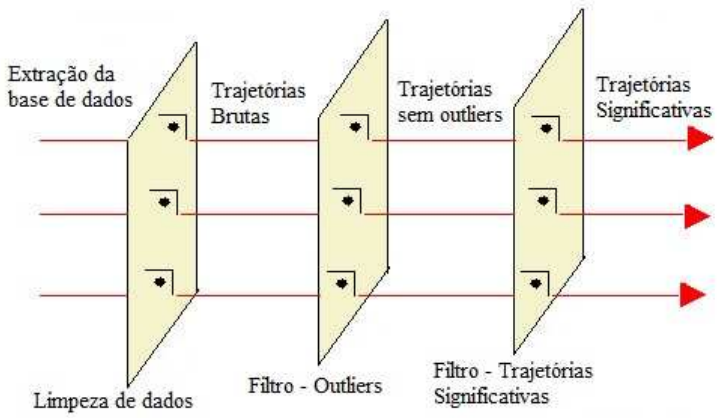


Figura16 – Fases no tratamento de dados

4.2.2 Trajetórias Significativas

Neste trabalho adotamos o conceito de trajetórias significativas com as trajetórias que apresentam forte relação com a atividade em que o veículo (ou objeto móvel) está proximamente envolvido.

Considerando-se que nas ocorrências registradas onde se flagrou o veículo transportando ilícitos ele estivesse em uma rota entre a origem e o destino para o

transporte ilegal, estabelecemos como trajetórias significativas o trecho entre a origem e o local onde foi flagrado cometendo o crime, de acordo com os seguintes critérios:

- A última trajetória percorrida pelo veículo, regredindo a partir do local de abordagem até o ponto mais distante, considerando o início da trajetória o ponto mais distante ou aquele em que registrar passagem sem ter registrado outras nas 24h anteriores.

4.3 Modelagem

Durante os experimentos foram empregados 2 processos de clusterização: K-*means*, e DBSCAN, além da extração de trajetórias significativas a partir do algoritmo descrito.

K-*means* necessita de uma definição *a priori* do número de clusters que se pretende dividir a amostra. Não é sensível a ruídos e considera todo e qualquer dados contido na amostra. Utilizamos aqui, alguns métodos estatísticos, baseado na distância euclidiana, para se determinar o número ótimo de clusters.

O DBSCAN é um método sensível à ruídos e não necessita a indicação do número de cluster *a priori,* definindo a partir de seu algoritmo o número apropriado. Identifica os valores *outliers*.

O Algoritmo para a extração de trajetórias significativas foi construído com base no conhecimento dos operadores em segurança que operam com o monitoramento de alvos suspeitos de atividades ilícitas, visando comparar com os resultados obtidos a partir dos modelos utilizados.

As amostras, por sua vez, serão submetidas a cada um dos métodos com três tratamentos diferentes: amostras brutas e amostras filtradas por outliers, comparando-se com as trajetórias significativas.

4.3.1 Experimentos

Trataremos aqui dos dados espaciais das trajetórias (id, x, y), adequados aos modelos empregados e comparáveis às informações obtidas a partir da construção de trajetórias significativas.

Inicialmente os dados utilizados, extraídos para cada uma das quatro amostras de trajetórias, foram normalizados, utilizando-se o método de clusterização *K-means,* para se verificar o comportamento das trajetórias de cada um dos grupos escolhidos através da análise de agrupamento dos pontos de registros de passagens de veículos pertencentes a cada grupo.

