Instituto Federal de Educação, Ciência e Tecnologia da Paraíba

Campus João Pessoa

Programa de Pós-Graduação em Tecnologia da Informação

Nível Mestrado Profissional

João Batista Firmino Junior

ANÁLISE COMPARATIVA DE MODELOS DE PREDIÇÃO DE DESTINO QUE USAM CADEIAS DE MARKOV E CADEIAS OCULTAS DE MARKOV

DISSERTAÇÃO DE MESTRADO

JOÃO PESSOA

2025

João Batista Firmino Junior

Análise Comparativa de Modelos de predição de Destino que usam Cadeias de Markov e Cadeias Ocultas de Markov

Dissertação apresentada como requisito parcial para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós-Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB.

Orientador: Prof. Dr. Francisco Dantas Nobre Neto

João Pessoa

2025

Dados Internacionais de Catalogação na Publicação (CIP)
Biblioteca Nilo Peçanha - *Campus* João Pessoa, PB.

F525a Firmino Junior, João Batista.

Análise comparativa de modelos de predição de destino que usam cadeias de Markov e cadeias ocultas de Markov / João Batista Firmino Junior. – 2025.
84 f. : il.

Dissertação (Mestrado em Tecnologia da Informação) – Instituto Federal de Educação da Paraíba / Programa de Pós-Graduação em Tecnologia da Informação (PPGTI), 2025.
Orientação: Prof. Dr. Francisco Dantas Nobre Neto.

1. Aprendizado de máquina. 2. Markov. 3.Tesselação. 4. Predição de destino. 5. Trajetória. I. Título.

CDU 004.8:519.22(043)

Bibliotecária responsável Lucrecia Camilo de Lima – CRB 15/132

MINISTÉRIO DA EDUCAÇÃO
SECRETARIA DE EDUCAÇÃO PROFISSIONAL E TECNOLÓGICA
INSTITUTO FEDERAL DE EDUCAÇÃO, CIÊNCIA E TECNOLOGIA DA PARAÍBA

PROGRAMA DE PÓS-GRADUAÇÃO *STRICTO SENSU*

MESTRADO PROFISSIONAL EM TECNOLOGIA DA INFORMAÇÃO

JOÃO BATISTA FIRMINO JÚNIOR

Análise Comparativa de Modelos de Predição de Destinos que usam Cadeias de Markov e Cadeias Ocultas de Markov

Dissertação apresentada como requisito para obtenção do título de Mestre em Tecnologia da Informação, pelo Programa de Pós- Graduação em Tecnologia da Informação do Instituto Federal de Educação, Ciência e Tecnologia da Paraíba – IFPB - Campus João Pessoa.

Aprovado em 30 de junho de 2025

Membros da Banca Examinadora:

Dr. Francisco Dantas Nobre Neto

IFPB - PPGTI

Dr. Francisco Petrônio Alencar de Medeiros

IFPB

Dr. Bruno Neiva Moreno

IFPB

Dr. Claudio de Souza Baptista

UFCG

João Pessoa/2025

Documento assinado eletronicamente por:

- **Francisco Dantas Nobre Neto**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 29/08/2025 16:45:59.
- **Claudio de Souza Baptista**, **PRESTADOR DE SERVIÇO**, em 30/08/2025 06:12:35.
- **Francisco Petronio Alencar de Medeiros**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 30/08/2025 09:29:48.
- **Bruno Neiva Moreno**, **PROFESSOR ENS BASICO TECN TECNOLOGICO**, em 02/09/2025 08:25:37.

Este documento foi emitido pelo SUAP em 28/08/2025. Para comprovar sua autenticidade, faça a leitura do QRCode ao lado ou acesse https://suap.ifpb.edu.br/autenticar-documento/ e forneça os dados abaixo:

Código 756101
Verificador: 3cebb7bbab
Código de Autenticação:

Este trabalho é dedicado à busca pelo conhecimento científico.

AGRADECIMENTOS

Os agradecimentos vão para minha família, pelo apoio, e para o IFPB. Em especial, para os docentes e discentes do Programa de Pós-Graduação em Tecnologia da Informação e ao grupo de pesquisa SIDE (*Semantics, Intelligence and Data Ecosystems*), pela troca de experiências e conhecimentos. Dentre os discentes, também enfatizo a influência de Msc. Janderson Dutra e, dentre os docentes do Programa de Pós-Graduação em Tecnologia da Informação (PPGTI), a do Dr. Francisco Dantas Nobre Neto, Professor orientador deste trabalho.

RESUMO

A predição de destinos é uma funcionalidade cada vez mais relevante em aplicações de mobilidade urbana, por oferecer ao usuário sugestões de rotas e lugares com base em padrões de deslocamento e, por vezes, informações com base em dados contextuais. Este trabalho realiza uma comparação entre dois modelos, sendo um com base em Cadeias de Markov, e outro com base em Cadeias Ocultas de Markov — aplicadas a um conjunto de dados reais de mobilidade veicular. A escolha metodológica foi fundamentada em um Mapeamento Sistemático da Literatura (MSL), dentre outras leituras. A análise foi realizada considerando conjuntos de etapas. O primeiro conjunto dessas etapas objetivou geração das amostras adequadas para os modelos; o segundo conjunto, com as predições e etapas para a comparação. A avaliação foi realizada com validação cruzada (K-Fold, k=10) e teste *t-student* para amostras pareadas desses *folds*, revelando que a performance dos modelos é sensível a um limite de corte das amostras para balanceamento dos dados conforme os veículos. Esse limite foi determinado para melhorar o valor das precisões, bem como a existência de mais rodadas em que os algoritmos funcionaram sem erros de construção dos modelos. Dessa forma, buscou-se responder à Questão de Pesquisa "Existe diferença quanto ao uso dos modelos de predição de destino com base em Cadeias de Markov e Cadeias Ocultas de Markov, no contexto de tráfego urbano e no uso de veículos individuais?". A resposta é que, após testes estatísticos, foi encontrado que há diferença estatística entre as duas técnicas avaliadas. Ou seja, globalmente, com Cadeias de Markov possuindo cerca de 59% de precisão; e com Cadeias Ocultas de Markov com aproximadamente 61% de precisão. Isso revelou que o modelo com base em Cadeias Ocultas de Markov, a partir do conjunto de dados utilizado, apresentou, com dados contextuais, melhores precisões na maior parte dos casos.

Palavras-chaves: Aprendizado de Máquina; Markov; tesselação; predição de destinos; trajetórias.

ABSTRACT

Destination prediction is an increasingly relevant functionality in urban mobility applications, offering users route and location suggestions based on displacement patterns and, at times, information derived from contextual data. This work conducts a comparison between two models: one based on Markov Chains and another based on Hidden Markov Chains — applied to a real vehicular mobility dataset. The methodological choice was grounded in a Systematic Literature Mapping (SLM), among other readings. The analysis was performed considering sets of stages. The first set of stages aimed at generating appropriate samples for the models; the second set involved predictions and comparison stages. The evaluation was conducted using cross-validation (K-Fold, k=10) and paired-sample *t-test* for these *folds*, revealing that model performance is sensitive to a sample cutoff threshold for data balancing according to vehicles. This threshold was determined to improve precision values as well as the occurrence of more rounds in which the algorithms functioned without model construction errors. Thus, this study sought to answer the Research Question: "Is there a difference in the use of destination prediction models based on Markov Chains and Hidden Markov Chains in the context of urban traffic and individual vehicle usage?" The answer is that, after statistical tests, a statistical difference was found between the two evaluated techniques. That is, globally, Markov Chains achieved approximately 59% precision, while Hidden Markov Chains achieved approximately 61% precision. This revealed that the Hidden Markov Chain-based model, using the employed dataset with contextual data, presented better precision in most cases.

Key-words: Machine Learning; Markov; tessellation; destination prediction; trajectories..

LISTA DE FIGURAS

LISTA DE TABELAS

LISTA DE ABREVIATURAS E SIGLAS

GPS	*Global Positioning System*
VED	*Vehicle Energy Dataset*
HMM	*Hidden Markov Models*
LSTM	*Long Short Term Memory*
FCM	*Fuzzy C-means*
RNN	*Recurrent Neural Network*
PPM	*Prediction by Partial Matching*
ESN	*Echo State Network*
FFNN	*FeedFoward Neural Network*
STI-GCN	*Spatial-Temporal Interaction-aware Graph Convolution Network*
GRU	*Gated Recurrent Unit*
CNN	*Convolutional Neural Network*

LISTA DE SÍMBOLOS

D	Parâmetro de distância
T	Parâmetro de tempo
SP	Subtrajetória
d(*,*)	Distância de dois pontos GPS

SUMÁRIO

1 INTRODUÇÃO

No contexto da mobilidade urbana, tanto do ponto de vista do poder público como do ponto de vista do cidadão comum e seus destinos, há desafios relacionados à temática de identificação de padrões de mobilidade.

Conhecer esses padrões de mobilidade são úteis, por exemplo, para encontrar anomalias que podem piorar o fluxo do trânsito; descobrir novas rotas em caso de uma via interrompida; minimizar a chance de acidentes ao reduzir o trânsito; encontrar destinos alternativos ou sugeridos automatizadamente - enfim, mapear a mobilidade de pessoas em seus automóveis como um caminho para a melhora da qualidade de vida em uma cidade.

Encontrar padrões a partir de eventos que permitam um melhor gerenciamento do fluxo de trânsito, em ambiente urbano, auxilia na construção e manutenção das Cidades Inteligentes. De acordo com Javidroozi, Shah e Feldman (2019), isso diz respeito a um conjunto de sistemas que usam dados em tempo próximo ao real, produzindo conhecimentos por diversos setores de uma cidade, dentre eles o setor de Transportes, apoiado por pesquisa acadêmica e certas técnicas computacionais.

Existem pesquisas de Mapeamento e de Revisão Sistemática da Literatura na área de predição de destinos e/ou de trajetórias como as de Junior, Dutra e Neto (2024), Graser et al. (2023) e Li et al. (2021), dentre outros trabalhos de pesquisa secundárias, que evidenciam os desafios no conhecimento das técnicas mais recentes na área de mobilidade urbana apoiadas por Tecnologia da Informação (TI), de forma clara, sistemática, auxiliando no reconhecimento de problemas e no desenvolvimento de soluções.

Assim, esse processo de revisar sistematicamente a literatura sobre técnicas preditivas de destinos ou de trajetórias auxiliou no que diz respeito a nortear esta pesquisa, uma vez que foi possível identificar quais técnicas estão sendo usadas para a construção de preditores de trajetórias. Além disso, se fez presente o desafio de se encontrar conjuntos de dados de locomoções reais, gratuitos, públicos, livres e de fácil acesso para predição de trajetórias e/ou destinos - que consistam em dados espaço-temporais, representativos da mobilidade humana em um período, em uma cidade, com automóveis - o que é difícil, devido a questões de privacidade.

Portanto, esta pesquisa visa realizar uma análise comparativa entre o uso de Cadeias de Markov e Cadeias Ocultas de Markov (HMM, do Inglês *Hidden Markov Chain*) em modelos de predição de destino, para uma base de dados de deslocamentos reais em um ambiente urbano, e obtidos de uma realidade cujo meio de transporte é o automóvel.

O motivo da escolha em se comparar modelos preditores com base em Cadeias de Markov e de HMM serão melhor detalhados nos próximos Capítulos. Porém, de forma breve, reside na característica de que um modelo com Cadeias de Markov realiza transição considerando apenas

o estado atual, sem ter que preservar um grande quantitativo de histórico (lugares já visitados).

Enquanto que, com relação ao uso da técnica, HMM, existe um diferencial em relação às Cadeias de Markov, uma vez que este tipo de modelo permite tratar tanto sobre eventos observados como sobre eventos escondidos, em que o primeiro é o que é visto, e o segundo são *tags* a serem pensadas como causas no modelo probabilístico (JURAFSKY; MARTIN, 2024). Ou seja, HMM inclui dados contextuais, o que pode ser relevante para predizer destinos com base não apenas na movimentação geográfica, mas também em dias da semana.

1.1 Motivação e Definição do Problema

Estudar os Transportes[1], no contexto das Cidades Inteligentes, é valioso no que diz respeito ao desenvolvimento das técnicas e métodos computacionais em se tratando de:

- Identificar padrões de movimentação a partir de dados espaço-temporais, para sugerir a criação de rotas menos congestionadas e/ou mais seguras, roteiros turísticos, roteiros para usuários com alguma necessidade especial, entre outras possibilidades;
- No âmbito científico, realizar comparações de técnicas no contexto de predição, que permitam trazer novas vantagens e conhecimento de desvantagens no uso de diferentes abordagens. Por exemplo: uso de Cadeias Ocultas de Markov seriam mais vantajosas que o de Cadeias de Markov?

Com base nessas premissas, tem-se, como primeira motivação, integrar as necessidades da sociedade com a mobilidade urbana, através de recomendações de locais para visitação; e, como segunda motivação, avaliar a relevância do uso de informações contextuais na predição, através da comparação inicial entre as duas técnicas markovianas.

Assim, a partir dessas técnicas (Cadeias de Markov e Cadeias Ocultas de Markov), úteis, conforme a literatura para predição de trajetórias e/ou de destinos, tem-se o seguinte problema de pesquisa:

- Dificuldade de encontrar trabalhos que comparem exclusivamente essas duas abordagens, no contexto de predição de destinos, e que realizem a comparação do desempenho de um modelo preditor com base em Cadeias de Markov e em HMM na análise de dados reais, com um método de organização dos dados útil a conjuntos de dados que possuam, no mínimo, campos com as coordenadas geográficas longitude e latitude e um ou mais campos referentes à informação temporal.

[1] Mais exatamente a modelagem matemática, com vias à implementação computacional, capaz de sustentar uma predição de destinos.

Sendo assim, a relevância desta proposta de pesquisa consiste, principalmente, para pesquisas aplicadas que usam Markov para predição de destinos de veículos urbanos individuais, geralmente de forma composta com outras técnicas, como é o caso, por exemplo, de trabalhos como os de Lassoued et al. (2017) e Amin et al. (2018), cujos trabalhos são descritos no Capítulo 3.

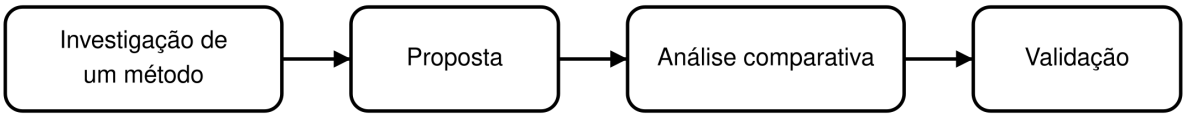
De forma clara, a Questão de Pesquisa a ser investigada é a seguinte: "Existe diferença quanto ao uso dos modelos de predição de destino com base em Cadeias de Markov e Cadeias Ocultas de Markov, no contexto de tráfego urbano e no uso de veículos individuais?".

Esse questionamento serve como base para as comparações com os resultados presentes na literatura, remetendo a um quadro comparativo final, que não servirá apenas para explorar uma possibilidade de uso dos dados, mas também abrindo caminho para futuros trabalhos com novas transformações desses dados.

O problema de pesquisa envolve a necessidade de uma análise comparativa relacionada à Questão de Pesquisa. A realização da análise comparativa deverá auxiliar na resolução do problema. Após a realização da comparação, a partir de uma base de dados reais, deve-se tornar mais claro um método de utilização de Cadeias de Markov e HMM com finalidade preditiva para destinos.

A Figura 1 sintetiza o fluxo de etapas desenvolvidas neste trabalho em um nível mais macro, iniciando-se com a busca e investigação de um método para trabalhar com os modelos markovianos, a elaboração da proposta, a realização da análise comparativa e a consequente validação dessa análise através de um teste paramétrico.

Figura 1 – Elaborações dos modelos preditivos.



Fonte: Elaboração própria.

1.2 Objetivos

1.2.1 Objetivo geral

Realizar uma análise comparativa dos modelos de predição de destino com base em Cadeias de Markov e Cadeias Ocultas de Markov, no contexto de usuários que usam seus veículos de maneira individual.

1.2.2 Objetivos específicos

- Analisar o contexto e a relevância no uso das técnicas escolhidas;
- Identificar a base de dados a ser usada para a análise comparativa;
- Definir os critérios para inclusão dos veículos e suas trajetórias na base de validação;
- Evidenciar a importância do balanceamento dos dados a partir da filtragem pela quantidade de repetições de pares Origem-Destino por veículo;
- Desenvolver e executar os modelos de predição com base em Cadeias de Markov e HMM.

1.3 Estrutura do Documento

Os capítulos subsequentes estão organizados da seguinte maneira:

- Após esta Introdução, os conceitos básicos deste trabalho são apresentado em detalhes no Capítulo 2, incluindo descrições e, quando necessário, com exemplos.
- No Capítulo 3 são apresentados os trabalhos relacionados, com a revisão da literatura.
- Quanto ao Capítulo 4, há a explicação sobre a Metodologia da Pesquisa, que inclui o Método desenvolvido como ferramenta para a análise comparativa.
- Finalmente, no Capítulo 5 são apresentados e discutidos os resultados desta pesquisa, com as técnicas avaliadas, os resultados em si, e a análise correspondente.
- As considerações finais e as propostas de continuação do trabalho são descritas no Capítulo 6.

1.4 Conclusão da Introdução

Em suma, conclui-se este capítulo com a explicação acerca da contextualização e problema de pesquisa, justificativa, escopo e objetivos. No próximo capítulo, será apresentada a fundamentação dos elementos norteadores e justificadores desta pesquisa.

2 FUNDAMENTAÇÃO TEÓRICA

Este estudo parte de conceitos básicos e de um entendimento sobre Cadeias de Markov e Cadeias Ocultas de Markov (HMM). Neste capítulo, serão apresentados os conceitos básicos dos objetos de estudo, que são referências para o método desenvolvido a partir do uso das referidas técnicas, seguindo-se sempre a pressuposição da propriedade markoviana.

2.1 Conceitos e Técnicas Básicas

Antes de serem avaliados os trabalhos relacionados, é preciso convencionar os termos padrão desta pesquisa, bem como as técnicas básicas.

2.1.1 Trajetórias e Tesselação com grades de Origem e de Destino

Definição 1 - Trajetória: consultando, primeiramente, Spaccapietra et al. (2008), tem-se que, em livre tradução, é o registro definido pelo usuário da evolução da posição (percebida como um ponto) de um objeto que se move no espaço durante um determinado intervalo de tempo, para atingir um determinado objetivo.

A partir desse conceito inicial, entende-se que uma trajetória representa uma movimentação espaço-temporal de algum usuário humano, animal ou evento, onde há uma sequência de pontos indicadores da posição do objeto em movimento no espaço com o *timestamp* em que o ponto foi coletado (SILVA; PETRY; BOGORNY, 2019). Para Li et al. (2021), considerando um objeto em movimento, é uma sequência finita, uma ordenação sequencial de pontos espaço-temporais, que é obtido por uma amostra de um rastreamento mediante algum dispositivo de localização.

Em síntese, trajetória é a representação de pontos em ordem cronológica, onde cada ponto P representa uma coordenada geoespacial e um *timestamp*[1], com o qual é possível extrair um conjunto de movimentos únicos que compartilham a propriedade de visitar a mesma sequência de lugares com tempos de viagem próximos, isso conforme é considerado em Silva et al. (2019), consultando parcialmente Giannoti et al. (2007).

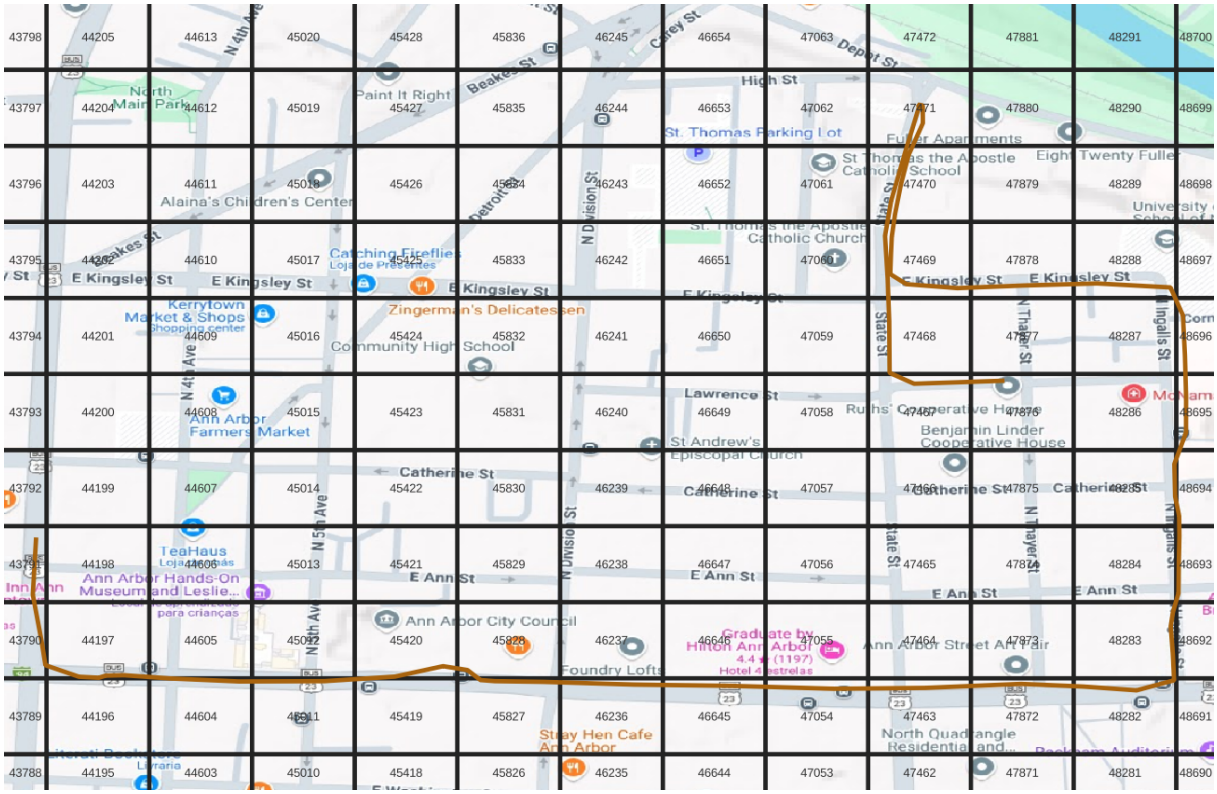
Definição 2 - Tesselação: o conceito de tesselação consiste em um modelo para melhor entender e processar o espaço geográfico (GOLD, 2016), basicamente segmentando uma área maior em partes menores. Há algumas configurações, como os diagramas de Voronoi, também conhecidos como Diagramas de Thyessen, sendo uma forma de particionamento do espaço em regiões (células) a partir de um número de pontos de observação existentes. Mais especificamente,

[1] Corresponde a um tipo de dado que representa o conjunto data, hora, minuto e segundos, e em parte dos casos, milissegundos e microssegundos. Esse conceito se torna mais claro em Tanimura (2022).

pode-se dizer que constituem o oposto dos centróides, que são a transformação de áreas em pontos (GAO, 2022). Neste trabalho, são utilizados retângulos de 100 m^2 para a criação das células. Conforme o citado autor (GOLD, 2016), em livre tradução: "é formado por células discretas com atributos próprios e, à medida que estão conectadas, com alguma representação de adjacência". Essa representação de adjacência é que fornece a noção de continuidade do fenômeno (a trajetória), seja de trechos que lhe são internos, até outras trajetórias.