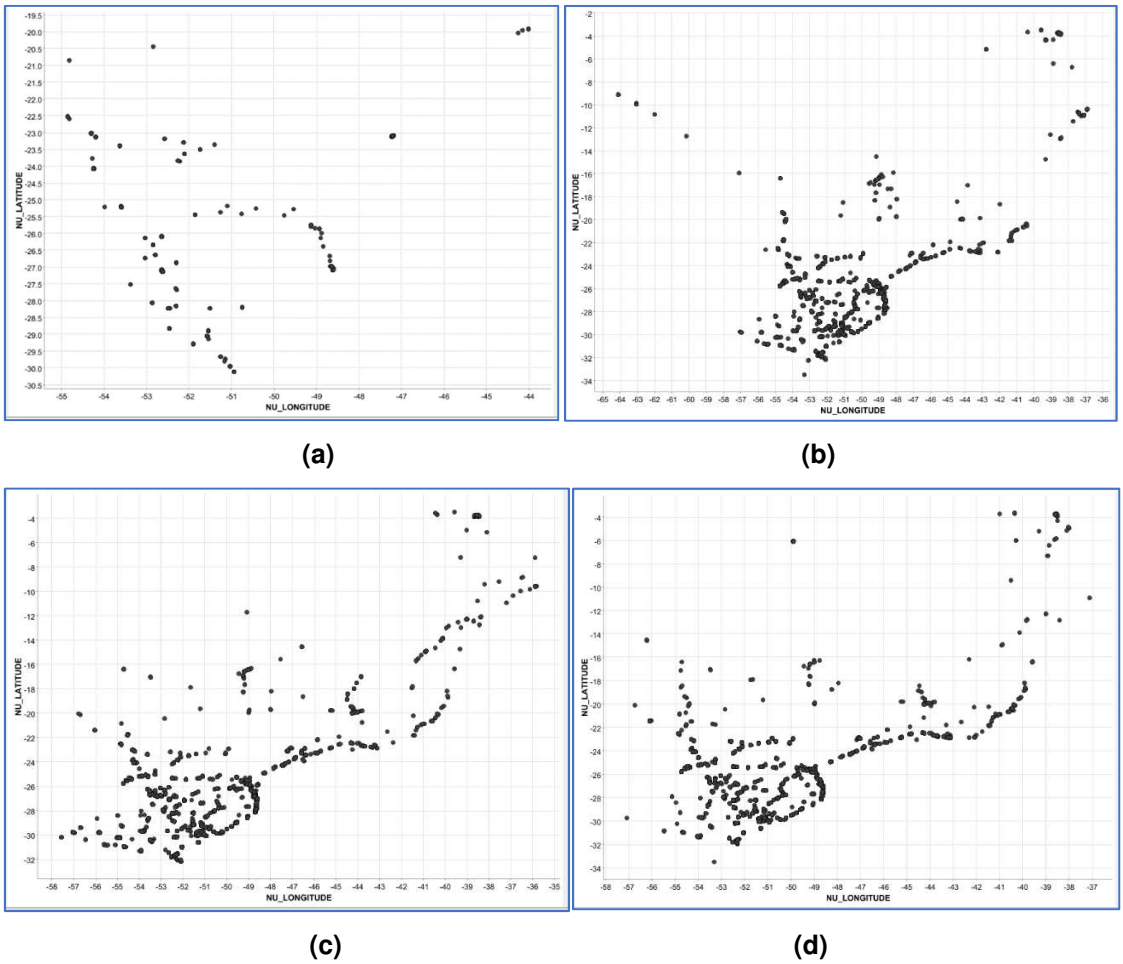


Gráfico 1 – Gráficos de dispersão dos pontos de passagem para veículos flagrados por (a) Tráfico de Armas, (b) Contrabando de Cigarros, (c) Tráfico de Cocaína e (d) Tráfico de Maconha

K-*means*

Para se implementar o método de clusterização K-*means* utilizou-se o pacote R {`factoextra`} [Kassambara, A. e Mundt, F., 2020] empregando-se a função `kmeans` para a formação dos clusters e `fviz_cluster` para construir os gráficos de agrupamentos.

Primeiramente utiliza-se aqui um método estatístico, baseado na distância euclidiana, para se determinar o número ótimo de clusters implementado através do pacote 'NbClust' do R [Charrad *at al.*, 2014]. Esse método utiliza mais de 30 índices *(Silhouette, C-index, Ratkowsky, McClain, Hubert, ...)* para estimar o número ótimo de clusters.

Tabela 6: Número de clusters sugerido a cada um dos grupos para o Método K-*means*

	Tráfico de Armas	Contrabando de Cigarros	Tráfico de Cocaína	Tráfico de Maconha
Sugestão de Nº de Clusters	3	4	5	5

O que geralmente acontece ao aumentar-se a quantidade de clusters no K-*means* é que as diferenças entre clusters se tornam muito pequenas, e as diferenças das observações intra-clusters vão aumentando. Então é preciso achar um equilíbrio em que as observações que formam cada agrupamento sejam o mais homogêneas possível e que os agrupamentos formados sejam o mais diferente uns dos outros.

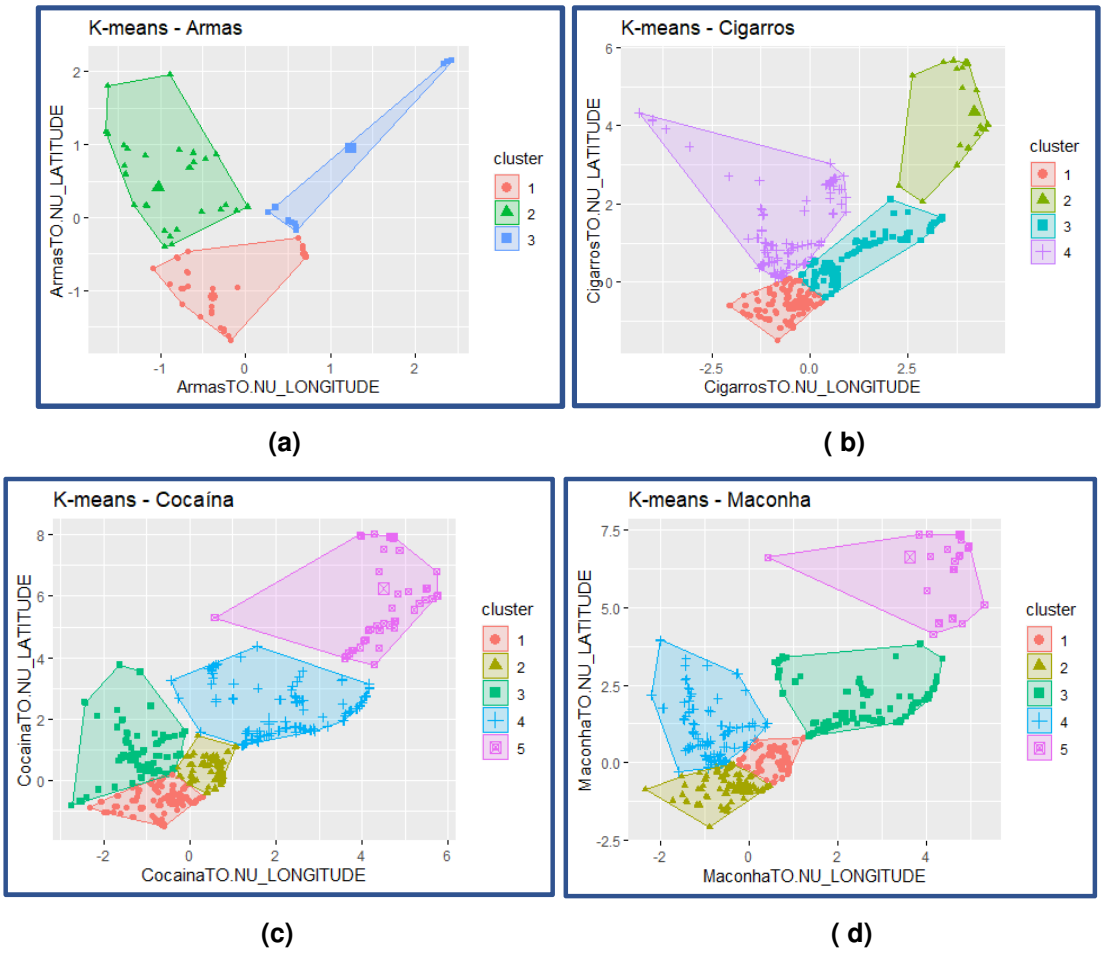


Gráfico 2 – Clusterização K-*means* para trajetórias de (a) Tráfico de Armas, (b) Contrabando de Cigarros, (c) Tráfico de Cocaína e (d) Tráfico de Maconha

A Tabela 7 mostra a distribuição de registros por cluster para cada um dos grupos. Pode-se observar que para cada grupo existe uma região, representada pelo Cluster1, com maior concentração de registros, variando sua região de concentração para cada grupo.

Tabela 7: Número de registros por cluster para cada um dos grupos utilizando-se o Método K-*means*

	CLUSTER 1	CLUSTER 2	CLUSTER 3	CLUSTER 4	CLUSTER 5	TOTAL
TRÁFICO DE ARMAS	*192*	*126*	*164*		-	*482*

CONTRABANDO DE CIGARROS	*3234*	*193*	*2068*	*1106*	*-*	*6601*
TRÁFICO DE COCAÍNA	*4896*	*3848*	*1506*	*498*	*113*	*10861*
TRÁFICO DE MACONHA	*2311*	*2927*	*518*	*2315*	*54*	*8125*

É possível medir-se a qualidade da classificação que K-*means* encontrou através da razão entre a soma das distâncias quadradas de cada observação com a média geral da amostra pela soma das distâncias quadradas das k médias para a média geral. É desejável que essa razão se aproxime de 1. Na Tabela 8 observa-se a os valores para cada um dos conjuntos de dados.

Tabela 8: Qualidade do Método K-*means* para cada um dos conjuntos de dados

	BETWEEN_SS / TOTAL_SS
TRÁFICO DE ARMAS	**83,70%**
CONTRABANDO DE CIGARROS	**84,20%**
TRÁFICO DE COCAÍNA	**86,80%**
TRÁFICO DE MACONHA	**83,80%**

DBSCAN

Para gerar os clusters abaixo através do método de clusterização DBSCAN utilizou-se os pacotes R {`dbscan`} [Hahsler M *at al*., 2019] e {`fpc`} [Henig, C., 2020].

Empregou-se aqui um método para se determinar o valor do *eps* ótimo. O método proposto consiste em calcular as distâncias dos k-vizinhos mais próximos em uma matriz de pontos. A ideia é calcular, a média das distâncias de cada ponto o seus k vizinhos mais próximos. O valor de k será especificado pelo usuário e corresponde a MinPts. Em seguida, essas distâncias k são plotadas em ordem crescente. O objetivo é determinar o "joelho", que corresponde ao parâmetro *eps* ideal. Um joelho corresponde a um limite onde uma mudança brusca ocorre ao longo da curva k-distância.

Utiliza-se aqui a função kNNdistplot[5] [Hahsler M *at al*., 2019] do pacote R {dbscan} para se desenhar o gráfico de distância k.

[5] https://github.com/mhahsler/dbscan

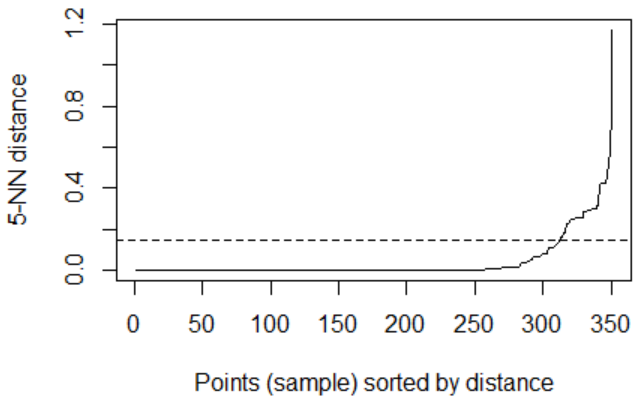


GRAFICO 3: Valor ótimo do parâmetro *eps* para a amostra Tráfico de Armas

Empregando-se o método acima, pudemos obter os valores para *eps* conforme demonstra a Tabela 9:

Tabela 9: Valores estimados para *eps* e valores para *MinPts* em cada um dos conjuntos de dados.

	eps	***MinPts***
Tráfico de Armas	***0,18***	***5***
Contrabando de Cigarros	***0,15***	***5***
Tráfico de Cocaína	***0,15***	***5***
Tráfico de Maconha	***0,15***	***5***

No Gráfico 4 é possível visualizar a representação do método de clusterização baseado em densidade DBSCAN para cada conjunto de dados. Nota-se que os pontos de menor densidade são representados por pontos pretos e considerados *outliers.*