A Tesselação é relevante para esta Pesquisa uma vez que ocorre a conversão dos pontos de origem e de destino para as regiões geográficas da Origem e do Destino. Isso pode ser ilustrado conforme na Figura 2 .

Figura 2 – Exemplo de tesselação de área com os rótulos representados por números, considerando Origem e Destino como qualquer dos pontos nos extremos da polilinha, desde que se refiram a posicionamentos diferentes entre eles.



Fonte: Elaboração própria.

Definição 3 - Origem e Grade de Origem: trata-se da primeira coordenada geográfica que representa o ponto de partida, com tempo zero, em que uma trajetória é considerada computável - para efeito de uso neste trabalho. Quanto à grade de origem, é uma área de interesse, com 100 metros quadrados, onde a Origem está inserida em seu centro, contendo o rótulo correspondente - capaz de substituir as coordenadas geográficas para efeito desta pesquisa. Por exemplo: um ponto de Origem pode ser convertido em grade de origem (célula), que é uma área de 100 metros quadrados, de onde partem as trajetórias com destino definido.

Definição 4 - Destino e Grade de Destino: trata-se da coordenada geográfica do destino de um usuário, numa determinada viagem, firmando-se como o término dessa viagem. Isso pode ser entendido conforme Zeng, Wang e He (2020), em que (...) "para uma trajetória de consulta $q = \{p_1, p_2, \ldots, p_m\}$, onde $1 \leq m \leq n-1$, a tarefa de predição de destino consiste em prever a localização (isto é, a longitude e a latitude) do ponto final p_n. ". Em resumo, é o par de coordenadas que indica o final da trajetória. Enquanto que a grade de destino é uma área de interesse, também com 100 metros quadrados, onde o Destino está inserido em seu centro, contendo o rótulo correspondente (um identificador numérico), substituindo as coordenadas geográficas.[2]

2.1.2 Predição ou Análise de Dados Preditiva

Definição 5 - Predição: a princípio, Kelleher, Namee e D'Arcy (2020) trazem o conceito de Análise de Dados Preditiva (para previsões), que corresponde à arte de construir e usar modelos que façam predições baseadas em padrões extraídos de dados históricos - no caso desta Pesquisa, de dados apenas de destinos. O ideal, neste caso, é que a predição não dependa somente de dados históricos, mas que consiga inferir para além deles.

2.1.3 Pares Origem-Destino

Definição 6 - Pares Origem-Destino: são as grades de início e fim das trajetórias, isto é, os pares de números que representam as grades. Para cada trajetória há dois pontos que correspondem aos limites de cada evento. Nesta Pesquisa, trabalha-se apenas com a predição dos rótulos de Destino ou, mais precisamente, com as áreas de 100 m^2 onde se situa um determinado Destino. Como exemplo: o par <150, 250> representa que a grade de origem está rotulada com o valor 150, enquanto o de destino com o rótulo 250.

2.1.4 Dados Espaciais

Definição 7 - Dados Geográficos: para Druck et al. (2004), esses dados - no contexto espacial - surgem no de representações computacionais que, a partir de um Sistema de Informação Geográfico (SIG), podem ser tratados (computacionalmente), armazenando a geometria e os atributos - mas que, no caso de dados geográficos, acrescenta-se um referenciamento espacial com base em um Sistema de Referência de Coordenadas (SRC) específico. Isso define um sistema de orientação e, mediante um código numérico, o EPSG (*European Petroleum Survey Group*), inclui o tipo de Projeção. No primeiro caso, trata-se da posição de um objeto ou fenômeno sobre a superfície da Terra; no segundo, a forma de representação cartográfica.

[2] Essas grades foram definidas no contexto deste trabalho. Ou seja, fazem parte de sua metodologia e são representadas por um identificador numérico, conforme cada geometria em forma de grade ou área.

Em relação aos dados, a representação ocorre pela presença dos campos de Longitude e de Latitude, que mais comumente se situa sob o código "EPSG:4326". Esse código é usado para fornecer posicionamento e navegação em escala global, sem ser apenas uma parcela regional.

2.1.5 Dados Contextuais ou Dados Temporais como Contexto

Definição 8 - Dados Temporais: são do tipo *datetime* ou *timestamp*, que se revelam mediante o formato de data acrescido de horas, minutos e segundos. Podem possuir um *timezone* específico (como o de uma porção do globo terrestre, uma zona) ou ser *Universal Time Coordinate* (UTC) ou *Global Time Greenwich Mean Time* (GMT), que são padrões internacionais. O primeiro foi inicialmente concebido no início dos anos 1960, para melhorar a disseminação de um sistema anterior, o UT1. Nele, é utilizado *Global Navigation Satellite Systems* (GNSS), de acordo com Arias e Guinot (2004). O segundo significa o tempo solar, de acordo com o Observatório Real de Greenwich, sendo este mais antigo que o UTC, mas que não deve ser utilizado para propósitos mais precisos, conforme Weintrit (2017). Os propósitos mais precisos podem ser deduzidos como aqueles que envolvam navegação de aviões, embarcações marítimas, veículos em geral.

Em relação aos dados contextuais, são aqueles dados comuns ao contexto onde os usuários estão inseridos, e dizem respeito ao local em que se situam os eventos, tais como: dados de clima, dados topográficos (ou seja, em relação ao terreno), sinais de trânsito, dentre outros. No trabalho de Liu et al. (2019), por exemplo, os autores consideram o número de táxis solicitado, a demanda de viagens (o *Local Spatial Context*), se todos os distritos são residenciais (que os autores chamam de *Global Relational Context*) e dados meteorológicos (que os autores chamam de *Temporal Evolution Context*).

2.1.6 Propriedade Markoviana

Em uma distribuição de probabilidade, o surgimento do dado atual irá influenciar apenas no dado imediatamente seguinte - cessando os seus efeitos para os dados subsequentes. Ou seja, essa distribuição é um processo markoviano, que somente existe “se o estado da variável aleatória na próxima instância de tempo depende apenas do resultado da variável aleatória no momento atual” (ANKUR; PANDA, 2018).

Uma vez entendida essa propriedade, os mesmos autores Ankur e Panda (2018) elencam suas principais categorias, das quais esta pesquisa pretende se concentrar nas que atuam, para Cadeias Ocultas de Markov, com um tempo constante em relação às mudanças de estado; e para HMM, com um tempo que é referido como parte de um contexto dessas mudanças de estado.

2.1.7 Cadeias de Markov e Destinos

Essa é uma técnica que permite predizer a probabilidade de mudança de estado, dada uma origem, até um destino. Isso pode ser interpretado através da leitura do trabalho de Grewal,

Krzywinski e Altman (2019a)[3], surgindo como um modelo onde os dados são estocásticos e de memória curta, onde, para se atingir uma predição, um modelo com Cadeias de Markov necessita de uma matriz de probabilidade de transição, onde figuram os estados na condição de repetidos - e os estados únicos, em si mesmos, sem as repetições.

Dessa forma, se um usuário esteve N vezes em determinadas áreas de interesse com o rótulo "lar", e se muito frequentemente a área de interesse posterior é "parque", a tendência é que esse padrão se repita, dada uma probabilidade.

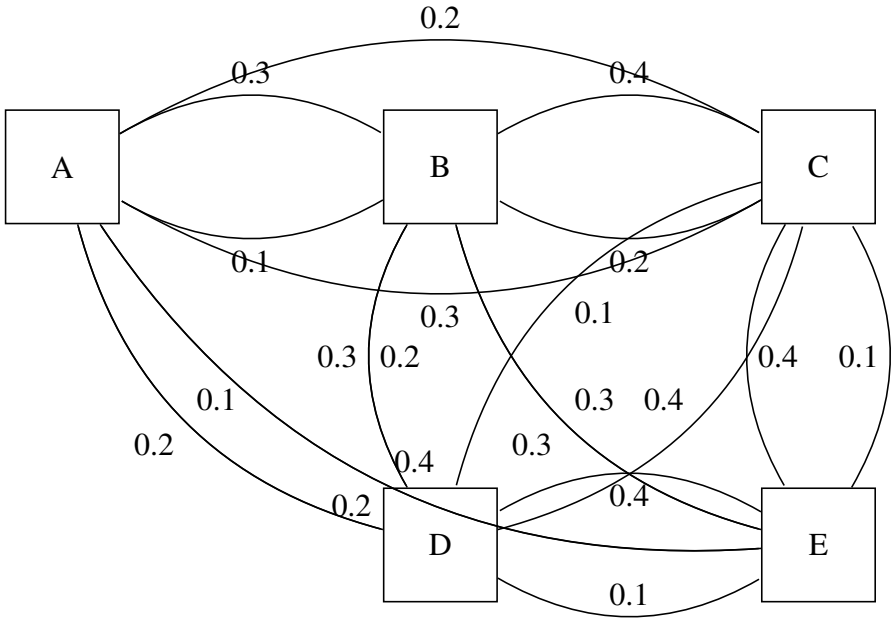
Outro exemplo: é possível imaginar que existem 5 estados: A, B, C, D e E (lembrando que A poderia ser lar e B, parque), que podem ser Origens e Destinos numa subtrajetória (entre 5 estados, 4 subtrajetórias). De A para A, pode-se dizer que há 0% de probabilidade de um indivíduo, objeto ou evento permanecer no mesmo lugar com o passar do tempo. De A diretamente para B, a probabilidade é de 0,3; de B até C, é de 0,2; de C até D, é de 0,1; e de D até E, é de 0,4 — sempre somando 1, ou $0{,}3 + 0{,}2 + 0{,}1 + 0{,}4$.

Na Figura 3, o gráfico apresenta sobreposições, mas, ainda assim, pela Tabela 1, é possível entender, com as células preenchidas, observando os valores indicando a probabilidade de transicionamento de um estado para o outro. Considerando-se, inclusive, que cada letra desses estados sejam coordenadas geográficas discretizadas (convertidas em valores textuais representando números) e, através da Tesselação, transformadas em áreas de interesse, sempre regulares. Pressupõe-se, então, não haver nenhum fator oculto a interferir nos valores de probabilidade de transição.

Em síntese, e com base em Grewal, Krzywinski e Altman (2019b), exuma Cadeia de Markov envolve uma matriz de probabilidades de transição junto com os estados. Assim, toda passagem, de um estado até outro, passa por uma progressão contínua até um destino final, que encerra essa progressão. Dessa forma, no gráfico da Figura 3 , têm-se os 5 estados ou nós de uma rede, e os arcos indicando probabiblidades em decimais. Enquanto que na Tabela 1, a Matriz de Probabilidade de Transição com dados fictícios e correspondentes ao gráfico ilustrado anteriormente.

[3] Os autores explicam que uma Cadeia de Markov envolve uma matriz de probabilidades de transição junto com os estados. Assim, toda passagem, de um estado até outro, passa por uma progressão contínua até um destino final, que encerra essa progressão.

Figura 3 – Ilustração de uma Cadeia de Markov.



Fonte: Elaboração própria.

Tabela 1 – Matriz de Transição de uma Cadeia de Markov.

Estado	**A**	**B**	**C**	**D**	**E**
A	0.0	0.3	0.2	0.1	0.4
B	0.1	0.0	0.4	0.2	0.3
C	0.3	0.2	0.0	0.1	0.4
D	0.2	0.3	0.4	0.0	0.1
E	0.2	0.3	0.1	0.4	0.0

Fonte: Elaboração própria.

Importante também está em ressaltar, antecipadamente, sobre a escolha de um limiar de corte para filtragem de amostras, no caso de Cadeias de Markov, durante a preparação dos dados finais para utilização desse algoritmo. Um modelo com Cadeias de Markov, assim, pode envolver um balanceamento entre a retenção de dados suficientes e a confiabilidade estatística das estimativas, evitando tanto o excesso quanto uma perda relevante de informação.

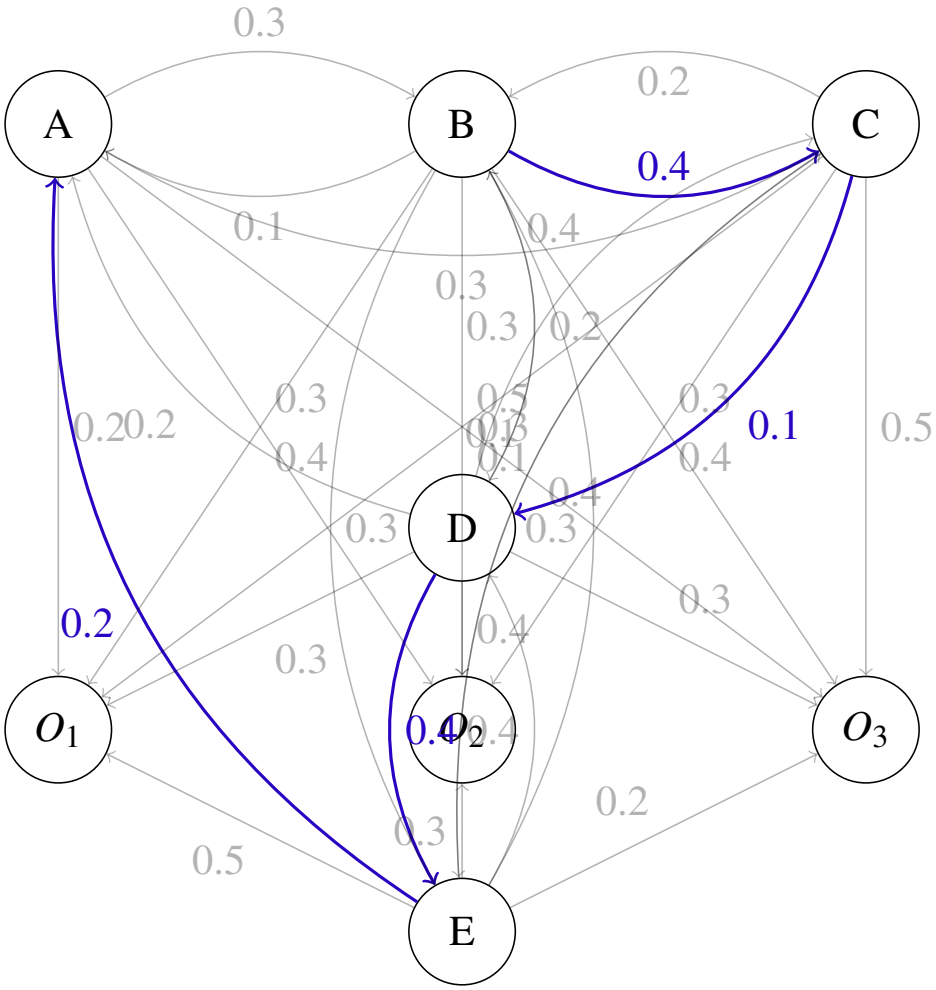
2.1.8 Cadeias Ocultas de Markov e Destinos

Esta seção abrange não apenas o que são Cadeias Ocultas de Markov e sua relação com os Destinos, mas também do que se trata o algoritmo de Viterbi e o conceito de "ocultos"ou "escondidos", presente no nome do model.

Assim, inicialmente, considerando que haja fatores ocultos, além daqueles incluídos numa matriz de probabilidade de transição, também com suas próprias probabilidades - chamadas

"de emissão- envolvidas, é possível exemplificar um caminho ou viagem pelo conjunto de arcos (ou linhas) em azul (na Figura 4), como se fossem subtrajetórias desde a Origem 'B' até o último Destino 'A'.

Figura 4 – Desenho de um modelo oculto de Markov com estados ocultos e observáveis, estando, em azul, uma viagem hipotética partindo do ponto B.



Fonte: Elaboração própria.

Observa-se que os Estados são A, B, C, D e E, seguindo, a princípio, mesma lógica de uma Cadeia de Markov. Porém, a palavra 'oculta' decorre dos estados Ocultos, além dos Símbolos associados, que indicam, mediante, por exemplo, "dia da semana"em que ocorreram as presenças nesses 5 estados (lugares), com suas probabilidades de ocorrência. Isto é, se há a informação de que um usuário possui 40 % de chances de sair do seu lar para um parque (estado) dado 20% de chance de ser às sextas-feiras (símbolo), qual seria a probabilidade de visitar uma sorveteria? Observe que o usuário visita um parque numa probabilidade vinculada ao fato de ser sexta-feira. Isso pode trazer variações, porém aqui atem-se apenas a esse paralelo entre parque e sexta-feira. Respostas podem ser encontradas para outras sexta-feiras ou para outras visitas a um parque, sem ser nesse dia. Nesta pesquisa, não há esse nível de complexidade ou variação

entre diferentes destinos e dias, apenas se pretende entender qual a probabilidade de se visitar novamente a sorveteria dado ser uma sexta-feira.

Lembrando que o conjunto de repetições de dias da semana é representado numa segunda matriz, a de emissão, que são, conforme na Figura 4 , "01", "02"e "03"(ou poderiam ser Segunda-Feira, Quarta-Feira e Sexta-Feira), localizados entre D e E, de cima para baixo, com algumas sobreposições.

Ou seja, mantendo-se a matriz de transição e os estados, há mais duas categorias que surgem para a predição de destinos a partir de Modelos Ocultos de Markov (ou HMM): as observações (os símbolos) e as probabilidades de emissão (o mesmo que na matriz de transição, mas cada estado será o estado oculto). Isso serve para - segundo Grewal, Krzywinski e Altman (2019a) - detectar com base em eventos não observados, como dia da semana e, até, se fosse o caso deste trabalho, o período do dia[4].

Dessa forma, tem-se a oportunidade de explicar como ocorre a decodificação do modelo: com o uso do algoritmo de Viterbi. A partir, por exemplo, da pergunta hipotética “Dado um conjunto de origens de treinamento influenciados indiretamente pelo dia, quais as probabilidades para um conjunto de rótulos de teste mais provável, confirmando um destino?”. A resposta é uma série de valores, após o uso do algoritmo citado anteriormente.

Entende-se, então, que o rótulo final, com base num conjunto de rótulos indicando os estados de origem, que possua maior frequência comparando o modelo gerado a partir dos dados de treino com um conjunto de dados de teste, será aquele com a maior probabilidade de ocorrer efetivamente. Afinal, as matrizes (de transição e de emissão) e os estados e símbolos únicos surgem na etapa de preparação, para comporem o modelo HMM, que é decodificado com Viterbi - que serve para localizar a sequência mais provável de estados ocultos, e cuja explicação mais detalhada se encontra em Rabiner (1989).

Ou seja, os estados ocultos são os rótulos, que, tal como em Cadeias de Markov, geram a matriz de transição, e os símbolos são o dia da semana (que, neste caso, geram a matriz de emissão). Isso significa que dos problemas fundamentais que um Modelo Oculto de Markov busca resolver, para este caso a solução de um deles é suficiente para esta pesquisa: encontrar a probabilidade maior de uma sequência de rótulos, indicando a maior repetição do rótulo determinador do destino como sendo o rótulo previsto. Para efeito desta pesquisa, a localização é entendida como os rótulos que indicam o "oculto"ou "escondido"do modelo HMM, enquanto que as variáveis visíveis ou diretas, são os dias da semana associados aos rótulos das grades de origem.

Assim, observam-se as mudanças de estado e de símbolos, nas tabelas, abaixo da Figura 4 e, nas Tabelas 2 e 3, respectivamente, as mudanças de estados e de símbolos.

[4] Evidentemente, um dia da semana pode ser uma informação muito comum. Seria possível utilizar alguma informação mais próxima da realidade do usuário.

Tabela 2 – Matriz de Transição de uma Cadeia Oculta de Markov, com a marcação das células que representam uma trajetória fictícia afetadas pelos estados visíveis.