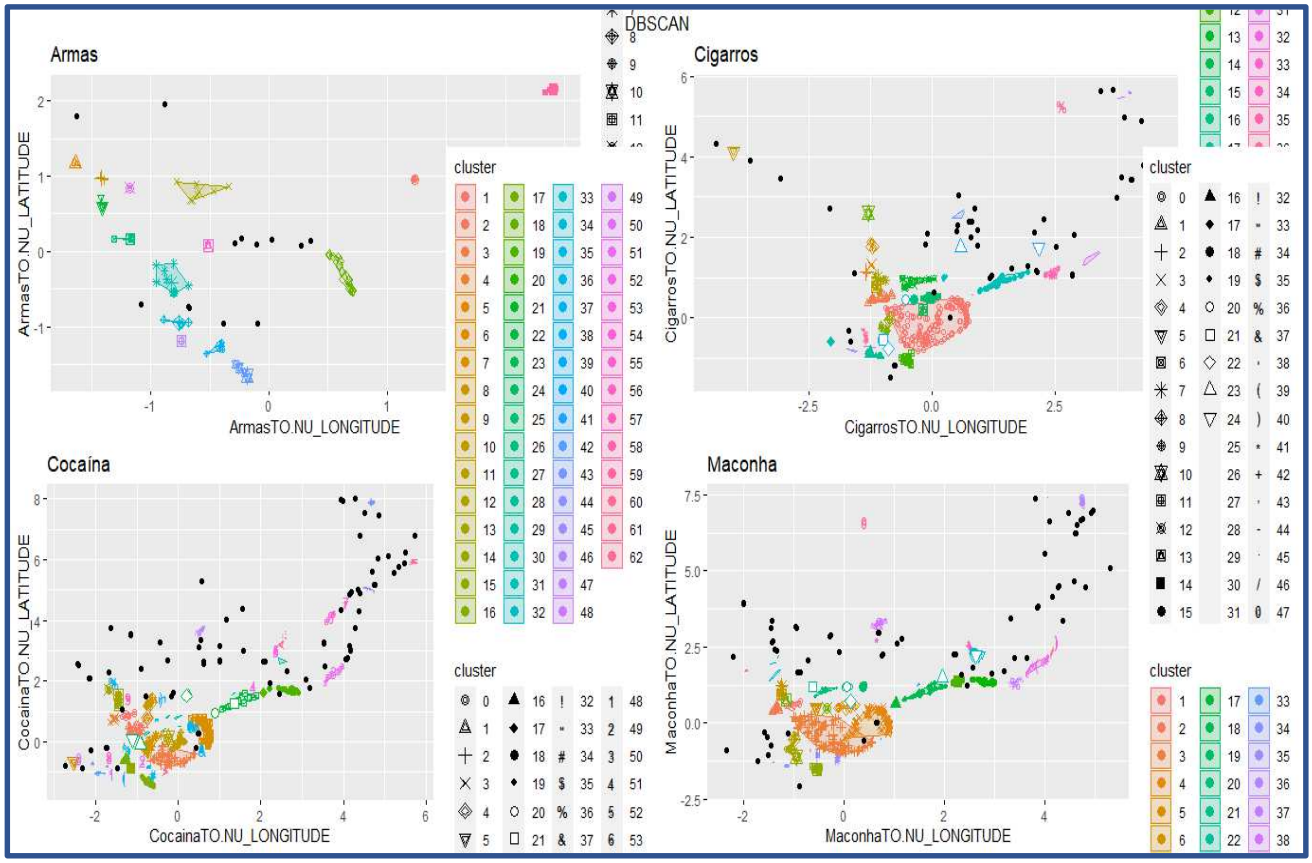


GRÁFICO 4: Clusterização DBSCAN para cada um dos conjuntos de dados

Trajetórias Significativas

Os critérios para se construir as trajetórias significativas foram estabelecidas a partir do conhecimento sobre o negócio dos agentes que operam com monitoramento e acompanhamento de trajetórias de alvos suspeitos de envolvimento em ilícitos. Seus critérios estão consolidados neste trabalho em 4.2.2 e o algoritmo utilizado para traçar-se as trajetórias foi implementado em R, utilizando-se pacotes como {`mapsapi`}, {`xml2`} e {`leaflet`}. O pacote {`mapsapi`} fornece uma interface para as APIs do Google Maps. As funções *mp_directions, mp_matrix e mp_geocode* são usadas para acessar as APIs *Directions, Matrix e Geocode*, respectivamente e permitem gerar as trajetórias a partir dos registros de passagens, com base nas trajetórias sugeridas pelo Google Maps.

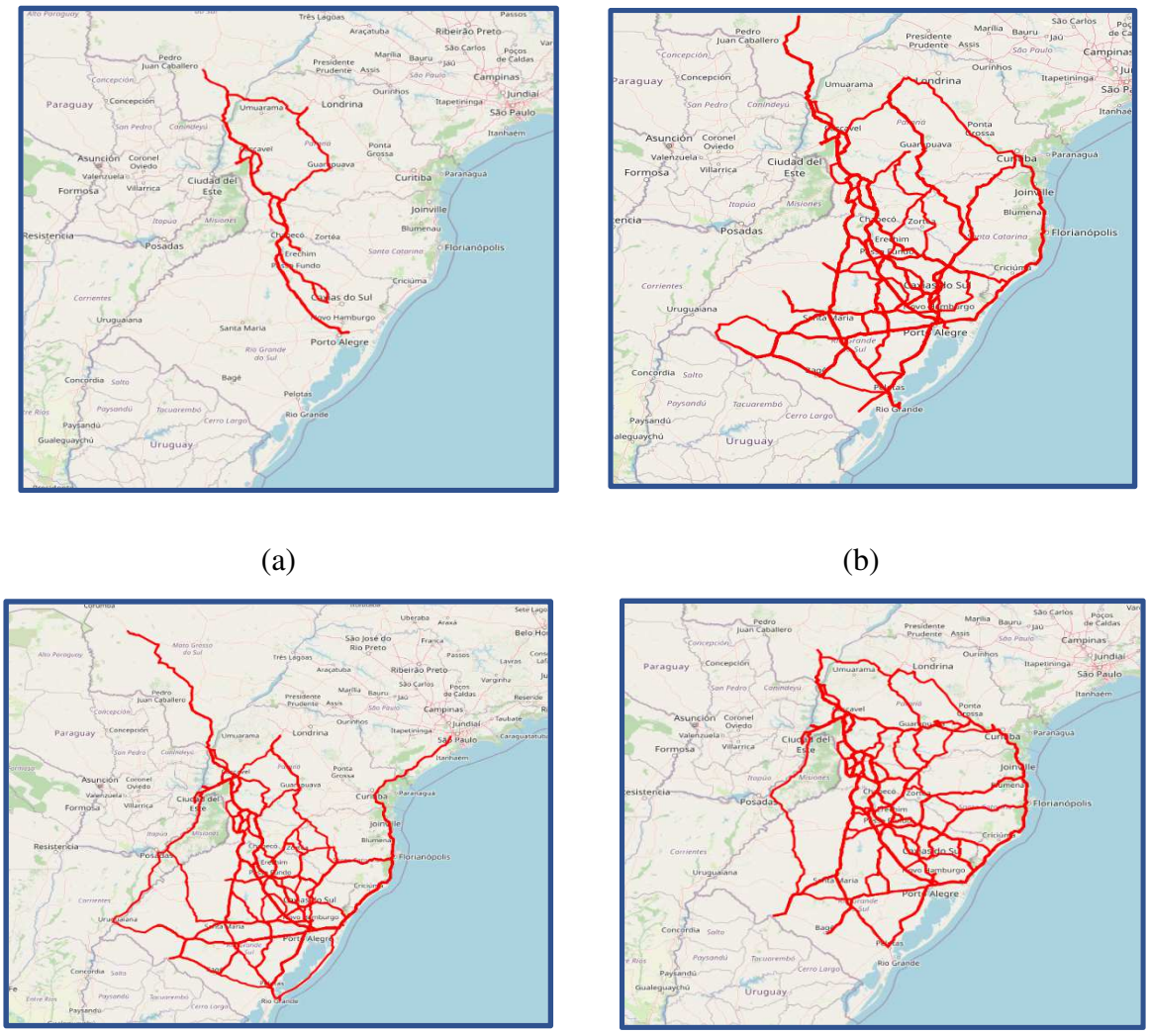


GRÁFICO 5: Trajetórias significativas considerando-se as principais rotas para cada grupo (a) Tráfico de Armas, (b) Contrabando de Cigarros, (c) Tráfico de Cocaína e (d) Tráfico de Maconha

5. Avaliação e Resultados

No experimento apresentado neste trabalho foram utilizados dois processos de clusterização bastante conhecidos: K-*means* e DBSCAN. Além disso, com o objetivo de comparar-se a eficiência dos modelos, foram construídas trajetórias significativas com base nos critérios indicados em 4.2.2.

5.1 Avaliação dos Modelos

O método K-*means*, que se baseia em parâmetros espaciais, como centroides e vizinhança de pontos, mostra-se útil para análises genéricas sobre grande quantidade de dados. Ele funciona bem para clusters compactos e bem separados. De forma geral, os métodos de particionamento (K-médias, agrupamento PAM) e agrupamento hierárquico são adequados para localizar clusters de formato esférico ou clusters convexos. Dados reais podem conter clusters de forma arbitrária (clusters não convexos: oval, linear, em forma de "S"), além de muitos *outliers* e ruídos.

O método K-*means*, por ser sensível a ruídos nos dados, mostra-se pouco eficiente para o processo de clusterização de trajetórias dos grupos de trajetórias aqui apresentados. DBSCAN, por ser um método de agrupamento baseado em densidade, é capaz de identificar melhor os agrupamentos de trajetórias, distinguindo dos valores que considera *outliers*. O método DBSCAN apresenta-se mais apropriada para tal aplicação. No entanto, por ser mais complexo, limita-se a menor quantidade de dados processados.

Antes de avaliar-se os resultados alcançados, importante destacar as características do conjunto de dados analisado. O conjunto de dados sobre trajetórias de veículos do grupo *tráfico de armas* apresenta poucos registros, devido ao pequeno número de ocorrências desse tipo. O conjunto de dados dos grupos sobre *tráfico de drogas (cocaína e maconha)* são os que apresentam os maiores volumes de dados, por isso adotou-se critérios de seleção para apreensões de maior volume.

Ao analisarmos os resultados do processo de clusterização através do DBSCAN, é possível observar que este método foi capaz de eliminar uma grande quantidade de *outliers,* concentrando os clusters onde há maior densidade de pontos. Porém, quando temos um volume grande de registros, distribuídos por várias regiões, mostrou-se necessário reduzir o valor do parâmetro *eps*, o que aumentou significativamente o número de clusters atribuído arbitrariamente pelo método.

O Gráfico 6 apresenta a representação do método DBSCAN para cada conjunto de dados, suprimidos os pontos *outliers*. Aqui é possível observar-se conjuntos de clusters que indicam as tendências de trajetórias, apresentando características próprias a cada tipo de atividade ilícita.

Tabela 10: Valores atribuídos para *eps* com *MinPts = 5* para cada um dos conjuntos de dados.

	Armas			Cigarros			Cocaína			Maconha		
eps	Pts	Clus	Out	Pts	Clus	Out	Pts	Clus	Out	Pts	Clus	Out
0,500	482	15	16	6601	41	71	10861	43	79	8125	37	81
0,300	482	21	29	6601	79	128	10861	86	122	8125	87	105
0,100	482	20	69	6601	149	252	10861	153	282	8125	166	254
0,050	482	20	75	6601	172	376	10861	181	412	8125	201	412
0,030	482	19	84	6601	192	424	10861	202	455	8125	226	458
0,010	482	22	91	6601	213	502	10861	223	533	8125	247	524
0,001	482	22	119	6601	235	603	10861	249	642	8125	276	606

Analisando-se a Tabela 10 é possível observar que para conjuntos menores de dados, com o conjunto Armas, tem-se uma rápida estabilização para o número de clusters atribuídos pelo algoritmo à medida que se reduz o valor do parâmetro *eps*. Porém, quando reduzimos o valor de tal parâmetro o algoritmo tende a considerar cada vez mais pontos como *outliers*.

Já para os conjuntos de dados maiores o algoritmo aumenta o número de clusters atribuídos à medida que se tem um valor para *eps* cada vez menor. A função proposta por Hahsler *at al.* (2019) é capaz de encontrar o valor ótimo para o parâmetro *eps* de forma que o algoritmo identifique as concentrações de registros de passagens sem que se tenha perda de informação, considerando-se um número cada ver maior de *outliers*.