Estado	A	B	C	D	E
A	0.0	0.3	0.2	0.1	0.4
B	0.1	0.0	0.4	0.2	0.3
C	0.3	0.2	0.0	0.1	0.4
D	0.2	0.3	0.4	0.0	0.1
E	0.2	0.3	0.1	0.4	0.0

Fonte: Elaboração própria.

Tabela 3 – Matriz de Emissão de uma Cadeia Oculta de Markov, com a marcação das células que representam uma trajetória fictícia afetadas pelos estados ocultos.

Estado	O1	O2	O3
A	0.2	0.3	0.5
B	0.4	0.3	0.3
C	0.1	0.4	0.5
D	0.3	0.4	0.3
E	0.5	0.3	0.2

Fonte: Elaboração própria.

Uma explicação mais clara, sobre o que são os rótulos ocultos e os rótulos visíveis, pode ser vista conforme na Tabela 4.

Tabela 4 – Componentes de um Modelo Oculto de Markov.

Componente	Descrição
Estados Ocultos	Variáveis não observáveis que representam o estado real do sistema
Matriz Probabilidade de Transição	A matriz com a probabilidade de mudança de estado
Símbolos	Observações visíveis emitidas pelos estados ocultos
Matriz Probabilidade de Emissão	A matriz com a probabilidade de mudança dos símbolos por estado

Fonte: Elaboração própria.

Em síntese, para finalizar este capítulo, uma Cadeia de Markov se constitui de uma matriz de transição de probabilidade que represente as repetições dos estados a serem integradas a cada um dos estados; enquanto uma Cadeia Oculta de Markov, além disso, considera os símbolos e uma matriz de probabilidade similar, porém de emissão. Assim, há uma integração entre matrizes

de transição e de emissão com os estados ocultos em particular e os símbolos, também em particular, respectivamente, na construção do modelo de cadeias ocultas[5].

E, em relação a Modelos Ocultos de Markov (HMM), também é importante nesta pesquisa considerar critérios adequados para filtragem de amostras durante a preparação dos dados.Esta configuração permite um equilíbrio entre a complexidade do modelo e a disponibilidade de dados, evitando tanto a subidentificação de parâmetros quanto o descarte desnecessário de sequências informativas.

2.1.9 Técnica de balanceamento de dados

Trata-se da busca, quando necessária, de uma técnica para balancear os dados de acordo com cada classe ou veículo.

Uma técnica possível considera os trabalhos de Anderson e Goodman (1957) para Cadeias de Markov, e Rabiner (1989) para HMM, e da penalização para adequação entre essas duas técnicas, dentre as opções escolhidas arbitrariamente, a filtragem foi feita com o valor de 50 trajetórias por veículo - não arbitrariamente, mas mediante o que foi definido através da função presente no Algoritmo 2. Além, claro, disso depender do conjunto de dados e da existência de certa quantidade de valores únicos dos rótulos das grades de destino.

Assim, os trechos relevantes para Cadeias de Markov, HMM e a penalização são pragmaticamente formuladas (mas, mediante bases estatísticas) e de acordo com os autores consultados e anteriormente citados, da seguinte forma:

- Critérios de adequação:

$$\text{adequacao\_markov} = \min\left(1.0, \frac{\text{limiar}}{10 \cdot \text{estados\_unicos}}\right) \tag{1}$$

$$\text{params\_hmm} = \text{n\_estados\_hmm}^2 + \text{n\_estados\_hmm} \cdot \text{estados\_unicos} \tag{2}$$

$$\text{adequacao\_hmm} = \min\left(1.0, \frac{\text{limiar}}{5 \cdot \text{params\_hmm}}\right) \tag{3}$$

- Pontuação combinada:

$$\text{pontuacao} = 0.5 \cdot \text{adequacao\_markov} + 0.5 \cdot \text{adequacao\_hmm} \tag{4}$$

- Penalizar perda de dados:

$$\text{fator\_veiculos} = \frac{\text{n\_veiculos}}{\text{len}(\text{df}[\text{coluna\_veiculo}].\text{unique}())} \tag{5}$$

$$\text{pontuacao\_final} = \text{pontuacao} \cdot (0.7 + 0.3 \cdot \text{fator\_veiculos}) \tag{6}$$

[5] Sublinhe-se que, no contexto deste trabalho, o objetivo é a utilização de um Modelo Oculto de Markov, através de HMM, numa etapa posterior ao uso de uma Cadeia simples, para comparação. E, para essa comparação, médias de precisões que dependem da realização de predição para entender o padrão Origem-Destino de um conjunto de usuários individuais representado por identificadores de veículos.

2.1.10 Subtrajetórias

Definição 9 - Subtrajetórias: o conceito é contextualizado, para o que Sun et al. (2024) se referem ao campo campo da segmentação de dados de trajetórias, - essa segmentação (geradora de partições de trajetórias) é definida como oriunda de algum dos quatro tipos de classificadores como os fundamentados em recursos, os em interpolação e os em agrupamentos. O *stay-point based segmentation* leva a supor subtrajetórias, que, a partir do particionamento de trajetórias em trechos menores, mediante pontos de parada. De forma mais clara, há:

> *Dados dois parâmetros de distância especificados pelo usuário D e parâmetro de tempo T, um ponto de permanência* $SP = \{pa \rightarrow pa+1 \rightarrow \ldots \rightarrow pb\}$ *é uma subtrajetória, satisfazendo as duas restrições a seguir: 1. restrição de tempo, ou seja,* $pb.t - pa.t \geq T$*; 2. restrição de distância, ou seja,* $d(pa, pi) \leq D$ *mas* $d(pa, pb+1) > D$ *para todo* $a < i \leq b$*, onde* $d(*,*)$ *é a distância de dois pontos GPS.*

Esse efeito é obtido através do método *Stop Splitting* da biblioteca *MovingPandas* da linguagem *Python*, que, através dos parâmetros de entrada (máximo diâmetro da parada, tempo mínimo da parada e comprimento mínimo das subtrajetórias retornadas como saída) sustenta-se pela abstração do caminho que seria percorrido se a distância e o tempo parados fossem utilizados para seguir o mesmo padrão do trajeto interrompido, de acordo com interpretação do conceito mediante a realização de experimentos.

2.2 Funcionamento dos Modelos de Predição de Trajetórias

Para efeito nesta Pesquisa, os modelos preditivos de Cadeias de Markov e HMM necessitam de dados de treinamento e de teste. Treinamento é o processo realizado para a construção dos modelos. São dados - nunca iguais aos de Teste - inseridos para que o modelo entenda como processar - são exemplos para generalização ou abstração. Uma vez que, de forma supervisionada, esse modelo entenda os dados, é possível executar a etapa de Teste, em que ocorre a previsão do destino (previsão da grade de destino a partir das grades de origem e de destino), sendo esta previsão comparada com a grade de destino real. Em caso de igualdade, considera-se que o modelo obteve um acerto (verdadeiro positivo), enquanto que uma diferença no grid de destino previsto com o destino real configura um erro (falso positivo). As diferenças nessas comparações são utilizadas para obtenção da estatística de Precisão.

Assim, treinamento envolve criar modelos preditivos que aprendam conforme as transições (Markov) ou que considere informações contextuais (HMM), a exemplo de dia da semana concatenado com o identificador da grade de origem. O modelo preditivo usa dados de deslocamentos já realizados (dados históricos). Enquanto o teste envolve, para ambos os modelos, a

situação em que dada uma origem deve-se predizer um destino. Markov usa somente o estado atual, enquanto HMM considera também o contexto (emissão, isto é, nesta Pesquisa, dia da semana e grade de origem concatenados).

Os modelos e seus componentes podem ser entendidos como artefatos, cuja materialização depende de um conjunto de etapas capaz de gerar resultados para, nesta pesquisa, a elaboração de um quadro que permita um entendimento claro para outros pesquisadores das diferenças entre as técnicas, no que diz respeito ao uso de dados com a estrutura da base de dados VED. Pressupõe-se, para o uso desses modelos neste contexto de trabalho, que são para veículos usados de forma individualizada - com dados anonimizados.

Deve-se observar também que o MSL realizado por Junior, Dutra e Neto (2024) trouxe uma ênfase em trabalhos com técnicas personalizadas de modelos de predição de destinos (uso de Markov com outros algoritmos, ou outras formas de composição), porém, ainda assim, houve casos com técnicas de Markov envolvidas isoladamente (como no caso de Araújo et al. (2019), ainda que modificando o modelo), o que traz a necessidade de evidenciar esse uso a partir de cenários simplificados. Além disso, mais adequação à natureza do comportamento humano é obtida dos dados, uma vez que cada elemento é dependente um do outro tendo por limite o seu estado imediatamente posterior, e não a totalidade dos estados. Isto é, o comportamento humano pode mudar a cada subtrajetória. O planejamento para a ida a um determinado destino pode passar por caminhos diferentes - esses, inesperados - ou mesmo haver uma mudança de ideia, necessitando-se recalcular a rota.

Além disso, no caso de HMM, também há uma consideração sobre o Tempo: este segue sequencialmente, para cada identificador de subtrajetória, tornando mais adequado uma técnica que considere sequências sucessivas.

De acordo com Chen e Hong (2012), depreende-se que, em geral, para estudos financeiros e econômicos, a propriedade de Markov é uma propriedade fundamental na análise de séries temporais e frequentemente utilizada com essas finalidades. A diferença está na utilização com dados geográficos, para a modelagem da movimentação humana. Essas referências indicam que há uma facilidade no uso de Cadeias de Markov - Ocultas ou não - com sequências de observações ao longo do tempo, compondo uma movimentação desde uma Origem até um Destino, o seu poder explicativo aumenta - principalmente considerando o aspecto imprevisível do comportamento humano.

Outro fator diz respeito ao poder explicativo de Markov mediante a analogia a uma distribuição de probabilidades percebida através de Ankur e Panda (2018), apresentando, conforme interpretação dessa obra, a um entendimento mais intuitivo, de caixa-branca, das técnicas preditivas - ao invés da dependência quanto à processamentos caixa-preta.

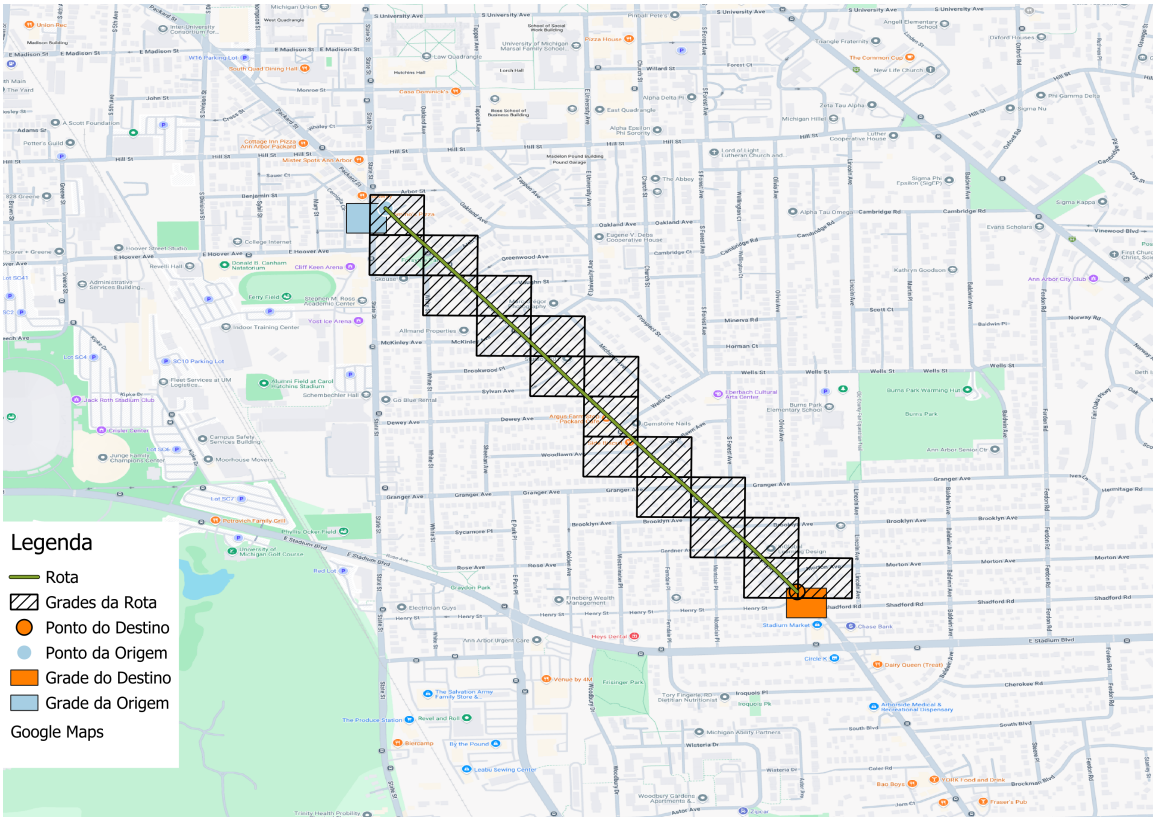
Em síntese, pode-se aqui entender Explicabilidade, para efeito desta pesquisa, como a propriedade de uma escolha de técnica de modelagem preditiva que permita uma maior

proximidade com um grau de objetividade científica, sendo um método que não dependa de um aparente acaso - a partir de um determinado nível de complexidade ou especificidade, como ocorreria com técnicas de Redes Neurais Profundas, conforme pode ser observado nas explicações sobre elas presentes em Kelleher, Namee e D'Arcy (2020).

Assim, a escolha por modelos markovianos apresenta as seguintes propriedades correspondentes a motivações: possuem clareza com o aspecto temporal dos dados do VED (clareza e adequação com o tipo dos dados); possuem fundamento no Mapeamento Sistemático da Literatura (fundamento na literatura científica, por ser utilizada); simplicidade de implementação, se comparada a Redes Neurais Profundas.

Mas, antes de aplicar os modelos com base em Cadeias de Markov e HMM, é preciso ilustrar a lógica do processo, mediante a Figura 5.

Figura 5 – Ilustração do funcionamento dos dois modelos.



Fonte: Elaboração própria.

Na Figura 5, tem-se uma ilustração simplificada, por ser resumida a um único par Origem-Destino, em que, dada uma origem em azul e um destino em laranja, têm-se uma trajetória representada espacialmente como uma rota.

Sua direção é para Sudeste e há diversos retângulos tracejados, com a mesma área, que indicam os estados para a Matriz de Probabilidade de Transição para Markov, cada um com uma probabilidade de transição. Além disso, para HMM, simultaneamente, pode-se considerar

que cada retângulo possui os estados de emissão, com a Matriz de Probabilidade de Emissão (contendo referência aos dias da semana para cada estado).

Em resumo: cada retângulo é um estado único, seja para Transição (lugar) ou Emissão (dia da semana e lugar de origem). Eles, enquanto objetos únicos, são considerados como estados e/ou símbolos; e contêm probabilidades para mudança de estado, formando, quando ordenados de cima para baixo e da esquerda para a direita, exatamente uma matriz (de transição ou de emissão).

2.3 Conclusão da Fundamentação Teórica

Conclui-se, então, este capítulo, com os conceitos embasadores de toda esta pesquisa, bem como com a explicação sobre o funcionamento dos modelos e da validação estatística.

Dessa forma, a Fundamentação Teórica serve de apoio para que haja um entendimento mais claro sobre os elementos e processos que subsidiam todo o trabalho, cuja sequência de tarefas, dos dados brutos com os períodos de tempo que formam trajetórias até as modelagens com Cadeias de Markov e HMM, seguem uma interpreção a partir da leitura de Hornsby e Cole (2007), em que é possível entender determinados dados mediante os seguintes atributos-chaves: identidade do objeto (um identificador de viagem e/ou de viajante), descrição do evento (algo como o tempo, mas que também podem incluir outros elementos de contexto) e a localização (como longitude e latitude organizados em campos).

O próximo Capítulo trará os trabalhos similares a estes, tencionando demonstrar os que se relacionam a este trabalho.

3 REVISÃO DA LITERATURA

A escolha do tema de pesquisa deste trabalho foi fundamentada em um Mapeamento Sistemático da Literatura (MSL), por Junior, Dutra e Neto (2024), que obteve um quadro comparativo de 33 artigos selecionados sobre o uso de modelos preditivos para trajetórias ou destinos, com base em uma adaptação da metodologia estabelecida por Kitchenham e Charters (2007).

A partir dessa metodologia adaptada à realidade do estudo, o protocolo seguiu as etapas:

- Etapa de Planejamento:

  Durante o planejamento, ficou definida a busca na Association for Computing Machinery (ACM), Digital Library, Institute of Electrical and Eletronics (IEEE), Xplore, Tandfonline, GEOINFO, IJCAI-17 e a ferramenta de busca da Sociedade Brasileira de Computação (SBC). Os estudos foram considerados de 2017 a 2023.

  Os estudos-alvo desses artigos existem a partir de um objetivo (predições de destinos ou de trajetórias - ou de ambos); mediante algoritmos personalizados, de Aprendizado Profundo, de Agrupamento ou dentro do escopo dos métodos de Markov.

- Etapa de definição dos termos de busca, após refinamento: “routes”, “route prediction”, “routes” e “route prediction”, para os artigos da SBC. Para outras bases, porém, o termo foi "trajectory prediction"OR "destination prediction"OR "route planning"AND "roads". Ou seja, no início os termos eram genéricos, por ainda estar sendo necessário entender o que buscar, para só depois se tornarem, conforme cada retorno das buscas, algo mais consolidado.

- Etapa de seleção dos trabalhos: foram filtrados 794 artigos nas buscas e, assim, definida a leitura de títulos e resumos.

- Etapa de filtragem: com a leitura de títulos e resumos, além do uso de critérios de exclusão e de inclusão, 80 artigos foram filtrados. Os critérios de inclusão foram escolher artigos em Inglês ou em Portugês; serem sobre estudos completos, finalizados; artigos publicados num intervalo de no máximo 5 anos até o momento do Mapeamento; abordando uso de meios de transporte urbanos como bicicletas, carros, ônibus e caminhadas. Enquanto que os critérios de exclusão foram uso de abordagens apenas indiretas em relação ao propósito de predição de destinos ou de trajetórias, tais como predição de velocidade ou de outros atributos sem envolver posicionamento geográfico, ou puramente sobre planejamento de rotas; estudos fora do tópico tal como redes e Internet; referentes a veículos autônomos ou visão computacional.

- Etapa de seleção dos trabalhos: seleção final de 33 artigos, após a leitura completa. Essa leitura revelou mais detalhadamente que alguns artigos que, aparentemente não estavam situados nos critérios de exclusão, na realidade estavam nesses critérios ou pareciam ambíguos. Além da inclusão de um critério de qualidade: uso de *datasets* reais para a validação dos modelos; acesso direto aos *scripts* e dados. Para, assim, o Mapeamento ser organizado numa planilha.
- Publicação dos resultados: escrita textual e revisão foram necessários.

Ou seja, um processo foi utilizado, permitindo uma análise do MSL delineando predições com técnicas de agrupamento, como é o caso de Besse et al. (2018) e os agrupamentos hierárquicos; redes neurais profundas, como é o caso de Shen et al. (2023); e modelos markovianos, como é o caso de Araújo et al. (2019); ou um conjunto de algoritmos integrando diferentes modelos, como é o caso de Zhang et al. (2018), que integra filtro de Kalman com Large Short Term Memory (LSTM). Entretanto, percebeu-se uma falta de foco em técnicas Markovianas puras, ou seja, sem estarem associadas a outras técnicas, mais recentes - além da ausência de uma análise mais básica sobre o quanto e de que forma o contexto interfere na precisão de um modelo, fazendo-se pensar na necessidade de mitigar isso.

3.1 Análise dos artigos do Mapeamento Sistemático da Literatura

Na Tabela 5 tem-se uma adaptação da tabela que consta no artigo do MSL, com os campos referentes aos autores, se utilizaram mais que um cenário (no que diz respeito à origem dos dados) sendo S para Sim e N para Não; tipo de predição: coletiva (C), individual (I) ou ambas (A); os algoritmos ou os modelos utilizados descritos; o contexto dos dados, urbano (U) ou de outros tipos (O);

Tabela 5 – Artigos que utilizaram dados contextuais.

Autores	**Mais de um Cenário**	**Tipo de Predição**	**Contexto dos Dados**	**Algoritmos/Modelos Utilizados**
Wang et al., 2017	S	C	U	Algoritmos de treinamento e previsão
Imai et al., 2018	N	I	U	Algoritmos de agrupamento
Vahedian et al., 2017	N	C	U	Algoritmos de aprendizagem e agrupamento
Sadri et al., 2018	S	I	U	PreHeat e o "TrAf"
Chen et al., 2019	N	C	U	LSTM
Ma e Xie, 2021	N	C	U	FCM (Fuzzy C-means) e LSTM
Fu e Lee, 2020	S	I	U	RNN e Gradiente Descendente
Barth et al., 2020	S	A	U	PPTS e OPTS
Liu et al., 2019	N	C	U	Algoritmos de Rede Neural Profunda
Tang et al., 2021	N	C	U	Descoberta de Grupo e P-PPM
Liang e Zhao, 2022	S	I	U	Geração de Trajetória e LSTM
Dai et al., 2019	N	I	O	LSTM
Rainbow et al., 2021	S	I	U	Algoritmo de Redes Neurais Profundas
Besse et al., 2018	S	C	U	Agrupamento Hierárquico
Jiang et al., 2022	N	I	U	Assistente de Decisão e Rebalanceamento
Fan e Yao, 2017	N	I	U	Pontos fixos e RBLS
Ning, 2021	N	C	U	LSTM
Zhang et al., 2018	N	I	U	ESN, LSTM e Filtro de Kalman
Wu et al., 2020	N	I	U	Algoritmo de Previsão de Trajetória
Ebel et al., 2020	S	I	O	LSTM
Lassoued, 2017	S	C	O	Previsão de Cluster
Qiao et al., 2018	S	C	O	PrefixTP
Bhuvaneswari et al., 2017	S	C	O	Algoritmo SAHDID
Selvaraj et al., 2021	S	I	U	LSTM
Choi et al., 2019	S	I	U	FFNN
Yuan e Li, 2019	N	I	O	DISON
Tong et al., 2021	N	I	U	Dijkstra, Expansão baseada em Tabu e Ambição
Ren et al., 2022	S	I	U	DBSCAN, BI e K-means
Santana e Campos, 2017	N	I	O	Agrupamento de pontos
Araújo et al., 2019	S	C	U	TEMMUS
Qin et al., 2023	N	C	U	UTA
Schen et al., 2023	S	I	U	STI-GCN com GRU e CNN
Wang et al., 2023	S	I	U	Lane Transformer

Fonte: Junior, Dutra e Neto (2024).