Comparando-se com trajetórias significativas para cada conjunto de dados, observa-se que o método é capaz de identificar as regiões de maior concentração de trajetórias. Além disso, pode ser capaz de identificar trajetórias não usuais como rotas intermediárias ou alternativas que normalmente não são consideradas como rotas significativas por operadores que realizam o monitoramento de trajetórias.

Considerando-se o custo computacional para se implementar um algoritmo de extração de trajetórias significativas baseado em regras estabelecidas *a priori* em um grande volume de dados, deve-se considerar o método DBSCAN como uma técnica capaz de se extrair características de trajetórias de forma menos onerosa em termos de processamento, sendo capaz de identificar mudanças de comportamento não detectadas por métodos que definem regras *a priori.*

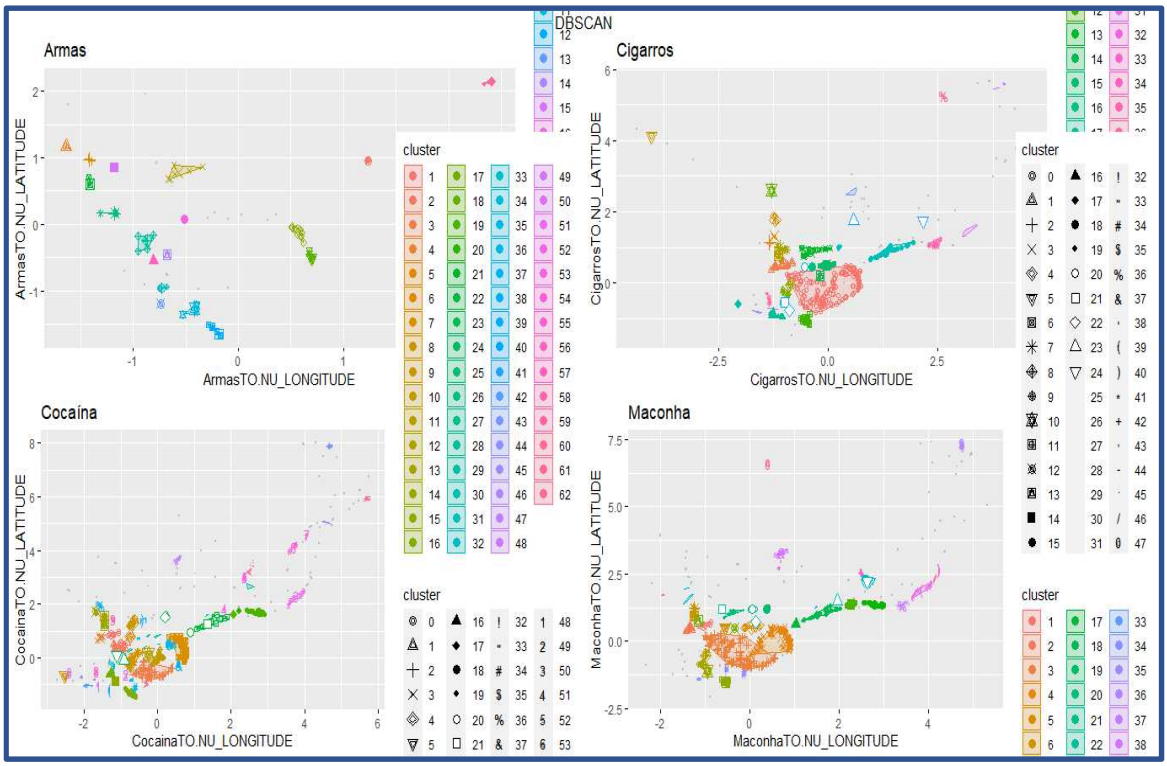


GRÁFICO 6: Clusterização DBSCAN para cada um dos conjuntos de dados, suprimidos os *outlier*

5.2 Resultados Alcançados

O Método DBSCAN de análise de cluster baseado em densidade mostra-se neste estudo bastante eficiente na extração de características para cada uma das atividades, sendo capaz de ressaltar as regiões de maior concentração de trajetórias e identificar trajetórias de menor frequência, restringindo às regiões de maior circulação.

Ele é capaz de identificar outliers e ressaltar a concentração de trajetórias de maior relevância para cada uma das atividades, indicando as trajetórias mais populares para cada grupo.

6. Conclusão

Neste trabalho foi possível observar que apesar das limitações de se explorar apenas a dimensão espacial no processo de clusterização, os resultados obtidos com o uso de técnicas de agrupamento espacial de aplicativos com ruído baseado em densidade (DBSCAN) podem se mostrar relevantes quando comparados com trajetórias significativas de dados de trajetórias reais em rodovias.

Importante ressaltar que as trajetórias significativas tratadas nesse trabalho não podem ser consideradas um conjunto completo de trajetórias importantes a serem consideradas em um processo de descoberta do conhecimento, mas sim um parâmetro por serem trajetórias já confirmadas como rotas de transporte de ilícitos. Muitos outros padrões podem ser extraídos das trajetórias através de um processo de clusterização como o DBSCAN aqui apresentado.

O tema *clustering* de trajetórias, especificamente explorando-se bases de dados de registros de passagens de veículos por pontos em rodovias através de ITS, mostra-se um campo a ser explorados e experimentado. Técnicas apropriadas podem ser desenvolvidas ou adaptadas para que se possa extrair conhecimento de um crescente volume de dados gerados por esses sistemas.

Conhecer os desafios e as limitações impostas por este tipo de aplicação mostra-se fundamental para o desenvolvimento e aperfeiçoamento de técnicas apropriadas para a descoberta do conhecimento sobre trajetórias de objetos móveis em rodovias.

6.1 Trabalhos Futuros

Com o desenvolvimento do presente trabalho, foi possível vislumbrar-se algumas possibilidades de trabalhos futuros relacionados aos temas propostos. Técnicas de clusterização bem ajustadas ao conjunto de dados explorado podem servir como subsídio a um modelo de classificação para implementar-se regras de indicação de abordagens policiais. Agregando-se dados de outras bases e informações produzidas dentro do órgão policial podem proporcionar a construção de classificadores que podem alcançar índices de assertividade bastante relevantes.