Dentre os trabalhos que se destacaram sobre técnicas com Markov, há o trabalho de Froehlich e Krumm (2015), o de Lassoued et al. (2017) e o de Araújo et al. (2019). Em comum, os três abordam técnicas markovianas, sendo que o primeiro enriquece com o entendimento da importância das trajetórias repetidas para melhores acurácias; o segundo, com a realização da clusterização antes de Markov, demonstrando a possibilidade de utilizar técnicas de agrupamento e predição com Markov; e o terceiro, com predições baseadas no dia da semana e na posição do objeto em movimento, baseando-se no nível de Entropia.

Assim, houve a utilização de conjuntos de dados reais (como o *Simulation of Urban MObility* no caso de Lassoued et al. (2017)) e, no caso de Araújo et al. (2019), com dados de Tóquio. Além disso, o Froehlich e Krumm (2015) utilizaram um conjunto de dados com 4468

viagens de 252 motoristas.

Nas seções seguintes, serão apresentados os trabalhos mais recentes que se utilizaram de técnicas com Markov, além dos trabalhos que se utilizaram de dados semânticos - esses últimos na condição de considerar que há pesquisas na área que se utilizam de informações não apenas o contexto espaço-temporal.

3.1.1 Artigos com Cadeias de Markov

Normalmente, artigos envolvendo dados de trajetórias e/ou de destinos, dentre aqueles explorados, atuam na composição de Cadeias de Markov com outros modelos. O primeiro caso pode ser evidenciado por Ling et al. (2010). Dessa forma, o propósito consiste em apresentar um sistema onde a rota pessoal de um usuário é predita utilizando um modelo probabilístico fundamentado sobre dados históricos de trajetória. Os autores usam um modelo de Markov numa das etapas da pesquisa, no momento em que consideram a diferença entre o método deles e com um modelo de Markov, onde há uma variável (um estado) igual à tupla contendo a região de interesse atual e a próxima região de interesse (essas regiões como equivalentes às células da Tesselação). A matriz de probabilidade (de transição) é atualizada - porém - não somente com a atual região de interesse, mas também com padrões de rotas que contêm as regiões de interesse fornecidas. Interpreta-se que há uma atualização entre a região de interesse atual com a iteração das regiões de interesse que são encontradas.

Uma parte desses autores, em outro artigo (LING; LV; GENCAI, 2010), consideraram tanto o uso de validação cruzada a partir de 10 subconjuntos dos dados, onde 90% foram dados de treino e 10% foram dados de teste, como uma comparação entre a técnica deles e um modelo básico de Markov, além de Cadeias de Markov de segunda ordem (que não é abordado no presente trabalho). O restante do processo foi próximo ao do artigo explicado anteriormente.

Outro artigo foi o de Zhou et al. (2020), que trabalham nessa pesquisa com o problema onde "dado um lugar de origem, um lugar atual e uma trajetória conectando-os, encontrar o destino final da jornada". Na pesquisa, eles afirmam quebrar a estrutura da matriz de transição direcionando-se aos itens essenciais para as transições eficientes entre células de um mapa, com ou sem desvios. Afirmam também que, consequentemente, podem atualizar a proporção entre as atualizações para probabilidades de transição de forma a atender às exigências do tráfego, em constante mudança - mantendo o custo dessas variações o tão baixo quanto possível.

É importante salientar que esses últimos pesquisadores referidos aqui utilizaram dados sintéticos, e utilizaram para a avaliação estatística precisão, eficiência e cobertura.

Quanto à Araújo et al. (2019), utilizaram, na proposta de um preditor de mobilidade, um modelo com base em uma versão temporal de Markov com similaridade de usuário. Ou seja, uma versão modificada de uma Cadeia de Markov.

Por fim, em Amin et al. (2018) um modelo de Markov é utilizado como uma das *ba-*

selines, enfatizando e demonstrando que tal modelo é popular no problema de predição de mobilidade, porém é também percebido que esse tipo de modelo não pode utilizar coordenadas GPS diretamente, considerando a tesselação do mapa em *grids* ou células uma forma de discretização de valores contínuos em estados.

3.1.2 Artigos com Cadeias Ocultas de Markov

Em relação a Cadeias Ocultas de Markov, Cho (2016) põem em prática uma Cadeia Oculta de Markov para cada caminho utilizando inter-localizações intermediárias, que vêm do reconhecimento de localização fase e prevê o próximo local selecionando o modelo que produz a maior pontuação de correspondência. O modelo, segundo o pesquisador no referido artigo, usa o smart-registros telefônicos, como modo de transporte, dia da semana e horário para prever os locais sensíveis ao contexto de forma confiável. Isso considerando o reconhecimento e predição de lugares.

Entretanto, o artigo referido acima não aborda diretamente Cadeias Ocultas de Markov para predição de destinos de veículos terrestres. Isso fica mais claro em Alvarez-Garcia et al. (2010), que constrói um sistema baseado na geração de uma Cadeia Oculta de Markov ou Modelo Oculto de Markov de dados históricos de log de GPS e lugares atuais para predizer o destino de usuários quando iniciam uma jornada. Para tal, os pesquisadores consideram um conjunto de N estados distintos, o estado inicial da distribuição que seria o lar do usuário como sendo o mais provável, V como subconjuntos de estados distintos ou símbolos, a matriz de transição de probabilidade e a matriz de distribuição de probabilidade dos símbolos observados (a matriz de emissão).

Além desse artigo, há o dos pesquisadores Qiao et al. (2015), em que existe um modelo que se propõe a "predizer caminhos de objetos em movimento no lugar de fatias de trajetórias padrões". Eles consideram trajetórias cujos objetos possuem velocidades que mudam rapidamente, ajustando o modelo dinamicamente e propondo uma seleção de parâmetros auto-adaptativa. As sequências de trajetórias mais prováveis são encontradas através do algoritmo de Viterbi.

Também é possível a referência a Lassoued et al. (2017), que apresentam "um modelo e algoritmo simples para prever destinos e rotas de motoristas, com base na entrada das últimas ligações rodoviárias visitadas como parte de uma viagem em curso.". Isso através de pré-processamento, agrupamento, treinamento e predição - essa última etapa através de Cadeias Ocultas de Markov onde se entende que um motorista não mude seu destino ou rota de repente. Portanto, o estado do agrupamento oculto é considerado independente do tempo, e a matriz de transição é simplesmente a matriz identidade. Ou seja, consideram como novidade mais visível apenas a matriz de emissão.

3.1.3 Trabalhos com dados contextuais e outras técnicas

Tendo-se em consideração que a distinção entre objetivos (se predição de destinos, predição de trajetórias ou de ambos os elementos) não altera substancialmente o tipo ou a estrutura do dado, observa-se que há algo crucial que se pode julgar valioso para predizer destinos ou mesmo trajetos inteiros: a vontade humana de acordo com um contexto. Para tal, a necessidade de dados que são significativos para cada pessoa, e que, diferente dos dados temporais ou contextuais, não são invariáveis independente da pessoa.

Em Junior, Dutra e Neto (2024), há um quadro de trabalhos com dados contextuais (Quadro 2), onde há as referências autoriais. Foram utilizadas identificações de pontos (como PoIs) e de áreas (como AoIs) de interesse no contexto dos dados dos trabalhos analisados. Além de considerar que "N"e "S"significam "Não"e "Sim", respectivamente; Direct (D) ou "Diretamente", Indirect (I) ou "Indiretamente"e Unknown (DESC) ou "Desconhecido".

Dessa forma, na Tabela 6, as colunas, respectivamente, foram: a primeira para identificação das Referências (autor e ano da publicação); a segunda explicando a presença ou ausência de PoIs ou AoIs; e a terceira sobre de onde foram gerados os dados, a Geração.

Tabela 6 – Artigos que utilizaram dados contextuais e outras técnicas.

Referências	**Pols/Aols**	**Geração**
Araújo et al. (2019)	N	I (Foursquare)
Barth et al. (2020)	S	I (preferência de rota)
Besse et al. (2018)	S	DESC
Jiang et al. (2022)	S	DESC
Lassoued et al. (2017)	S	D (OpenStreetMap)
Liu et al. (2019)	S	I (Didi Chuxing, Uber e Grab)
Qin et al. (2023)	S	D (espaço circundante)
Rainbow et al. (2021)	N	D (por tipo de objeto)
Chang et al. (2022)	N	D (considerando dados de altitude)
Santana e Campos (2017)	N	D (através de georreferenciamento)
Tang et al. (2021)	S	D (dados esparsos como horário de partida)
Tong et al. (2021)	N	D (através de perfis de viagem)
Wang et al. (2023)	S	D (através do número de segmentos de estrada)
Wu et al. (2020)	N	I (ambiente ao redor do pedestre)

Fonte: Junior, Dutra e Neto (2024).

Os trabalhos elencados na Tabela 4 trazem um melhor detalhamento do que significam dados mais próximos da condição humana, únicos, que melhor se adequam ao conceito de experiência como é o caso do de trajetória. Não são apenas padrões espaciais ou espaço-temporais, mas contextos mais completos.

3.2 Diferenças entre os artigos analisados e diferencial em relação a esta pesquisa

Na Tabela 7, têm-se, respectivamente, as diferenças entre os artigos analisados, e o diferencial deste trabalho em relação aos artigos. Os campos são para identificação ou descrição de Autores (referências), qual o Modelo utilizado, o Contexto dos Dados (suas origens, se são reais ou mais de um conjunto de dados, ou mesmo a ausência clara de informações sobre suas origens e características), a Validação (as técnicas estatísticas empregadas para validar) e o Diferencial desses trabalhos capaz de destacá-los no sentido de servirem de referência a esta pesquisa.

Tabela 7 – Análise comparativa dos artigos em relação ao trabalho proposto.

Autores	**Modelo**	**Contexto dos Dados**	**Validação**	**Diferencial**
Wang et al., 2023	Modelo com Transformer	Definido pelo uso do conjunto de dados Argoverse	*Minimum Average Displacement Error*, *Minimum Final Dis- placement Error* and the MR (*Miss Rate*)	Dados reais recentes, modelagem e validação simplificada
Sadri et al., 2018	Modelo próprio com segmentação multi-critério	Dois conjuntos de dados reais	Apropriada para um modelo próprio, como taxa de segmentação	Dados reais recentes, e uso de Markov como preditor (e não apenas como meio de preparação dos dados)
Dai et al., 2019	Inovou com um LSTM Espacial	Conjunto de dados da rodovia I-80	*Mean Absolute Deviation* e *Root Mean Square*	Conjunto de dados reais recentes
Lassoued, 2017	Markov Oculto e Clusterização	Cidade de Dublin pelo OpenStreetMap (OSM) e dados sintéticos	O principal está na comparação entre um conjunto de dados real e um sintético	Foco na comparação entre dois modelos markovianos, um considerando apenas as localidades e o outro considerando também um marcador temporal
Santana e Campos, 2017	Algoritmo próprio, que agrupa ou isola pontos (abordagem sem Aprendizado de Máquina)	Falta clareza no artigo, sobre o conjunto de dados	Uso de *Accuracy* como validador estatístico	Uso de dados reais

Fonte: Elaborado pelo autor.

Sendo assim, o diferencial mais relevante deste trabalho em relação aos aqui relacionados diz respeito à maneira específica como os dados foram amostrados, além da forma como foi feito o balanceamento dos dados, que se mostrou necessário tanto para melhora dos resultados como para a viabilidade deles. Secundariamente, um conjunto de dados reais e relativamente recentes (a saber, o *Vehicle Energy Dataset*, conhecido pela sigla VED), e uma validação apenas com a

Precisão com o intuito implícito de entender quando um modelo pode ser mais preciso que o outro ao incluir informação contextual.

3.3 Conclusão da Revisão da Literatura

Para finalizar este capítulo, explica-se que foram apresentados, descritos e explicados alguns dos trabalhos do MSL, porém não todos devido à extensão das obras analisadas. Entretanto, foi possível ter uma compreensão acerca das peculiaridades de cada modelo e conjunto de dados, e deles com este trabalho. O foco não foi tão voltado a artigos com técnicas de agrupamento ou com redes neurais profundas, porém esses trabalhos existem e são bastante relevantes, mas o diferencial deste trabalho também é se propor a uma comparação mais simples, entre modelos markovianos visando, assim, uma comparação.

No próximo capítulo, referente à Metodologia, deverá ficar mais claro como essa comparação se tornou possível e, no capítulo seguinte, os resultados obtidos e analisados.

4 METODOLOGIA

A Metodologia deste trabalho consiste na explicação sobre as etapas do desenvolvimento dos modelos computacionais e manejo dos dados de deslocamentos.

As etapas são sobre a preparação dos dados brutos (na seção 4.1), que contempla a concatenação e análise inicial desses dados, além das derivações de novos campos, e a finalização da engenharia dos dados. Em seguida, existe a etapa de Segmentação (na seção 4.2) e a da Consolidação (na seção 4.3), até a Geração do *Dataset* Final (na seção 4.4).

Após essas etapas, torna-se possível a elaboração dos modelos preditivos e sua validação, cujo funcionamento foi explicado no Capítulo da Fundamentação Teórica. A materialização desse método prossegue na seção 4.5, com a implementação dos modelos de predição.

Dessa forma , neste capítulo, são explicados os modelos computacionais de predição de trajetórias construídos. A explicação será tanto em nível conceitual como também em termos de ferramentas utilizadas.

Na predição de destinos, foram utilizadas duas técnicas frequentes para a construção de modelos de predição de trajetórias: uma com base em Cadeias de Markov, utilizado em; e outra com base em HMM - cujos exemplos de uso se encontram em Araújo et al. (2019) e Lassoued et al. (2017), dentre outros.

Além disso, foi definido o conjunto de dados para realizar uma análise comparativa da performance. Dentre as bases de dados disponíveis na literatura, optou-se por aquela com dados reais de deslocamento, e que foram coletados em período recente. Esses dados possuem uma estrutura que inclui campos de longitude e de latitude, e campos que se referem à marcação do tempo. Dessa forma, foi utilizada a base de dados ABC para realizar a análise comparativa dos dois modelos de predição desenvolvidos *Vehicle Energy Dataset* (VED)[1]. Essa baseapresenta informações dos anos 2017 e 2018, voltados à área de Ann Arbor, da parte central até um pouco além dos seus limites. Além do fato de que já vieram anonimizados - não sendo possível determinar usuários específicos nem lugares associados a alguém.

4.1 Preparação dos Dados Brutos

Uma vez que os dados foram obtidos, esta etapa consistiu em concatenar arquivos CSV e derivar as colunas de data e dia da semana.

[1] Disponível em <https://github.com/gsoh/VED>.

4.1.1 Concatenação e Análise Inicial

A concatenação surgiu diante da necessidade de unificar os dados, que provêm de 54 arquivos sequenciados por períodos semanais, em que o nome dos arquivos indicam seu conteúdo da seguinte forma: "VED_mmddyy_week.csv". Por exemplo: no campo *DayNum* os números, decimais, começam de 1 até o último arquivo, com um decimal que começa com 375. E essa unificação serviu para a análise inicial dos dados. Na Figura 6 o dia de início dos experimentos do VED.

Lembrando que essa junção é resultante de dezenas de arquivos no referido formato CSV, após a descompactação, e se subdivide em dados estáticos (metadados) e em dados dinâmicos.

Figura 6 – O início do campo *DayNum* do primeiro dos 54 arquivos.

	DayNum
0	1.586651
1	1.586651
2	1.586651
3	1.586651
4	1.586651
...	...
489409	7.992231
489410	7.992231
489411	7.992231
489412	7.992231
489413	7.992231

Fonte: Elaboração própria.

Para dados dinâmicos, os arquivos foram utilizados conforme na Tabela 8. A configuração tabular original desse arquivo final, para o veículo cujo identificador é "531", pode ser evidenciada com as colunas mais importantes marcadas em vermelho (DayNum, VehId, Trip, Timestamp(ms), Latitude[deg] e Longitude[deg]).

Tabela 8 – Dados de veículos do Vehicle Energy Dataset

DayNum	VehId	Trip	Timestamp(ms)	Lat[deg]	Long[deg]	Speed [kph]	MATS/ RPM	Engine Load[%]	Absolute Power	AC Power	AC Power	Heater	HV Batt SOC[%]	Short T Fuel[1%]	Short T Fuel[2%]	Long T Fuel[1%]
173.89	531	1091	0	42.26	-83.73	17.0	132400	17342	13.73	NaN	NaN	NaN	NaN	NaN	NaN	NaN
173.89	531	1091	100	42.26	-83.73	30.0	132400	17342	13.73	NaN	NaN	NaN	NaN	NaN	NaN	4.68
173.89	531	1091	700	42.26	-83.73	39.0	132400	9242	13.73	NaN	NaN	NaN	NaN	NaN	NaN	4.68
173.89	531	1091	800	42.26	-83.73	39.2	137699	9242	13.73	NaN	NaN	NaN	NaN	NaN	NaN	4.68
173.89	531	1091	1100	42.26	-83.73	46.0	217899	9242	13.73	NaN	NaN	NaN	NaN	NaN	NaN	4.68
								...								
7.78	531	629	0	46.01	42.24	17.0	2.78	144.28	14.11	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7.78	531	629	100	46.01	42.24	33.0	3.91	105.20	14.11	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7.78	531	629	200	46.01	42.24	11.0	2.78	105.20	14.11	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7.78	531	629	300	46.52	42.24	52.0	3.91	105.20	14.11	NaN	NaN	NaN	NaN	NaN	NaN	NaN
7.78	531	629	400	46.01	42.24	32.0	3.91	105.20	14.11	NaN	NaN	NaN	NaN	NaN	NaN	NaN

Fonte – Elaborada pelo autor (2025).

Coordenadas de longitude e latitude estão no Sistema de Referência de Coordenadas (EPSG[2] “4326”) para coordenadas geográficas decimais, englobando todo o planeta, mas centrado em localidade específica desta pesquisa: Ann Arbor, Michigan, EUA. Foi necessário convertê-los num formato de melhor leitura para o computador, o *Geometry*, que é próximo do WKT ou *Well-Known Text*, mas com uma configuração própria e adequada para a estrutura de dados GeoDataFrame, representando dados binários.

O campo VehId representa o identificador dos veículos, enquanto *Trip* representa o identificador único das trajetórias.

4.1.2 Derivações de Novos Campos

Quanto ao aspecto temporal, com o auxílio dos metadados para entendimento contextualizado sobre o que cada campo significa, além do *DayNum* e do *Timestamp(ms)*, foi possível projetar um “*datetime*”, coluna derivada daquelas duas colunas, servindo tanto para o processamento da segmentação as trajetórias, como para a derivação do campo *day*.

Assim, a engenharia de dados necessária envolveu essencialmente a criação do campo *datetime*, convertendo a coluna *Timestamp(ms)* para *timestamps* com o parâmetro "D", indicando que essa última coluna representa "dias". Em seguida, com base na leitura da documentação do VED, determina-se primeiro de novembro de 2017 como o início da contagem dos dias, com base no campo *DayNum*.

Em seguinda, com a definição do *timestamp* mais antigo e a do mais recente, é feita uma interpolação em que, num *dataframe* inicialmente vazio chamado "t", com o método *date_range*,

[2] *European Petroleum Survey Group*, que serve como identificador único da posição e projeção de uma feição espacial no contexto geográfico

é criada uma série de datas igualmente espaçadas (num mesmo intervalo) com base no tamanho do *dataframe*.

Ao final, os valores de "t"são atribuídos à coluna nova do primeiro *dataframe*, havendo uma conversão novamente para o formato *datetime* e o arredondamento desses dados para que sejam considerados apenas até os segundos.

A seguir, na Tabela 9, têm-se as colunas e os dados derivados, além de 2 colunas, Day e Period, obtidas com a ajuda do PyMove, expostos em vermelho.

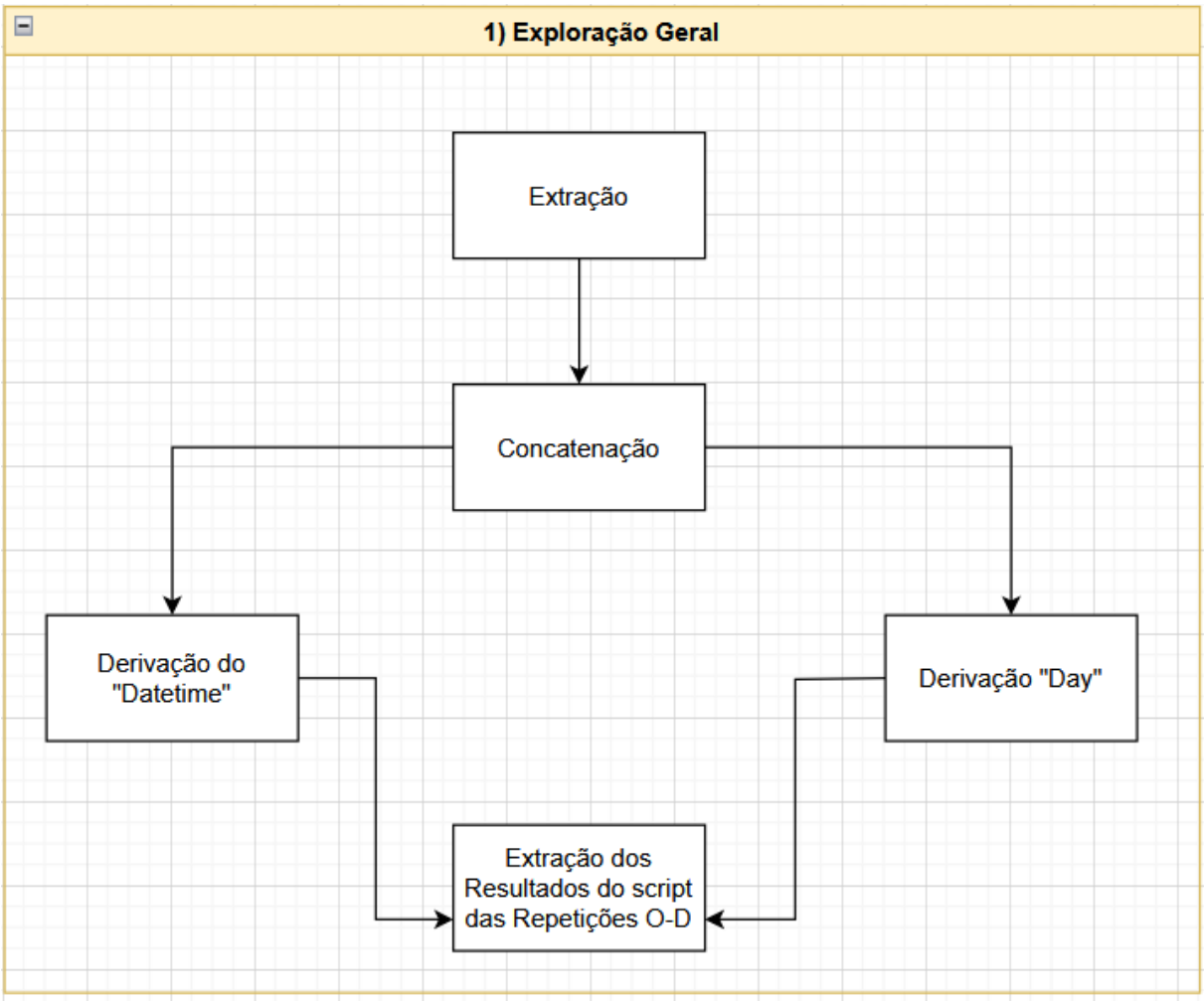
Tabela 9 – Dados de trajetos veiculares com informações temporais

VehId	**Trip**	**lat**	**lon**	**datetime**	**day**
560	2	42.252974	-83.674152	2017-11-02 12:05:45	Thursday
560	2	42.252974	-83.674152	2017-11-02 12:05:57	Thursday
560	2	42.252974	-83.674152	2017-11-02 12:06:08	Thursday
560	2	42.252974	-83.674152	2017-11-02 12:06:20	Thursday
560	2	42.252974	-83.674152	2017-11-02 12:06:32	Thursday
		...		...	...
371	4411	42.282303	-83.734510	2018-11-11 11:35:46	Sunday
371	4411	42.282303	-83.734510	2018-11-11 11:35:58	Sunday
371	4411	42.282303	-83.734510	2018-11-11 11:36:10	Sunday
371	4411	42.282303	-83.734510	2018-11-11 11:36:22	Sunday
371	4411	42.282303	-83.734510	2018-11-11 11:36:34	Sunday

Fonte – Elaborada pelo autor, com base no VED.

Em síntese, com essas tarefas iniciais, foi possível obter a ordenação a partir do novo campo *Datetime* derivado pela engenharia dos dados e a derivação do campo *Day*, com os dias da semana em Inglês, mediante a ferramenta (biblioteca Python) *PyMove*.

Figura 7 – Processo de Obteção das Repetições Origem-Destino.



Fonte: Elaboração própria.

Entretanto, houve a necessidade de segmentação dessas *Trips*, sendo, com base nesse campo, obtidas as subtrajetórias. Essa necessidade surgiu da possibilidade de obtenção de uma melhor representatividade de viagens que um usuário individual realizaria. E é nesse caminho da segmentação de que trata a próxima etapa.

4.2 Segmentação

Na Figura 8, tem-se a preparação dos dados, que se iniciou através das trajetórias - a servirem de fonte para a extração das subtrajetórias.

Com essas trajetórias e os parâmetros 100 metros, 30 minutos e 1000 metros realizou-se um processo de *stay-point based segmentation* a partir da ferramenta *MovingPandas* (Moving-Pandas Contributors, 2024). Procedimento em que os 100 metros existem como o valor máximo de deslocamento, os 30 minutos como o tempo mínimo de parada e os 1000 metros para filtrar

subtrajetórias com no mínimo esse comprimento[3].

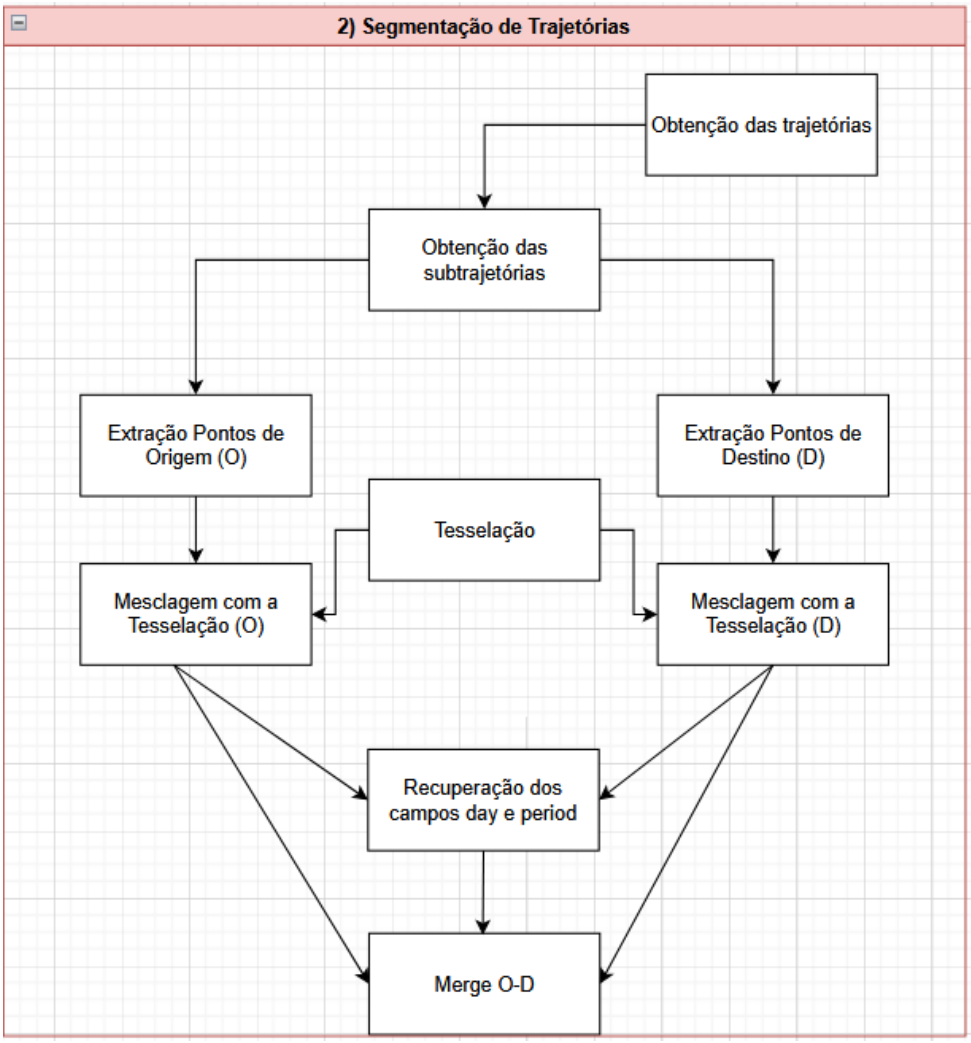
Tratando-se de segmentar as trajetórias, surgem as subtrajetórias, que, para esta pesquisa, consiste no particionamento de trajetórias por meio da abstração de momentos em que a trajetória é registrada como pontos sem movimentação (o tempo passa, mas existe a permanência dentro de uma região delimitada por 30 minutos). Por exemplo: uma trajetória 132 torna-se 132.1, 132.2 e 132.3, com base na separação entre os momentos vistos na 132.1 até um momento em que se perde a continuidade temporal, e se retorna no início da 132.2, seguindo-se a mesma lógica para a 132.3 e quantas mais subtrajetórias existirem até a finalização do segmento-pai 132, desde que possua, no mínimo, 1000 metros de comprimento.

Assim, a sequência de tarefas desta etapa, apresenta-se como: Obtenção das trajetórias com o *MovingPandas*[4], a partir dos dados originais e, mediante essas trajetórias, a obtenção das subtrajetórias. Em seguida, ocorre a Extração dos Pontos de Origem (O) e de Destino (D). Com a Tesselação em grades de 100 metros quadrados, há a mesclagem ou combinação desses pontos de Oridem e de Destino com os rótulos das grades que passam a representá-los, com a inclusão das colunas *datetime* e *day*. Ao final, ocorre uma junção desses rótulos referentes às células de origem e de destino. A saída corresponde ao resultado dessa junção.

[3] rata-se de um procedimento citado e analisado em Sun et al. (2024).

[4] Para análise de trajetórias e engenharia dos dados, foram utilizadas as bibliotecas MovingPandas (MovingPandas Contributors, 2024), Scikit-Mobility (Scikit-mobility Team, 2024) e PyMove (PyMove Development Team, 2024).

Figura 8 – Preparação dos dados, centrada na obtenção das subtrajetórias.

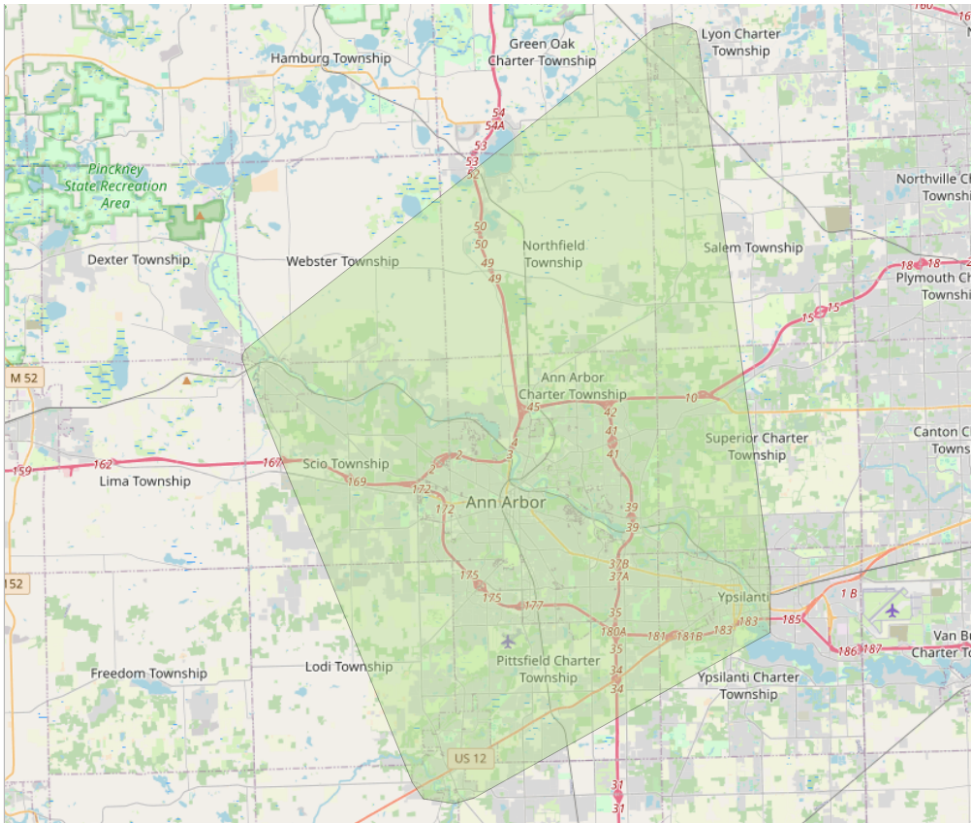


Fonte: Elaboração própria.

Assim, essas subtrajetórias estão no mesmo formato que as trajetórias utilizadas como fonte, sendo também necessário extrair células de Origem e de Destino, conforme visto, paralelamente à tesselação.

Essa tesselação é realizada tendo por parâmetro a região da Figura 8. Essa região abrange a cidade de Ann Arbor e 4 distritos vizinhos, conforme representação na Figura 9.

Figura 9 – Região de estudo.



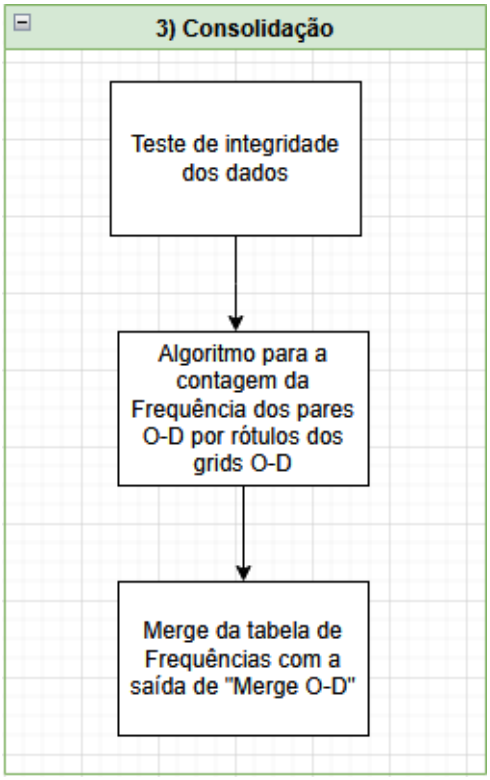
Fonte: Elaboração própria.

Por fim, ao término desta Fase, tem-se a Consolidação, que é descrita na próxima parte.

4.3 Consolidação

Na Figura 10, apresenta-se o teste de integridade dos dados - que consistiu em repetições sistemáticas do processo para validar a correta aplicação do método da biblioteca *MovingPandas* para a geração de subtrajetórias, a execução do algoritmo para a nova contagem de frequência de repetições por rótulos Origem-Destino das subtrajetórias, e ao final, a junção da tabela de contagem de repetições com a saída de uma junção das grades de Origem e de Destino.

Figura 10 – Consolidação, com teste de integridade.



Fonte: Elaboração própria.

Ou seja, a saída da Exploração Geral foi apenas Exploratória, para conhecer os dados. Depois, verificou-se a necessidade de repetição do algoritmo de contagem de repetições, dessa vez com as subtrajetórias e não com as trajetórias completas.

A saída dessa nova Exploração Geral e a saída da Preparação (etapa seguinte) é que foram utilizadas na Consolidação, com o uso do Algoritmo para a contagem da Frequência dos pares O-D e, ao final, como já dito, a junção da tabela de Frequências com a saída da Fase de Preparação (ou seja, a saída do processo repetido da Fase de Exploração, com a saída da Fase Exploração).

Dessa forma, esse processo gerou a Figura 11.

Figura 11 – Matriz das amostras.

Fonte: Elaboração própria.

A referida Figura da Matriz de amostras, implica em células cujos valores são o quantitativo de veículos que possuem 2+ trajetórias repetidas com 2+ repetições, 3+ trajetórias repetidas com 3+ repetições até 6+ trajetórias repetidas com 6+ repetições.

Com os 99 veículos da última célula da matriz - que são aqueles com 6+ trajetórias repetidas que possuem 6+ repetições -, foi iniciada uma segunda fase da pesquisa, que consiste no *pipeline* das análises estatísticas, predições e comparações. Mas, antes dessas etapas, foi necessário continuar preparando os dados de amostra.

4.4 Geração do *Dataset* Final

Para as modelagens com Cadeias de Markov e HMM e posterior execução das predições, foi realizada uma filtragem oriunda de um balanceamento dos dados, para mitigar problemas relacionados a dados insuficientes em termos de validação estatística ou cenários em que os algoritmos sequer possuem os elementos mínimos para processamento (conforme avisos de falha encontrados, no uso da biblioteca Python pyDTMC).

Dessa forma, inicia-se a explicação destas Etapas finais, com uma explicação acerca do balanceamento dos dados.

4.4.1 Balanceamento dos dados: obtenção da distribuição das repetições geográficas por veículo

Inicialmente, foi utilizado um algoritmo para processar os dados finais de Origens e de Destinos. No Algoritmo 1, pode-se observar o que é feito com os arquivos de Origem e de Destino dos veículos encontrados na célula 6+ por 6+ referida anteriormente, no formato *GeoDataFrame*:

Algoritmo 1 Processamento e Análise de Padrões de Movimento Geoespacial

Require: Arquivos geoespaciais de partida e chegada
Ensure: Análise de contagem de repetições de padrões de movimento
1: $partida \leftarrow$ CarregarArquivoGeospacial("partidafinal99_novo.gpkg")
2: $chegada \leftarrow$ CarregarArquivoGeospacial("chegadafinal99_novo.gpkg")
3: $mtdf1 \leftarrow$ FusionarDataFrames($partida, chegada$, "VehId", "id")
4: $mtdf1.pares \leftarrow$ Concatenar($mtdf1.tile\_ID\_x$, "_", $mtdf1.tile\_ID\_y$)
5: $mtdf1 \leftarrow$ SelecionarColunas($mtdf1$, ["VehId", "id", "pares", "tile_ID_x", "tile_ID_y", "day_x"])
6: **procedure** ANALISARPADRÕESDEMOVIMENTO($mtdf1$)
7: $contar \leftarrow$ OrdenarDataFrame($mtdf1$, ["tile_ID_x", "tile_ID_y", "day_x"])
8: $pares\_contados \leftarrow$ AgruparPorContagem($contar$, ["VehId", "tile_ID_x", "tile_ID_y", "day_x"])
9: $pares\_filtrados \leftarrow$ Filtrar($pares\_contados$, contagem > 1)
10: ExportarParaCSV($pares\_filtrados$, "pares_desmembrados_original.csv")
11: **return** $pares\_filtrados$
12: **end procedure**
13: **procedure** VISUALIZARDADOS(df)
14: $df\_max \leftarrow$ AgruparPorSoma(df, "VehId", "contagem")
15: ExibirValorMáximo($df\_max.contagem.soma()$)
16: ExibirTop5Valores(df_max, "contagem", decrescente)
17: CriarGráficoBarra(df_max, "VehId", "contagem")
18: AdicionarRótulos(gráfico, $df\_max.contagem$)
19: ConfigurarEixos(gráfico, "ID do Veículo", "Contagem Geográfica")
20: DefinirTítulo(gráfico, "Contagem Geográfica por ID de Veículo - Valores Agrupados de Repetição")
21: ExibirGráfico()
22: **end procedure**
23: $dados\_filtrados \leftarrow$ AnalisarPadrõesDeMovimento($mtdf1$)
24: VisualizarDados($dados\_filtrados$)

Fonte: Elaboração própria.

O Algoritmo 1 apresenta o método desenvolvido para geração da amostra final de dados, a ser utilizada para Cadeias de Markov e HMM. A partir de dois arquivos no formato GPKG (*GeoPackage*), referentes às partidas e às chegadas, é realizada uma fusão dos dados mediante as chaves de identificação. Este processo é executado conforme demonstrado nas linhas 1-2, em que os arquivos geoespaciais de partida e chegada são carregados utilizando a biblioteca GeoPandas. Na linha 3, é realizada uma junção (*merge*) entre partida e chegada utilizando as chaves `VehId` e `id`, onde `VehId` refere-se ao identificador do veículo e `id` às subtrajetórias.

Em seguida, na linha 4, é criada uma nova coluna `pares` que concatena os identificadores das grades espaciais de origem e destino (`tile_ID_x` e `tile_ID_y`), estabelecendo uma representação única para cada par origem-destino. Na linha 5, realiza-se uma seleção das colunas relevantes para a análise, mantendo apenas `VehId`, `id`, `pares`, `tile_ID_x`, `tile_ID_y` e `day_x`.

O procedimento `AnalisarPadrõesDeMovimento` (linhas 7-13) implementa a lógica principal de processamento. Na linha 8, os dados são ordenados pelos identificadores das grades e pelo dia das partidas. Em seguida, na linha 9, é realizado um agrupamento por `VehId`, `tile_ID_x`, `tile_ID_y` e `day_x`, calculando-se a frequência de ocorrência de cada combinação. A linha 10 implementa a filtragem das repetições, selecionando apenas os registros com contagem

superior a 1, representando padrões de movimento recorrentes. Os resultados são exportados para um arquivo CSV na linha 11.