Ao trabalharmos com informações contendo um certo grau de incerteza ou classes não muito bem definidas, parece conveniente utilizar-se classificadores baseados em lógica Fuzzy, com uso de algoritmos de clusterização como o FCM (Fuzzy C-means clustering) Dunn(1973).

Considerar as informações temporais, como data/hora de passagem de veículos pelos pontos de registro, bem como o tempo ou velocidade média de deslocamento entre os pontos pode tornar o processo de agrupamento mais complexo, extraindo-se informações mais sofisticadas, o que exige um esforço computacional maior para a implementação dos modelos. Algoritmos modificados a partir do DBSCAN como o T-DBSCAN [Chen *et al*., 2014] ou TRACLUS [Lee *at al*., 2007] podem aproveitar essas informações para extrair-se conhecimentos associados às informações espaço-temporais.

Referências

M. C. Gonzalez, C. A. Hidalgo, and A. L. Barabasi. Understanding individual human mobility patterns. Nature, 453(7196):779–782, 2008.

W. Kellerer, C. Bettstetter, C. Schwingenschlogl, P. Sties, and K. E. Steinberg. (auto) mobile communication in a heterogeneous and converged world. IEEE Personal Commu- nications, 8(6):41–47, 2001.

M. Holyoak, R. Casagrandi, R. Nathan, E. Revilla, and O. Spiegel. Trends and missing parts in the study of movement ecology. Proceedings of the National Academy of Sciences, 105(49):19060–19065, 2008.

Transporte, C.-C. N. Do (2017). Transporte rodoviário: desempenho do setor, infraestrutura e investimentos. *CNT Confederação Nacional do Transporte*, p. 70.

Central Intelligence Agency (CIA), The World Factbook, https://www.cia.gov/library/publications/the-world-factbook/geos/br.html [accessed on Jul 20]

IBGE (2016). Pesquisa Anual de Serviços (PAS) - Informativo. *IBGE Instituto Brasileiro de Geografia e Estatística,* p. 8

Gudmundsson, Joachim; Laube, Patrick; Wolle, Thomas (2012). *Computational movement analysis.* In: Kresse, Wolfgang; Danko, David M. *Springer Handbook of Geographic Information. Dordrecht: Springer, 725-741*.

Masciari *at al*. 2013

MJSP Apresentação do Ministro: Projetos Estratégicos para o Combate aos Crimes de Corrupção, Crime Organizado e Crimes Violentos, Câmara dos Deputados, 2019, https://www2.camara.leg.br/atividade-legislativa/comissoes/comissoes-permanentes/cspcco/audiencias-publicas/APRESENTACAOCAMARAMJSP_VERSAOCOMINTROfinal.pptx [Acessed on Jul 20]

Organisation for Security and Co-Operation in Europe (2017). *OSCE Guidebook Intelligence-Led Policing*. v. 13

LAROSE, Daniel T.; LAROSE, Chantal D. **Discovering knowledge in data: an introduction to data mining**. John Wiley & Sons, 2014.

SORENSEN, Paul; ECOLA, Liisa; WACHS, Martin. **Mileage-based user fees for transportation funding: A primer for state and local decisionmakers**. Rand Corporation, 2012.

SUSSMAN, Joseph. **Introduction to transportation systems**. 2000.

QURESHI, Kashif Naseer; ABDULLAH, Abdul Hanan. A survey on intelligent transportation systems. **Middle-East Journal of Scientific Research**, v. 15, n. 5, p. 629-642, 2013.

MARSH. Cuttin-edge products. Nov 2010; Available from: http://www.marshproducts.com/.

GOTTFRIED, Interpretation–BMI B.; AGHAJAN, H. Progress in movement pattern analysis. **Behaviour monitoring and interpretation-BMI: Smart environments**, v. 3, p. 43, 2009.

Gudmundsson J., Laube P., Wolle T. (2008) Movement Patterns in Spatio-temporal Data. In: Encyclopedia of GIS. Springer, Boston, MA

T. Hägerstrand. What about people in regional science. Papers of the Regional Science Association, 24:7–21, 1970

H. J. Miller. Modelling accessibility using space-time prism concepts within geographical information systems. International Journal of Geographical Information Systems, 5(3):287– 301, 1991.

F. Schmid, K.-F. Richter, and P. Laube. Semantic trajectory compression, 2009.

D. Bernstein and A. Kornhauser. An introduction to map matching for personal navigation assistants. Technical report, New Jersey TIDE Center, 1996.

C. Du Mouza and P. Rigaux. Mobility patterns. Geoinformatica, 9(4):297–319, 2005.

SANTOS, Irineu Júnior Pinheiro dos. TRACTS: um método para classificação de trajetórias de objetos móveis usando séries temporais. 2011.

LAUBE,P., DENNIS,T., FORER,P., AND WALKER, M. Movement beyond the snap- shot: Dynamic analysis of geospatial lifelines. Computers, Environment and Urban Systems 31, 5 (2007), 481–501. doi:10.1016/j.compenvurbsys.2007.08.002.

SPACCAPIETRA, Stefano *et al*. A conceptual view on trajectories. **Data & knowledge engineering**, v. 65, n. 1, p. 126-146, 2008.

SPINSANTI, Laura; BERLINGERIO, Michele; PAPPALARDO, Luca. Mobility and Geo-Social Networks. 2013.

Miranda Junior, P. O. De and Abreu, J. F. De (2018). Mobile Intelligence - Um arcabouço para análise e visualização de dados tempo-espaciais originados por dispositivos móveis - DOI 10.5752/P.2316-9451.2013v1n2p45. *Abakós*, v. 1, n. 2, p. 45–66.