O procedimento `VisualizarDados` (linhas 14-22) organiza a análise estatística e visualização dos resultados. Na linha 15, é realizado um novo agrupamento, desta vez somando as contagens por `VehId`, para identificar quais veículos apresentam maior recorrência de padrões. As linhas 16-17 exibem estatísticas básicas, incluindo o valor máximo total e os cinco veículos com maior número de padrões recorrentes.

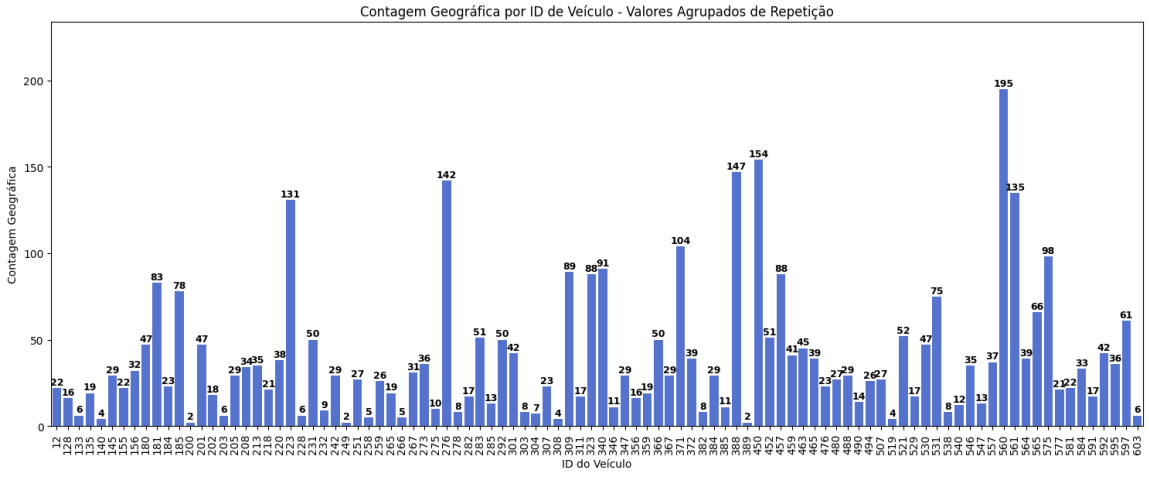
A visualização gráfica é configurada nas linhas 18-21, criando um gráfico de barras que representa a contagem geográfica por ID de veículo. O gráfico é enriquecido com rótulos de dados, títulos apropriados e configurações de eixos para facilitar a interpretação dos resultados.

Finalmente, nas linhas 23-24, o algoritmo executa a análise de padrões e realiza a visualização dos dados filtrados, permitindo a identificação de comportamentos recorrentes nos deslocamentos dos veículos analisados.

Esta metodologia permite identificar eficientemente tanto repetições geográficas (padrões de origem-destino) quanto temporais (padrões associados a dias específicos), fornecendo meios para análises mais aprofundadas sobre comportamentos de mobilidade.

Na Figura 12, pode-se observar o gráfico com base nas repetições presentes na coluna contagem para cada veículo.

Figura 12 – Distribuição das repetições geográficas por veículo.



Fonte: Elaboração própria.

4.4.2 Balanceamento dos dados: filtragem

Em relação ao balanceamento dos dados da amostra, sobre a técnica de filtragem empregada a partir da necessidade de balanceamento dos dados.

Algoritmo 2 Filtro de Trajetórias com Limiar Automático

```
1: procedure FiltrarTrajetosComLimiarAutomatico(df)
2:     coluna_veiculo ← "VehId"
3:     coluna_destino ← "tile_ID_y"
4:     n_estados_hmm ← 5
5:     limiares ← [10, 15, 20, 25, 30, 35, 40, 50]
6:     resultados ← [ ]
7:     for cada limiar em limiares do
8:         df_temp ← FiltrarPorGrupo(df, coluna_veiculo, λx : |x| ≥ limiar)
9:         if |df_temp| = 0 then
10:            continuar com próximo limiar
11:        end if
12:        n_veiculos ← |ValoresÚnicos(df_temp[coluna_veiculo])|
13:        seq_por_veiculo ← ContarPorGrupo(df_temp, coluna_veiculo)
14:        media_seq ← Média(seq_por_veiculo)
15:        estados_unicos ← |ValoresÚnicos(df_temp[coluna_destino])|
16:        adequacao_markov ← min(1.0, limiar/(10 · estados_unicos))
17:        params_hmm ← n_estados_hmm^2 + n_estados_hmm · estados_unicos
18:        adequacao_hmm ← min(1.0, limiar/(5 · params_hmm))
19:        pontuacao ← 0.5 · adequacao_markov + 0.5 · adequacao_hmm
20:        fator_veiculos ← n_veiculos/|ValoresÚnicos(df[coluna_veiculo])|
21:        pontuacao_final ← pontuacao · (0.7 + 0.3 · fator_veiculos)
22:        Adicionar {limiar, n_veiculos, pontuacao_final} a resultados
23:    end for
24:    resultados_ordenados ← Ordenar(resultados, por pontuacao_final, decrescente)
25:    limiar_ideal ← resultados_ordenados[0].limiar
26:    df_filtrado ← FiltrarPorGrupo(df, coluna_veiculo, λx : |x| ≥ limiar_ideal)
27:    n_veiculos_original ← |ValoresÚnicos(df[coluna_veiculo])|
28:    n_veiculos_filtrado ← |ValoresÚnicos(df_filtrado[coluna_veiculo])|
29:    ExibirEstatísticas(n_veiculos_filtrado, n_veiculos_original, |df_filtrado|, |df|)
30:    return df_filtrado
31: end procedure
```

Fonte: Elaboração própria.

Ou seja, no Algoritmo 2, o principal está nas linhas: 16, para cálculo de adequação para Cadeias de Markov; 17 e 18, para HMM; além da linha 19 com o cálculo da pontuação combinada, e das linhas 20 e 21, que correspondem ao ajuste da pontuação da perda de dados, com um fator de penalização dessa perda. Dessa forma, 50 foi o limiar determinado a partir do conjunto de opções "10, 15, 20, 25, 30, 35, 40, 50", e, de 99 veículos elencados dentre os que possuem ao menos 6 rotas, com ao menos 6 repetições cada, foram utilizados de fato 23 veículos para Cadeias de Markov e HMM. Os registros foram de 3776 para 2129 devido ao corte de valores somados por veículo inferiores a 50.

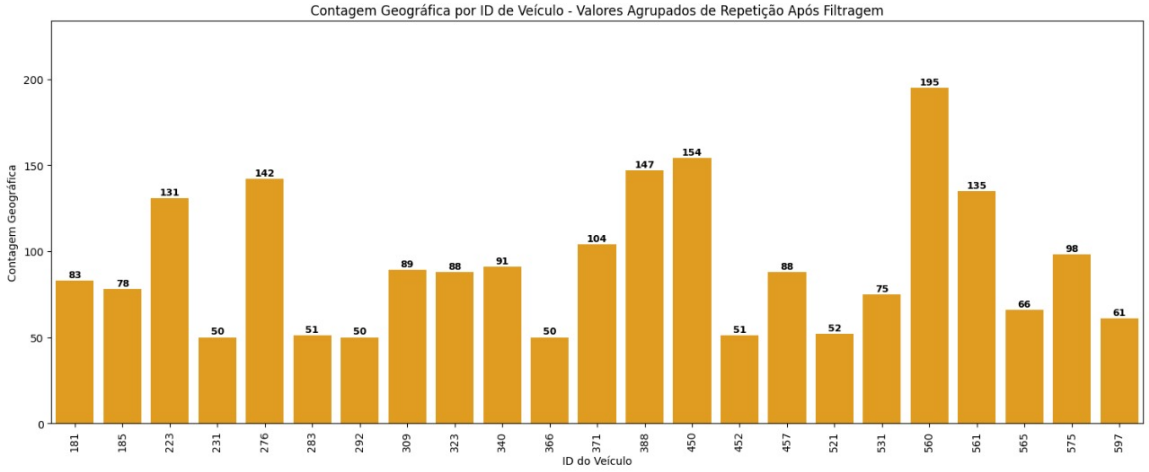
Entretanto, valores diferentes de 50 foram também testados, além de também dos dados terem sido testados com os modelos sem filtragens. Isso levou ao motivo da escolha, ou seja, que, sem um limiar, a quantidade de veículos que performam bem com os algoritmos de forma a existirem menos perdas é diferente para Markov e HMM, o que torna a validação estatística dos dados problemática, sendo que exigem um pareamento de veículos. Por exemplo: utilização de t-Student ou t-Student para conjuntos de dados com X e Y veículos, respectivamente. O correto é que sejam X e X ou Y e Y. Além disso, para cada *fold* nem sempre os modelos, principalmente HMM, geram predições, sendo o cenário em que esse problema é minimizado decorrente do limiar 50.

Os valores abaixo de 50, testados preliminarmente, não costumam trazer precisões médias muito relevantes em termos quantitativos (são, em geral, mais baixas para os dados amostrados);

ao mesmo tempo em que se pretendeu encontrar um filtro que equilibre performance além de coerência na consecução dos experimentos. Performance no sentido de valores mais altos, e coerência com o que foi definido originalmente com base no uso de fórmulas ancoradas em conceitos e explicações presentes no capítulo da Fundamentação Teórica.

O resultado pode ser observado na Figura 13 com base nas repetições presentes na coluna "contagem"para cada veículo no limiar de 50 (não há barras com valores inferiores a 50).

Figura 13 – Distribuição das repetições geográficas por veículo, após filtragem com o Algoritmo 2.



Fonte: Elaboração própria.

É importante destacar, dessa forma, que a amostra definitiva foi a de 2129 registros - ou seja, dos 3776 com a filtragem através do limiar 50, restou essa quantidade de registros de viagens.

A seguir, serão apresentadas as implementações dos modelos de predição com base em Markov e HMM.

4.5 Implementação

A primeira etapa consiste na captação das mesmas amostras tanto para Markov como para HMM (Markov Oculto). Dessa forma, após a filtragem das repetições agrupadas pelo limiar 50, tem-se um único conjunto de dados.

Uma amostra desse conjunto de dados final consta na Tabela 10.

Tabela 10 – Amostra do conjunto de dados de trajetos de veículos.

Veículo	**grid_origem**	**grid_destino**	**dia**	**contagem**
181	66449	75312	Thursday	15
185	69505	78354	Monday	5
223	30421	42530	Wednesday	17
231	21285	25818	Sunday	7
276	77130	74546	Tuesday	4
283	78322	51969	Thursday	2
292	66449	75312	Monday	9
309	75736	51150	Tuesday	2
323	75312	51558	Tuesday	2
340	21628	41349	Monday	2
366	73701	78354	Sunday	6
371	75364	50739	Wednesday	2
388	66449	66047	Tuesday	2
450	69852	77895	Saturday	2
452	66071	64033	Thursday	4
457	75312	66449	Saturday	6
521	48204	47040	Saturday	9
531	31691	39346	Tuesday	3
560	78354	78354	Saturday	10
561	66449	50738	Sunday	2
565	31284	30950	Sunday	4
575	78354	77928	Tuesday	2
597	21933	44191	Wednesday	2

Fonte: Elaborado pelo autor.

Quanto ao Dicionário de Dados dos dados utilizados para Markov e HMM, existe conforme na Tabela 11.

Tabela 11 – Estrutura dos campos do conjunto de dados.

Campo	**Tipo**	**Descrição**
Veículo	int64	Identificador único de cada veículo.
grid_origem	int64	Identificador da grade de origem (partida).
grid_destino	int64	Identificador da grade de destino (chegada).
dia	object	Dia da semana do trajeto.
contagem	int64	Quantidade de vezes que o trajeto foi realizado.

Fonte: Elaborado pelo autor.

A segunda etapa, trata-se do uso do modelo de Markov. A terceira, trata-se do uso do modelo de HMM. Ao final, é feito o teste de Shapiro-Wilk para determinar a normalidade dos valores de *mean_precision* de cada conjunto de resultados (para os dois modelos). Ao determinar a normalidade ou não desses valores, vem a quarta e última etapa, em que é aplicado o teste de validação.

4.5.1 Modelo com Base em Cadeias de Markov

Esse Modelo considerou apenas a sucessão ou conjunto de rótulos referentes às grades de origens (que se sucedem conforme uma progressão temporal), para predizer os rótulos referentes às grades de destinos, a propriedade de Markov, em que uma ocorrência depende apenas da ocorrência imediatamente anterior e é a única causa da ocorrência seguinte. Além disso, utilizou-

se de um código baseado, primeiramente, na configuração dos estados únicos (rótulos das grades de origens e de destinos onde ocorreram a presença do veículo em movimento), e dos estados de teste, para Kfold (sendo k=10) com embaralhamento ou *shuffle*. Ou seja, para todos os deslocamentos de um determinado usuário foram particionados em 10 grupos, diferentes entre si, com seus próprios dados de deslocamento. Além disso, foi elaborada uma função para a matriz de probabilidades de transição, cujo conceito foi descrito na Fundamentação Teórica.

Em seguida, um método integrou os estados únicos de treino por 10 vezes para cada um dos 23 veículos, também por 10 vezes. Trata-se do método *MarkovChain()* do PyDTMC.

Segue o pseudocódigo para o modelo com base em Cadeias de Markov, no Algoritmo 3.

Algoritmo 3 Análise de Cadeias de Markov com Validação Cruzada Repetida

Require: DataFrame df com colunas 'Veículo', 'grid_origem', 'grid_destino'
Ensure: Resultados de avaliação por veículo em *all_results*
$n\_splits \leftarrow 10$; $n\_repeats \leftarrow 10$; $\alpha\_values \leftarrow [0.001, 0.01, 0.05, 0.1]$; $all\_results \leftarrow \{\}$
for cada $(vehicle\_id, vehicle\_data)$ em $df.groupby$('Veículo') **do**
  if $|vehicle\_data| < 10$ **then continue**
  end if
  $states \leftarrow$ sorted(unique($vehicle\_data$['grid_origem'] $\cup$ $vehicle\_data$['grid_destino']))
  if $|states| < 2$ **then continue**
  end if
  $all\_repeat\_results \leftarrow []$; $all\_fold\_precisions \leftarrow []$
  for $repeat \in \{0, \ldots, n\_repeats - 1\}$ **do**
    $kf \leftarrow$ KFold(n_splits, shuffle = True, random_state = $42 + repeat$)
    $\alpha\_results \leftarrow \{\}$
    for α em α_values **do**
      $fold\_precisions \leftarrow []$
      for $(fold, (train\_idx, test\_idx))$ em $enumerate(kf.split(vehicle\_data))$ **do**
        $train \leftarrow vehicle\_data.iloc[train\_idx]$; $test \leftarrow vehicle\_data.iloc[test\_idx]$
        if $|train| < 5$ ou $|test| < 2$ **then continue**
        end if
        $trans[s_1][s_2] \leftarrow \alpha \; \forall s_1, s_2 \in states$ ▷ Pseudocontagem
        for linha em $train$ **do**
          $trans$[linha.grid_origem][linha.grid_destino] += 1
        end for
        $P[i, j] \leftarrow trans[s_i][s_j] / \sum_k trans[s_i][s_k] \; \forall i, j$ ▷ Normalizar
        $mc \leftarrow$ MarkovChain($P, states$)
        $y\_true \leftarrow []$; $y\_pred \leftarrow []$
        for linha em $test$ **do**
          if linha.grid_origem $\in states$ **then**
            $pred \leftarrow mc.predict(1,$ linha.grid_origem$)$
            $y_true.append($linha.grid_destino$)$
            $y\_pred.append(pred[1])$ se $|pred| > 1$ senão $pred[0]$
          end if
        end for
        if $|y\_true| > 0$ **then**
          $prec \leftarrow$ precision_score(y_true, y_pred, average = 'weighted')
          $fold\_precisions.append(prec)$
        end if
      end for
      if $|fold\_precisions| \geq 3$ **then**
        $\alpha\_results[\alpha] \leftarrow$ {folds : $fold\_precisions$, mean : mean($fold\_precisions$)}
      end if
    end for
    if $\alpha\_results \neq \emptyset$ **then**
      $best\_\alpha \leftarrow \arg\max_\alpha \alpha\_results[\alpha]$.mean
      $best\_result \leftarrow \alpha\_results[best\_\alpha]$
      $all\_fold\_precisions.append(best\_result$.folds$)$
      $all\_repeat\_results.append($\{alpha : $best\_\alpha$, mean : $best\_result$.mean, folds : $best\_result$.folds\}$)$
    end if
  end for
  if $all\_repeat\_results \neq \emptyset$ **then**
    $precisions \leftarrow [r$.mean para r em $all\_repeat\_results]$
    $\mu \leftarrow$ mean($precisions$); $\sigma \leftarrow$ std($precisions$)
    $stat, p \leftarrow$ shapiro($precisions$) se $|precisions| \geq 3$ senão (None, None)
    $best\_\alpha \leftarrow$ mode($[r$.alpha para r em $all\_repeat\_results]$) ▷ Mais frequente
    $all\_results[vehicle\_id] \leftarrow$ {mean : μ, std : σ, shapiro_p : p, best_alpha : $best\_\alpha$,
      repeat_results : $all\_repeat\_results$, all_folds : $all\_fold\_precisions$}
  end if
end for
Calcular teste Shapiro-Wilk global com todos os valores de fold
Gerar DataFrames com resultados, visualizações e salvar múltiplos CSVs
return *all_results*

Fonte: Elaboração própria.

O Algoritmo 3 implementa o modelo para Cadeias de Markov de primeira ordem com validação cruzada repetida. A abordagem combina múltiplas técnicas para garantir a confiabilidade estatística dos resultados.

A partir da linha 1, o algoritmo define os parâmetros principais: 10 *folds* para validação cruzada (n_splits), 10 repetições independentes ($n\_repeats$), e quatro valores de suavização Laplace ($\alpha\_values = [0.001, 0.01, 0.05, 0.1]$). O dicionário $all\_results$ armazenará os resultados consolidados de cada veículo. Nas linhas 2 a 5, para cada veículo no conjunto de dados, o algoritmo primeiro verifica a viabilidade da análise. Veículos com menos de 10 registros (linha 3) ou com menos de 2 estados únicos (linha 6) são descartados, pois não fornecem dados suficientes para uma modelagem confiável. Os estados são extraídos da união das colunas de origem e destino (linha 5).

Nas linhas 8 a 39, implementa-se a estratégia de validação cruzada repetida. A linha 8 inicializa as estruturas $all\_repeat\_results$ e $all\_fold\_precisions$, sendo esta última crucial para armazenar todas as precisões individuais dos *folds* para análises estatísticas posteriores. Cada uma das 5 repetições (linha 9) utiliza uma semente aleatória diferente ($42 + repeat$) para o embaralhamento do K-Fold (linha 10), garantindo diversidade no particionamento dos dados.

Nas linhas 11 a 31, o algoritmo implementa uma busca exaustiva pelos melhores hiperparâmetros. Para cada valor de α (linha 12), realiza-se a validação cruzada completa. A matriz de transição é inicializada com pseudocontagem α (linha 18) para evitar probabilidades zero, e as transições observadas são acumuladas (linhas 19-21). A normalização (linha 22) garante que cada linha da matriz some 1, criando uma cadeia de Markov válida (linha 23).

Nas linhas 25 a 29, o modelo realiza predições no conjunto de teste. O método `predict` retorna uma sequência, sendo necessário tratar diferentes tamanhos de retorno (linha 29). A precisão ponderada é calculada comparando as predições com os valores reais (linha 33).

Nas linhas 32 a 45, o algoritmo primeiro completa todos os *folds* para cada α (linha 34), depois seleciona o melhor α baseado na média de precisão entre todos os *folds* (linha 38). As precisões dos *folds* da melhor configuração são armazenadas em $all\_fold\_precisions$ (linha 44), e na linha 45, são armazenados todos os melhores resultados de repetições.

Nas linhas 48 a 54, após completar todas as repetições para um veículo, o algoritmo consolida os resultados estatísticos. Caso haja resultados válidos em $all_repeat_results$ (linha 48), é extraída a lista de precisões médias de cada repetição (linha 49). A média global (μ) e o desvio padrão (σ) dessas precisões são calculados (linha 50), fornecendo uma medida da performance geral e sua variabilidade entre repetições.

Quanto ao teste de Shapiro-Wilk (linha 51), sabendo-se que é aplicado quando há pelo menos 3 valores de precisão, verificando se as médias das repetições seguem uma distribuição normal. O melhor α global é identificado como o valor modal - ou seja, o que aparece com maior frequência entre as melhores configurações de cada repetição (linha 52), garantindo que a escolha final reflita a configuração mais consistentemente bem-sucedida. Todos esses resultados são armazenados na estrutura $all_results$ para o veículo atual (linhas 53-54).

Na linha 58 há a geração de múltiplos arquivos CSV contendo diferentes granularidades

dos resultados - desde resumos por veículo até valores individuais de cada *fold*.

É importante salientar que o treinamento foi realizado com os rótulos das grades de origem e de destino, e a etapa de teste utiliza apenas os rótulos de origem para predizer o destino, havendo a comparação, para o cálculo da precisão, dos resultados com os rótulos reais de destino não utilizados durante o treinamento.