LAUBE, Patrick; IMFELD, Stephan. Analyzing relative motion within groups oftrackable moving point objects. In: **International Conference on Geographic Information Science**. Springer, Berlin, Heidelberg, 2002. p. 132-144.

LAUBE, Patrick; IMFELD, Stephan; WEIBEL, Robert. Discovering relative motion patterns in groups of moving point objects. **International Journal of Geographical Information Science**, v. 19, n. 6, p. 639-668, 2005.

FENG, Zhenni; ZHU, Yanmin. A survey on trajectory data mining: Techniques and applications. **IEEE Access**, v. 4, p. 2056-2067, 2016.

Tanuja, V. and Govindarajulu, P. (2016). A Survey on Trajectory Data Mining. *International Journal of Computer Science and Security (IJCSS)*, v. 10, n. 5, p. 195.

PARENT, Christine *et al*. Semantic trajectories modeling and analysis. **ACM Computing Surveys (CSUR)**, v. 45, n. 4, p. 1-32, 2013.

FERRERO, Carlos Andres; ALVARES, Luis Otávio; BOGORNY, Vania. Multiple aspect trajectory data analysis: research challenges and opportunities. In: **GeoInfo**. 2016. p. 56-67.

Sung *at al*. 2012

ZHENG, Yu. Trajectory data mining: an overview. **ACM Transactions on Intelligent Systems and Technology (TIST)**, v. 6, n. 3, p. 1-41, 2015.

GENG, Xiaoliang; ARIMURA, Hiroki; UNO, Takeaki. Pattern mining from trajectory gps data. In: **2012 IIAI International Conference on Advanced Applied Informatics**. IEEE, 2012. p. 60-65.

BIAN, Jiang *et al*. A survey on trajectory clustering analysis. **arXiv preprint arXiv:1802.06971**, 2018.

KHAING, Hnin Su; THEIN, Thandar. An efficient clustering algorithm for moving object trajectories. In: **Proceedings of the 3rd International Conference on Computational techniques and Artificial Intelligence (ICCTAI'14)**. 2014. p. 11-12.

CASSIANO, Keila Mara; PESSANHA, José Francisco Moreira. ANÁLISE ESPECTRAL SINGULAR COM CLUSTERIZAÇÃO BASEADA EM DENSIDADE NA MODELAGEM DE SÉRIES TEMPORAIS.

CHEN, Wen; JI, M. H.; WANG, J. M. T-DBSCAN: A Spatiotemporal Density Clustering for GPS Trajectory Segmentation. **International Journal of Online Engineering**, v. 10, n. 6, 2014.

GAFFNEY, Scott; SMYTH, Padhraic. Trajectory clustering with mixtures of regression models. In: **Proceedings of the fifth ACM SIGKDD international conference on Knowledge discovery and data mining**. 1999. p. 63-72.

LEE, Jae-Gil; HAN, Jiawei; WHANG, Kyu-Young. Trajectory clustering: a partition-and-group framework. In: **Proceedings of the 2007 ACM SIGMOD international conference on Management of data**. 2007. p. 593-604.

MAZIMPAKA, Jean Damascène; TIMPF, Sabine. Trajectory data mining: A review of methods and applications. **Journal of Spatial Information Science**, v. 2016, n. 13, p. 61-99, 2016.

DEMŠAR, Urška *et al*. Analysis and visualisation of movement: an interdisciplinary review. **Movement ecology**, v. 3, n. 1, p. 1-24, 2015.

Abula *at al*. 2008

GIDOFALVI, Gyozo; HUANG, Xuegang; PEDERSEN, Torben Bach. Privacy-preserving data mining on moving object trajectories. In: **2007 International Conference on Mobile Data Management**. IEEE, 2007. p. 60-68.

DEMŠAR, Urška et al. Analysis and visualisation of movement: an interdisciplinary review. **Movement ecology**, v. 3, n. 1, p. 1-24, 2015.

PINHEIRO, Patricia Peck. **Proteção de Dados Pessoais: Comentários à Lei n. 13.709/2018-LGPD**. Saraiva Educação SA, 2020.

BOGORNY, V.; BRAZ, F. J. Introdução a Trajetórias de Objetos Móveis: conceitos, armazenamento e análise de dados. **Univille**, 2012.

MACHADO, Diego; DONEDA, Danilo. Proteção de dados pessoais e criptografia: tecnologias

criptográficas entre anonimização e pseudonimização de dados. Revista dos Tribunais. vol.

998. Caderno Especial. p. 99-128. São Paulo: Ed. RT, dezembro 2018

ZHENG, Yu. Trajectory data mining: an overview. **ACM Transactions on Intelligent Systems and Technology (TIST)**, v. 6, n. 3, p. 1-41, 2015.

FENG, Zhenni; ZHU, Yanmin. A survey on trajectory data mining: Techniques and applications. **IEEE Access**, v. 4, p. 2056-2067, 2016.

BIAN, Jiang *et al*. A survey on trajectory clustering analysis. **arXiv preprint arXiv:1802.06971**, 2018.

OSSAMA, Omnia; MOKHTAR, Hoda MO; EL-SHARKAWI, Mohamed E. An extended k-means technique for clustering moving objects. **Egyptian Informatics Journal**, v. 12, n. 1, p. 45-51, OSSAMA 2011.

KHAING, Hnin Su; THEIN, Thandar. An efficient clustering algorithm for moving object trajectories. In: **Proceedings of the 3rd International Conference on Computational techniques and Artificial Intelligence (ICCTAI'14)**. 2014. p. 11-12.

Charrad, Malika, Nadia Ghazzali, Véronique Boiteau, and Azam Niknafs. 2014. "NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set." *Journal of Statistical Software* 61: 1–36. http://www.jstatsoft.org/v61/i06/paper.

HAHSLER, Michael; PIEKENBROCK, Matthew; DORAN, Derek. dbscan: Fast density-based clustering with r. **Journal of Statistical Software**, v. 91, n. 1, p. 1-30, 2019.