4.5.2 Modelo com Base em Cadeias Ocultas de Markov - ou HMM

No contexto deste trabalho, o uso de HMM volta-se aos padrões de movimentação humana, principalmente ao ser mais enriquecida no momento em que o modelo, além da matriz de transição e dos estados únicos para treino, exige a matriz de emissão e os símbolos a serem combinados aos estados ocultos. Ou seja, os símbolos surgem na condição da consideração de cada dia da semana mais a grade de origem, - dessa forma referentes às partidas de cada veículo - e são combinados pelo referido método, à matriz de emissão (que, tal como a matriz de transição, deve ser quadrática, cada linha somar 1, e ser sobre a probabibilidade de mudança de posição de um rótulo para o outro). Um exemplo simples pode ser considerado tomando-se emissão: <segunda-feira, 150>, em que segunda-feira representa o dia de início de uma movimentação, e 150 o rótulo de um grid de origem hipotético.

A predição foi feita considerando os estados ocultos de teste, para um modelo construído com base no algoritmo de Viterbi, e as comparações, para a obtenção das Precisões, com a variável ytest.

O pseudocódigo referente ao desenvolvimento do modelo de predição com base em HMM está no Algoritmo 4.

Assim, tem-se que o Algoritmo 4 implementa predições mediante HMM com validação cruzada repetida, integrando informações temporais (contextuais de acordo com os dias da semana) e espaciais (rótulos das grades de origem) para predição de trajetórias. O treinamento é realizado para os estados ocultos considerando os rótulos das grades de destino; e para os símbolos observáveis considerando a composição dos rótulos das grades de origem com os dias da semana. O teste ocorre com os rótulos das grades de destino não utilizados no treinamento, comparados, para a obtenção das precisões, com os rótulos das grades de destino reais.

Dessa forma, nas linhas 1-7, o algoritmo começa definindo os parâmetros de validação cruzada na linha 1, com 10 *folds* e 10 repetições, além de quatro valores de suavização α. Para cada veículo no conjunto de dados (linha 2), verifica-se se há dados suficientes (linha 3), exigindo pelo menos 10 observações e 2 estados distintos. Os estados são definidos como os *grids* únicos de origem e destino (linha 5), enquanto os símbolos observáveis são criados concatenando o dia da semana com o *grid* de origem (linha 6).

Algoritmo 4 Análise de Modelos Ocultos de Markov com Validação Cruzada Repetida

Require: DataFrame df com colunas 'Veículo', 'grid_origem', 'grid_destino', 'dia'
Ensure: Resultados de avaliação por veículo em *all_results*
1: $n\_splits \leftarrow 10$; $n\_repeats \leftarrow 10$; $\alpha\_values \leftarrow [0.001, 0.01, 0.05, 0.1]$
2: **for** cada $(vehicle\_id, vehicle\_data)$ em $df.groupby$('Veículo') **do**
3: **if** $|vehicle\_data| < 10$ ou $|states| < 2$ **then continue**
4: **end if**
5: $states \leftarrow$ unique($vehicle\_data$[origem] $\cup$ $vehicle\_data$[destino])
6: $vehicle\_data$[simbolo] $\leftarrow$ 'dia:' + $vehicle\_data$[dia] + '_grid:' + $vehicle\_data$[origem]
7: $symbols \leftarrow$ unique($vehicle\_data$[simbolo]); criar mapeamentos estado/símbolo $\leftrightarrow$ índice
8: $all\_repeat\_results \leftarrow []$; $all\_fold\_precisions \leftarrow []$
9: **for** $repeat \in \{0, \ldots, n\_repeats - 1\}$ **do**
10: $kf \leftarrow$ KFold(n_splits, seed $= 42 + repeat$); $\alpha\_results \leftarrow \{\}$
11: **for** cada α em α_values **do**
12: $fold\_precs \leftarrow []$
13: **for** $(train, test)$ em $kf.split(vehicle\_data)$ **do**
14: **if** $|train| < 5$ ou $|test| < 2$ **then continue**
15: **end if**
16: **try:**
17: **Estimar transições entre estados ocultos:**
18: $T[i, j] \leftarrow \alpha$; contar transições $destino_t \rightarrow destino_{t+1}$; normalizar
19: **Estimar emissões:**
20: $E[i, j] \leftarrow \alpha$; contar emissões $destino \rightarrow$ símbolo(dia+origem); normalizar
21: **Estimar distribuição inicial dos dados:**
22: $\pi[i] \leftarrow$ contar destinos iniciais por dia no treino $+\alpha$; normalizar
23: $hmm \leftarrow$ HMM$(T, E, \pi, states, symbols)$
24: $y\_true \leftarrow []$; $y\_pred \leftarrow []$
25: **for** cada observação no teste **do**
26: $obs \leftarrow$ símbolo atual (dia+origem); $real \leftarrow$ destino real
27: $\_, path \leftarrow hmm.viterbi([obs], \pi)$ ▷ π baseado nos dados
28: **if** *path* é lista/array não vazio: $pred \leftarrow path[-1]$
29: **elif** *path* é string: $pred \leftarrow path$
30: **else:** $pred \leftarrow \arg\max_i E[i, obs]$ (fallback)
31: $y\_true.append(real)$; $y\_pred.append(pred)$
32: **end for**
33: **if** $|y\_true| > 0$: $fold\_precs.append$(precision(y_true, y_pred))
34: **except: continue**
35: **end for**
36: **if** $|fold\_precs| \geq 3$ **then** $\alpha_results[\alpha] \leftarrow \{$mean$(fold_precs), fold_precs\}$
37: **end if**
38: **end for**
39: **if** $\alpha\_results \neq \emptyset$ **then**
40: Selecionar α com maior precisão média
41: Armazenar resultados e precisões dos folds
42: **end if**
43: **end for**
44: **if** $all\_repeat\_results \neq \emptyset$ **then**
45: Calcular μ, σ, Shapiro-Wilk, moda do α
46: $all\_results[vehicle\_id] \leftarrow$ estatísticas consolidadas
47: **end if**
48: **end for**
49: Análise global Shapiro-Wilk; visualizações; salvar folds individuais
50: **return** *all_results*

Fonte: Elaboração própria.

Nas linhas 8-43, para cada repetição (linha 9), o algoritmo cria uma nova partição K-Fold (linha 10) e testa diferentes valores de α (linha 11). Dentro de cada *fold* (linha 13), o modelo HMM é construído estimando três componentes: a matriz de transição T entre estados ocultos (linhas 17-18), a matriz de emissão E que relaciona estados a observações (linhas 19-20), e a distribuição inicial π (linhas 21-22). A suavização de Laplace com parâmetro α é aplicada para evitar probabilidades zero.

Nas linhas 24-36, para cada observação no conjunto de teste (linha 25), o algoritmo de Viterbi é aplicado (linha 27) para encontrar a sequência mais provável de estados ocultos. Caso o Viterbi falhe, utiliza-se um método de *fallback* baseado na máxima probabilidade de emissão (linha 30). A precisão é calculada comparando os destinos preditos com os reais (linha 33).

Nas linhas 39-47 Após avaliar todos os valores de α em todos os *folds*, seleciona-se o α com maior precisão média (linha 40). Os resultados de todas as repetições são consolidados (linhas 44-46), calculando-se estatísticas como média, desvio padrão e teste de Shapiro-Wilk para normalidade.

Por fim, nas linhas 49-50, o algoritmo conclui com análises globais, incluindo teste de Shapiro-Wilk sobre todos os resultados, geração de visualizações e salvamento dos resultados individuais por *fold* para análises posteriores.

4.6 Análises estatísticas

A validação das modelagens desenvolvidas, que corresponde à sequência em código dos dados de amostra, seguidamente para Cadeias de Markov e Cadeias Ocultas de Markov, tendo, em seu escopo, apenas o suficiente para uma avaliação estatística adequada.

Nessa etapa final, assim, foi feito primeiramente o teste de Shapiro-Wilk, confirmando a Normalidade dos dois conjuntos de precisões médias (para Cadeias de Markov e HMM). Para, em seguida, ser aplicado *t-Student* nos resultados.

Algoritmo 5 Análise Comparativa Markov vs HMM com Teste t-Student Pareado

Require: CSVs com resultados individuais dos folds para Markov e HMM
Ensure: Análise estatística comparativa e visualizações
1: $markov\_df \leftarrow$ carregar 'resultados_markov_todos_folds_individuais.csv'
2: $hmm\_df \leftarrow$ carregar 'resultados_hmm_todos_folds_individuais.csv'
3: **Calcular médias por veículo:**
4: $markov\_means \leftarrow$ agrupar $markov\_df$ por veículo e calcular média das precisões
5: $hmm\_means \leftarrow$ agrupar hmm_df por veículo e calcular média das precisões
6: $paired\_data \leftarrow$ merge($markov\_means$, hmm_means) por veículo
7: $paired\_data$[difference] $\leftarrow$ $paired\_data$[markov] $-$ $paired\_data$[hmm]
8: **Teste global:** $t\_stat, p\_global \leftarrow$ ttest_ind($paired\_data$[markov], $paired\_data$[hmm])
9: **Garantir pareamento correto dos folds:**
10: $fully\_paired \leftarrow$ merge($markov\_df$, hmm_df) por [veículo, repeat, fold]
11: $p\_values \leftarrow []$; $n\_pairs \leftarrow []$; $test\_type \leftarrow []$
12: **for** cada $vehicle$ em $paired\_data$ **do**
13: $vehicle\_paired \leftarrow fully\_paired[vehicle]$
14: **if** $|vehicle\_paired| > 1$ **then**
15: $\_, p\_val \leftarrow$ ttest_ind($vehicle\_paired$[markov], $vehicle\_paired$[hmm])
16: $p\_values.append(p\_val)$; $n\_pairs.append(|vehicle\_paired|)$
17: $test\_type.append$('paired')
18: **else**
19: $markov\_v \leftarrow markov\_df[vehicle]$[precision]
20: $hmm\_v \leftarrow hmm\_df[vehicle]$[precision]
21: **if** $|markov\_v| > 1$ e $|hmm\_v| > 1$ **then**
22: $\_, p\_val \leftarrow$ ttest_ind($markov\_v$, hmm_v) ▷ Teste não-pareado
23: $p\_values.append(p\_val)$; $n\_pairs.append(0)$; $test\_type.append$('unpaired')
24: **else**
25: $p\_values.append$(NaN); $n\_pairs.append(0)$; $test\_type.append$('none')
26: **end if**
27: **end if**
28: **end for**
29: $paired\_data$[p_value] $\leftarrow$ p_values; $paired\_data$[n_pairs] $\leftarrow$ n_pairs
30: $paired\_data$[test_type] $\leftarrow$ $test\_type$
31: **Visualização 1 - Precisões por veículo:**
32: Criar barras horizontais para Markov (azul) e HMM (vermelho)
33: Adicionar valores nas barras e linhas de média global
34: **Visualização 2 - P-values por veículo:**
35: Criar barras horizontais com escala log
36: Colorir: vermelho se $p < 0.05$, cinza caso contrário
37: Marcar testes não-pareados com asterisco
38: Adicionar linha de significância em $p = 0.05$
39: **Resumo estatístico:**
40: $n\_significant \leftarrow$ contar veículos com $p < 0.05$
41: $n\_paired \leftarrow$ contar testes pareados realizados
42: Imprimir estatísticas: médias globais, p-value global, proporção de significativos
43: Salvar $paired\_data$ em 'analise_markov_hmm_pareado.csv'
44: Salvar visualizações em 'precisao_pvalues_markov_hmm_horizontal.png'
45: **return** análise completa e arquivos gerados

Fonte: Elaboração própria.

O Algoritmo 5 tem por finalidade uma análise estatística comparativa entre os modelos de Cadeias de Markov e HMM, utilizando conforme citado antes, o teste *t-student* pareado, para avaliar se há diferença estatisticamente significativa entre as precisões com os dois modelos.

Nas linhas 1 e 2, carregam-se os arquivos CSV contendo todas as precisões individuais dos *folds* de cada modelo. Estes arquivos foram gerados pelos algoritmos anteriores e contêm dados granulares de cada *fold*, de cada repetição, para cada veículo.

Das linhas 3 a 7, preparam-se os dados para análise. Calculam-se as médias das precisões por veículo para cada modelo (linhas 4-5), realiza-se a junção dos dados para garantir correspondência entre veículos (linha 6), e calcula-se a diferença entre as precisões de Markov e HMM (linha 7).

Na linha 8, executa-se o teste *t-student* pareado global, comparando as médias de todos os veículos entre os dois modelos. Este teste fornece o valor de p, de forma global, que indica se há diferença significativa considerando todos os veículos em conjunto.

Das linhas 9 a 10, implementa-se uma etapa relevante: o pareamento correto dos dados. A junção é realizada considerando não apenas o veículo, mas também a repetição e o *fold* específicos, garantindo que as comparações sejam feitas entre resultados obtidos com os mesmos particionamentos dos dados.

O *loop* principal (linhas 12 a 28) realiza testes individuais por veículo. Para cada veículo:

- Se há dados pareados suficientes (linha 14), aplica-se o teste *t-student* para amostras independentes (linha 15), registrando o p-value, número de pares e tipo de teste (linhas 16-17).
- Caso contrário, tenta-se um teste *t-student* não-pareado como *fallback* (linhas 19-23), útil quando o pareamento completo não é possível.
- Se nenhum teste é viável, registra-se NaN (linha 25).

Nas linhas 29-30, os resultados dos testes individuais são incorporados ao *dataframe* principal, incluindo os valores de p, número de pares e tipo de teste utilizado.

A visualização ocorre em duas partes principais:

Gráfico 1 - Precisões por veículo (linhas 32-33): Cria-se um gráfico de barras horizontais comparando as precisões médias de Markov (azul) e HMM (vermelho) para cada veículo. Incluem-se os valores exatos nas barras e linhas verticais indicando as médias globais de cada modelo.

Gráfico 2 - P-values por veículo (linhas 35-37): É utilizada a escala logarítmica para melhor visualização dos valores de p, que são demonstradas em várias ordens de magnitude. Barras vermelhas indicam diferenças significativas ($p < 0.05$), enquanto barras cinzas indicam não-significância. Testes não-pareados são marcados com asterisco para transparência metodológica.

Das linhas 40 a 41, calculam-se estatísticas resumidas: número de veículos com diferença significativa, proporção de testes pareados vs. não-pareados, e outras métricas relevantes. Estas informações são apresentadas tanto nos gráficos quanto no console.

Finalmente, nas linhas 43-44, os resultados são persistidos em arquivo CSV para análises posteriores e as visualizações são salvas em alta resolução (300 DPI) para uso em publicações.

O algoritmo prioriza testes pareados quando possível, mas adaptando-se com testes não-pareados quando necessário. Dessa forma, permite interpretar com base, principalmente nos Gráficos, a partir dos resultados com Cadeias de Markov e HMM e suas precisões médias orga-

nizadas por *folds* - ainda que, no primeiro gráfico, cada precisão esteja, para melhor visualização, organizada por veículo.

4.7 Conclusão da Metodologia

Assim, conclui-se este Capítulo que explicou os procedimentos adotados em dois conjuntos de etapas: o primeiro, culminado na obtenção das amostras; e o segundo, no uso dessas amostras para os modelos. No próximo Capítulo, são apresentados os Resultados com base nesses procedimentos descritos.

5 RESULTADOS

Em uma breve contextualização acerca dos dados, pode-se dizer que, a partir de Oh, Leblanc e Peng (2022), autores da pesquisa original que resultou na criação da base de dados do VED, demonstrou-se que eles podem ser utilizados para identificar trajetórias, que poderão fazer parte de uma amostra, para a avaliação experimental dos modelos propostos de predição de destinos. Entretanto, para o caso desta pesquisa, tem-se que, a partir de uma análise, tais dados podem ser estudados para a obtenção de trajetórias, considerando as colunas de coordenadas geográficas, identificador de viagem e *timestamp* (em microssegundos).

Os dados foram coletados em acordo entre a Universidade de Michigan e o Laboratório Nacional de Idaho. Isso com o objetivo de estudar o comportamento dos usuários quanto ao consumo de energia e os potenciais de economia de tecnologias *eco-driving*.

Os mesmos autores evidenciam que são 383 veículos[1] em trajetos percorridos entre o período de 1 de novembro de 2017 e 9 de novembro de 2018, em Ann Arbor (cidade ao sul de Michigan). Os veículos são de diferentes tipos, mas neste trabalho esses tipos não são relevantes. Basicamente, variam de veículos de passageiros (carros comuns) até caminhões leves. O total percorrido, durante 1 ano e 8 dias foi de 373964 milhas.

Com os metadados formados por uma tabela [2] e o conjunto de dados em formato "csv", houve um processo de desidentificação para o trabalho dos pesquisadores, que consistiu em: *Random Fogging*, *Geofencing* e *Major Intersections Bounding*, de acordo com Oh, Leblanc e Peng (2022).

Ainda sobre os dados, houve algumas dificuldades, a princípio, principalmente com a forma de leitura do *timestamp* - e isso definiu o início do percurso dos experimentos, com o tratamento prévio dos dados através da derivação da coluna *datetime*.

Esse conjunto possui dois tipos de arquivos tabulares: dados dinâmicos e dados estáticos. Os dados estáticos podem ser vistos como metadados. Contêm, segundo Oh, Leblanc e Peng (2022), o tipo do veículo (em geral, se é elétrico, híbrido ou convencional), a classe do veículo (carro comum, SUV ou caminhão leve), configuração do motor, cilindrada do motor, transmissão, rodas e peso.

Os dados utilizados nesta Pesquisa têm por escolha 99 dos veículos da célula 6+ por 6+ conforme visto na matriz já apresentada na Metodologia, na Figura 11. Porém, apenas 23, desses 99, foram efetivamente utilizados na amostragem. O motivo dessa filtragem é devido tanto à diminuição da complexidade do processamento, de 384 para 99 veículos, como - e principalmente

[1] Na verdade, 384, conforme correção realizada em <https://github.com/gsoh/VED>.

[2] Dados estáticos com tipo do veículo, classe do veículo, configuração do motor, cilindrada do motor, transmissão, rodas e peso.

- devido à consideração das subtrajetórias e daqueles veículos com maior número de rotas - estas possuindo o maior número de repetições, permitindo uma melhora no desempenho das precisões partindo-se do pressuposto de que quanto mais repetições melhor o desempenho por existirem mais dados históricos para as predições.

Na Tabela 12, está apresentada a quantidade de trajetórias geradas para cada veículo, antes da filtragem pelo limiar 50:

Tabela 12 – Contagem Geográfica por ID de Veículo - Valores Agrupados de Repetição

Veículo	**Contagem**	**Veículo**	**Contagem**	**Veículo**	**Contagem**
12	22	273	36	459	41
128	16	275	10	463	45
133	6	276	142	465	39
135	19	278	8	476	23
140	4	282	17	480	27
145	29	283	51	488	29
155	22	285	13	490	14
156	32	292	50	494	26
180	47	301	42	507	27
181	83	303	8	519	4
184	23	304	7	521	52
185	78	307	23	529	17
200	2	308	4	530	47
201	47	309	89	531	75
202	18	311	17	538	8
203	6	323	88	540	12
205	29	340	91	546	35
208	34	346	11	547	13
213	35	347	29	557	37
218	21	356	16	560	195
220	38	359	19	561	135
223	131	366	50	564	39
228	6	367	29	565	66
231	50	371	104	575	98
232	9	372	39	577	21
242	29	382	8	581	22
249	2	384	29	584	33
251	27	385	11	591	17
258	5	388	147	592	42
259	26	389	2	595	36
265	19	450	154	597	61
266	5	452	51	603	6
267	31	457	88	N/A	N/A

Fonte: Elaborado pelo autor.

Após a filtragem pelo limiar 50, têm-se conforme a Tabela 13.

Tabela 13 – Contagem Geográfica por ID de Veículo - Valores Agrupados de Repetição

Veículo	**Contagem**
181	83
185	78
223	131
231	50
276	142
283	51
292	50
309	89
323	88
340	91
366	50
371	104
388	147
450	154
452	51
457	88
521	52
531	75
560	195
561	135
565	66
575	98
597	61

Fonte: Elaborado pelo autor.

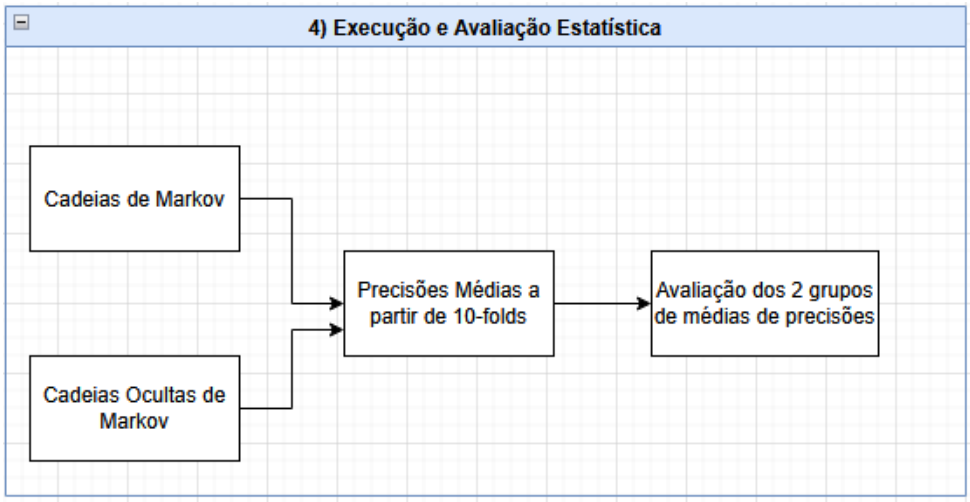
Buscou-se atuar sobre um conjunto elaborado de dados de forma equivalente. Para ambos os casos, foram considerados as repetições somente geográficas, e não temporais.

Na próxima seção, há uma descrição sobre o planejamento experimental.

5.1 Planejamento Experimental

Na Figura 14, há o diagrama de como foi executada a avaliação para 23 veículos, com 10-folds. Para uso de HMM, o algoritmo utilizado foi o de Viterbi (considerando neste caso que para essa avaliação, apenas esse modelo gerou um resultado normalizado, gerando a necessidade de uso de t-student, conforme instruções de Sirqueira et al. (2020)).

Figura 14 – Finalização.



Fonte: Elaboração própria.

Ao final, o Método permite a avaliação com *t-student*, levando a um quadro final passível de interpretação e validação, contendo os pares com significância estatística e os pares sem significância estatística.

Outro fator é a Questão de Pesquisa e a Hipótese, de que há diferença significativamente relevante entre as técnicas. A Questão desta Pesquisa consiste em "Existe diferença quanto ao uso dos modelos de predição de destino com base em Cadeias de Markov e HMM, no contexto de tráfego urbano e uso de veículos individuais, com relação a uma métrica como a ?". Segue-se, assim, a configuração de teste a determinar se H0 = Não há diferença ou H1 = Há diferença.

E, quanto à escolha de Precisão no lugar de, por exemplo, Acurácia, deveu-se a uma questão de escolha de pesquisa, considerando-se a fórmula da Precisão em que os verdadeiros positivos são divididos pela soma dos verdadeiros positivos com os falsos positivos. Neste caso é considerada a reprodutibilidade, a consistência dos valores, e não necessariamente a proximidade com um valor verdadeiro.

5.2 Resultados Obtidos

Os resultados constam na Figura 15. O gráfico à esquerda consiste na Precisão Média por Veículo, e as barras se dividem entre Markov, na cor verde, e HMM, na cor vermelha. No gráfico à direita, têm-se os valores de p por veículo, estando em vermelho apenas os valores de p estatísticamente relevantes. Por fim, a linha tracejada indica o limite entre valores de p significativos e não significativos.

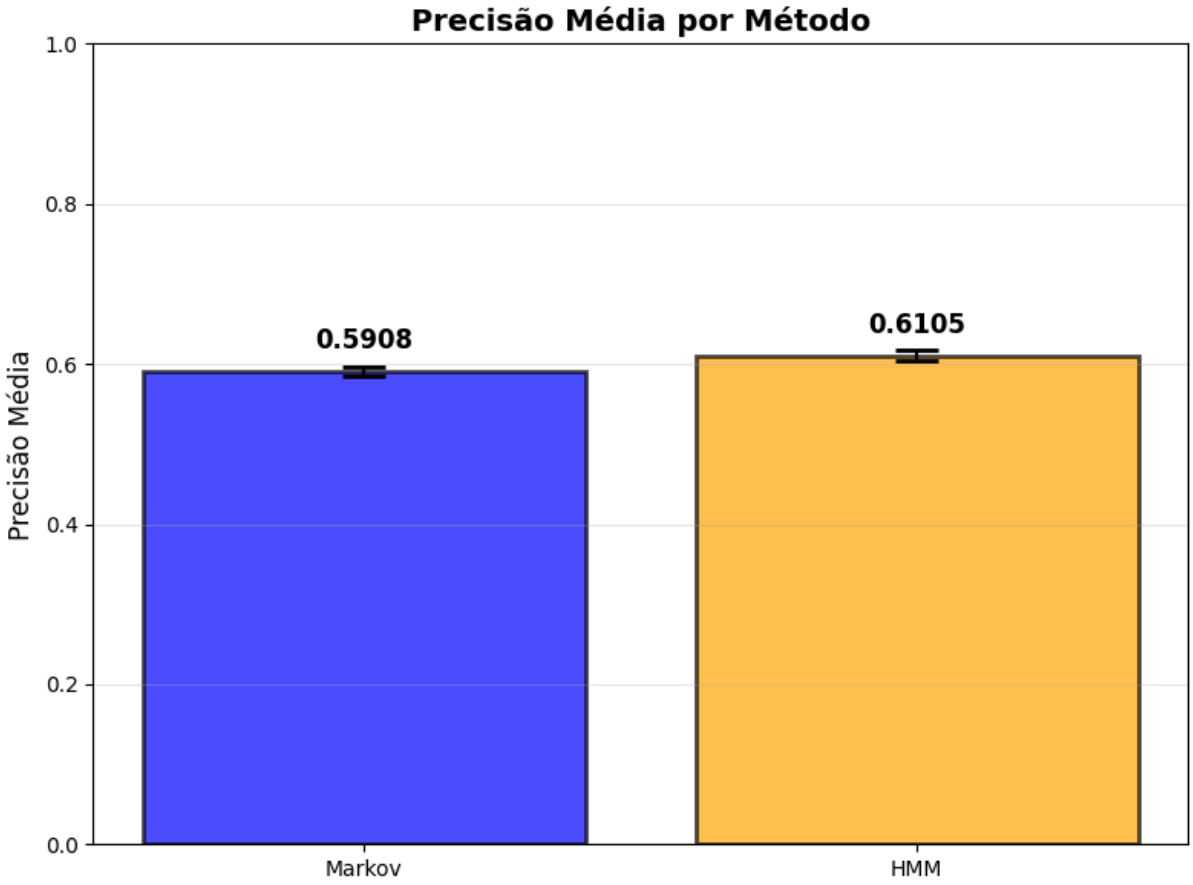
Figura 15 – Precisões Médias para Cadeias de Markov e HMM, e os valores de p.

Fonte: Elaboração própria.

Nesse caso, visualmente, percebem-se que as médias das precisões para cada veículo parecem melhores em se tratando do algoritmo que implementa o modelo HMM. Considerando que as médias globais das precisões foram conforme na Figura 16, essa percepção se torna mais

evidente. Globalmente, considerando os 23 veículos, o valor global de p foi de 0,022088, ou seja, relevante estatisticamente.

Figura 16 – Precisões Médias Globais para Cadeias de Markov e HMM.

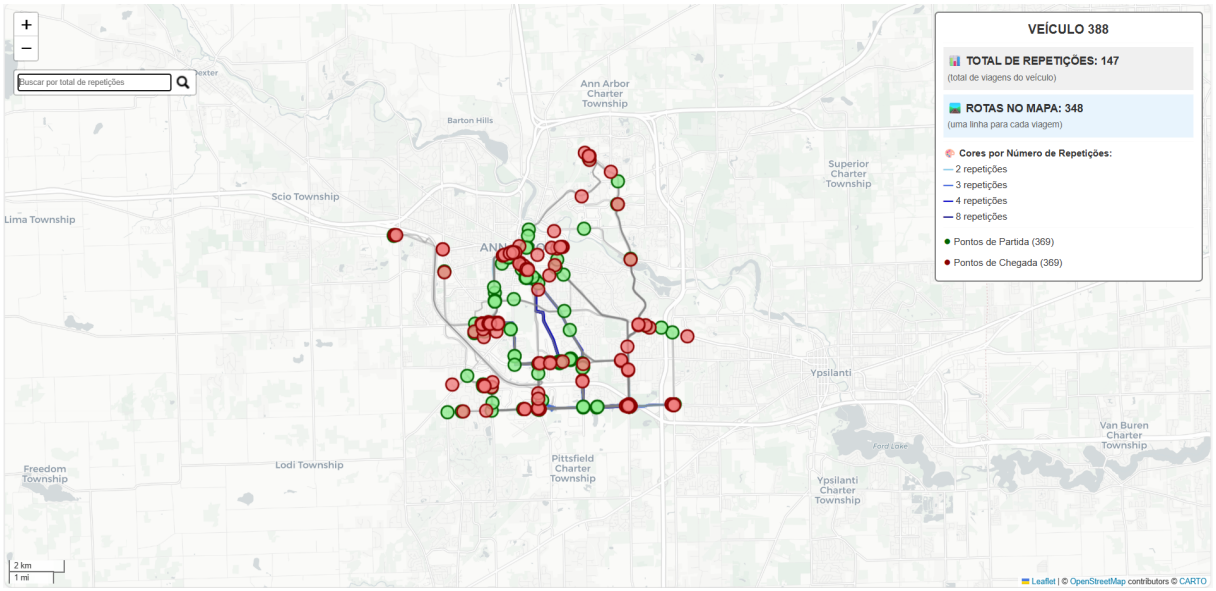


Fonte: Elaboração própria.

5.3 Visualização

Com base em uma visualização a partir da biblioteca *Folium*, foi possível uma orientação prévia sobre o comportamento dos dados para uma pequena parcela dos dados do Veículo 388, para só após ser realizado o processamento definitivo. A Figura 17, mostra esse mapa interativo.

Figura 17 – Mapa interativo.



Fonte: Elaboração própria.

5.3.1 Quadro Comparativo e Análise dos Resultados

Basicamente, os resultados são consolidados na Tabela 14, cujas linhas em vermelho são as que atestam H1 (a hipótese de que há diferenças estatisticamente significativas entre as médias das precisões com Cadeias de Markov e HMM).

Tabela 14 – Precisões médias e p-valores para veículos em Cadeias de Markov e HMM

Veículo	Markov	Markov Oculto	p-value
181	0.8144	0.6985	1.61e-04
185	0.4393	0.2655	1.30e-10
223	0.4934	0.6608	7.72e-08
231	0.3267	0.5261	4.08e-08
276	0.5222	0.4278	1.08e-04
283	0.5190	0.5547	3.73e-01
292	0.8003	0.6687	3.84e-04
309	0.5175	0.5856	2.83e-02
323	0.6508	0.7938	2.24e-05
340	0.6697	0.7213	1.17e-01
366	0.7434	0.7210	5.05e-01
371	0.3568	0.5246	2.61e-09
388	0.3826	0.5128	7.66e-11
450	0.8761	0.8298	8.25e-03
452	0.8698	0.8742	8.98e-01
457	0.4342	0.4005	2.03e-01
521	0.5566	0.2136	5.58e-22
531	0.8165	0.9346	9.23e-07
560	0.2850	0.0172	1.84e-51
561	0.5563	0.7111	1.68e-08
565	0.9034	0.9762	4.26e-05
575	0.5123	0.6546	3.00e-06
597	0.5431	0.7690	1.10e-09

Fonte: Elaboração própria.

Entende-se, desses resultados da Tabela 14, que 5 veículos apresentaram diferenças não estatisticamente significativas, com uma diferença no máximo um pouco maior que 5% entre as precisões médidas do veículo 340, entre Markov e Markov Oculto (HMM). Esse resultado jultifica, mais consolidadamente, o uso da filtragem em 50 repetições (valores abaixo desse patamar tendem a precisões médias mais baixas, conforme experimentos preliminares e, consequentemente, foram descartados neste trabalho), por trazer resultados por veículo também relevantes e explicáveis estatisticamente, num intervalo de valores, entre 0 e 1, condizente com os p-values. Por exemplo: para os casos não significativos (que foram 5), a diferença costuma ser de aproximadamente desde menos de 1% (é caso do veículo 452) a pouco mais que 5% entre Markov e HMM, provavelmente devido a como os algoritmos dos modelos interpretaram, em particular, os dados, considerando ao menos Markov ou HMM.

Além disso, HMM traz informações contextuais, mais aderentes à natureza humana conforme tempo e espaço. Dessa forma, é preciso o cuidado para não considerar "a melhor" técnica de forma puramente quantitativa e sem correspondência à realidade dos fenômenos estudados, mas as particularidades de cada uma para cada conjunto de amostras e os objetivos

preditivos de acordo com movimentações de seres humanos.

Isso implica que é necessário aprofundar a pesquisa, posteriormente, para dados contextuais e semânticos. Porém, é possível dizer que o Método desenvolvido é apropriado para uma pesquisa básica sobre como devem ser estruturados os dados de amostra para conhecer os algoritmos preditivos.

5.4 Conclusão dos Resultados

Em termos de finalização deste Capítulo, foram apresentados e explicados os resultados de acordo com os cenários avaliados e explicados inicialmente. Esses cenários voltaram-se às técnicas a serem avaliadas com os dados elaborados: Cadeias de Markov e HMM. E, em relação aos *scripts*, estão disponíveis, junto aos dados elaborados e aos relatórios simplificados, em: Github: Insumos-Mestrado.

Quanto às Considerações Finais, será visto como esse Método pode auxiliar em uma continuidade de pesquisas que prevejam destinos para seres humanos, além de questões como lições aprendidas e ameaças encontradas na validação desta pesquisa. E, em relação aos Apêndices A e B, respectivamente, há uma demonstração gráfica dos grafos, para certo número de estados, considerando o processamento com o pyDTMC e os dados do VED para o veículo 388.

6 CONSIDERAÇÕES FINAIS

Esta pesquisa foi centrada em realizar uma análise comparativa entre Cadeias de Markov e HMM, com foco em predição de destinos, no contexto dos veículos urbanos individuais (automóveis de diferentes portes), com o intuito de auxiliar outras pesquisas de natureza básica ou aplicada.

As próximas seções tratam das Conclusões e Contribuições(seção 6.1), Desafios (seção 6.2) e Propostas de Trabalhos Futuros (seção 6.3), finalizando este trabalho.

6.1 Conclusões e Contribuições

Para contemplar o objetivo principal deste trabalho, o de "Realizar uma análise comparativa dos modelos de predição de destino com base em Cadeias de Markov e Cadeias Ocultas de Markov, no contexto de usuários que usam seus veículos de maneira individual", foi criado um conjunto de etapas bem definidas, de modo a manter coesão e responsabilidades para cada componente.

- A principal contribuição deste trabalho consiste em apresentar uma análise comparativa entre duas técnicas de predição: Markov e HMM, demonstrando diferenças entre estas, embasadas por testes estatísticos;
- Com seus resultados, foi possível perceber a utilidade para a criação de subtrajetórias, a partir do método *Stop Splitting* presente no *MovingPandas*, em sua versão 0.19.0, para a segmentação de trajetórias mediante um conjunto de parâmetros definido para o caso do VED. Essa percepção surgiu da regularidade dos resultados, no sentido de permitirem uma comparação adequada;
- Há espaço, para esta pesquisa, para a inclusão de outras variáveis contextuais, a exemplo de verificar se o dia da trajetória é dia de semana ou de fim de semana, para enriquecimento contextual, desde que sendo respeitadas as restrições para o conjunto de funções elaboradas para HMM, no *script* onde há as predições;
- Entende-se, também, que há abertura para enriquecimento semântico para trabalhos sobre Predição de Destinos, principalmente sendo possível integrar o VED a dados do *OpenStreetMap* mediante bibliotecas como OSMNx.
- Outra contribuição diz respeito às formas de visualização de movimentação de objetos trazidos com o auxílio do *Folium* para mapas interativos, considerando que a natureza pontual dos registros, antes, torna-se linear com o uso do *MovingPandas*;

- Derivação da coluna *datetime* com o auxílio de métodos da biblioteca Pandas, a partir de algumas inferências fundamentais, explicadas neste texto. Além da derivação das colunas referentes a dias e turnos com o auxílio do *PyMove*;
- Elaboração de uma forma de encontrar trajetórias repetidas, que auxiliaram, para as técnicas escolhidas, na análise preditiva[1].

6.2 Desafios

Esta seção apresenta os limites ou desafios deste trabalho, e considera:

- A definição de uma semântica das trajetórias (ou seja, o que de fato foi uma viagem, relacionada a determinada época para um determinado destino existente nessa época);
- Acesso a um *dataset* gratuito, anonimizado, ou a dados em tempo quase real a partir de sensores, em comum acordo com a Secretaria de alguma Prefeitura brasileira;
- Encontrar, a partir de diferentes contextos e bases de dados, um formato compatível das amostras finais para submeter aos modelos referentes às Cadeias de Markov e HMM;
- Tesselação sem considerar diferenças hierárquicas de escala. Ou seja, a grade é fixa para qualquer escala, independentemente do tamanho da trajetória;

6.3 Propostas de Trabalhos Futuros

Esta seção contempla os trabalhos futuros que podem ser originados a partir deste trabalho. Abrangendo, desde o contexto e os modelos escolhidos, até novas variáveis semânticas e/ou formas de tesselação e preparação das amostras.

Primeiramente, referente ao conjunto de dados escolhido, é uma proposta considerar processamento com base em sensores, em tempo próximo ao real, a partir de convênios com órgãos de prefeituras. Além disso, os dados contextuais a serem incluídos podem variar, como na possibilidade de inclusão de períodos do dia, topografia, se é um feriado ou dia comum de trabalho etc.

Quanto aos dados semânticos, por exemplo, novas pesquisas com base nesta podem enriquecer a partir da resolução espacial das Origens e dos Destinos, que vão determinar objetivos para uma determinada trajetória, como "chegar a Campina Grande", "chegar a um bairro de Campina Grande" ou "chegar a um determinado endereço, considerando ser um Domingo ou o período da tarde".

[1] Isso ocorreu com o uso do primeiro *script*, referente à contagem de repetições de trajetórias, presente na url referida na Seção 5.4.

Sobre a visualização, também é possível desenvolver uma interação com dados enriquecidos para predição e às próprias predições. O que significaria um sistema, por exemplo, que aceitassa ampla variedade de *datasets*, com dados de sensores em tempo próximo ao real, com os campos de Longitude e Latitude ou *Geometry* e a contagem temporal (um campo de *datetime*), para uma variedade de modelos, considerando dados contextuais e semânticos variados.

Quanto ao conjunto de dados específicos do VED, é possível investigar a possibilidade de uso com dados mais esparsos, apenas *checkins* e *checkouts* de táxis. Isso não diretamente com o uso do Método desenvolvido, mas como uma possibilidade de busca por inferência de trajetórias com base em dados mais esparsos. Por fim, em relação à discretização dos dados, a utilização de outras técnicas de codificação, mais elaboradas, para *geohashing*, que faz o mesmo, mas considerando não grades de mesma resolução e convertendo as coordenadas geográficas em códigos alfanuméricos cujo tamanho do código possa equivaler a uma resolução espacial mais refinada ou grosseira, ou seja, respectivamente, para áreas menores ou maiores. Isso, naturalmente, levará a modificações em todo o Método. Dois exemplos são as bibliotecas Geohash2 e OpenLocationCode. A primeira, presente no PostGIS e no *Oracle Spatial*, dentre suas funções espaciais.

Por fim, em relação a este trabalho, para predizer destinos, é também possível uma pesquisa que considere outras técnicas de segmentação. Ou seja: predição de trajetórias/destinos com estudos de comportamento humano. Isso reunindo todo esse conjunto e desenvolver sistemas preditivos, no caso da proposição de uma pesquisa aplicada, com uma visualização similar ao que pode ser promovido para Web ou *Mobile*. Isto é, um estudo reunindo padrões de comportamento de motoristas, dados que podem ser mais esparsos, outras formas de discretização de coordenadas geográficas, técnicas de anonimização, utilização de predições coletivas ao invés de individuais, para, com mapas interativos, haver uma melhor interação entre usuário e sistema. Com outro módulo do mesmo sistema, o pesquisador ou administrador enxergaria e poderia interagir com a escolha de mais técnicas preditivas, técnicas de amostragem e localidades, para facilitar seu trabalho de administração ou de pesquisa.

Enfim, são sugestões de trabalhos futuros que ainda precisariam de adequada validação. Mas, com esta pesquisa, já é possível vislumbrar possibilidades nesses quatro conjuntos temáticos.

REFERÊNCIAS BIBLIOGRÁFICAS

ALVAREZ-GARCIA, J. A. et al. Trip destination prediction based on past gps log using a hidden markov model. *Expert Systems with Applications*, v. 37, p. 8166–8171, 2010. Citado na página 37.

AMIN, S. et al. What will you do for the rest of the day? an approach to continuous trajectory prediction. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, v. 2, n. 4, 2018. Citado 2 vezes nas páginas 17 e 36.

ANDERSON, T. W.; GOODMAN, L. A. Statistical inference about markov chains. *Ann. Math. Statist.*, Institute of Mathematical Statistics, v. 28, n. 1, p. 89–110, mar 1957. Citado na página 28.

ANKUR, A.; PANDA, A. *Hands-On Markov Models with Python*. [S.l.]: Packt Publishing Ltd, 2018. Citado 2 vezes nas páginas 22 e 30.

ARAúJO, F. R. D. et al. TEMMUS: A Mobility Predictor based on Temporal Markov Model with User Similarity. *Simpósio Brasileiro De Redes De Computadores E Sistemas Distribuídos (SBRC)*, 2019. Citado 5 vezes nas páginas 30, 34, 35, 36 e 41.

ARIAS, E. F.; GUINOT, B. *COORDINATED UNIVERSAL TIME UTC : HISTORICAL BACKGROUND AND PERSPECTIVES*. Bureau International des Poids et Mesures and SYRTE, Observatoire de Paris: [s.n.], 2004. Citado na página 22.

BESSE, P. C. et al. Destination prediction by trajectory distribution-based model. *IEEE Transactions On Intelligent Transportation Systems: A Publication Of The IEEE Intelligent Transportation Systems Council*, v. 19, p. 2470–2481, 2018. Citado na página 34.

CHEN, B.; HONG, Y. Testing for the markov property in time series. *Econometric Theory*, v. 28, p. 130–178, 2012. Citado na página 30.

CHO, S.-B. Exploiting machinelearning techniques for location recognition and prediction with smartphone logs. *Neurocomputing*, v. 176, p. 98–106, 2016. Citado na página 37.

DRUCK, S. et al. *Análise espacial de dados geográficos*. Brasília: Embrapa: [s.n.], 2004. Citado na página 21.

FROEHLICH, J.; KRUMM, J. Route prediction from trip observations. *SAE Technical Paper*, v. 6, n. 3, 2015. Citado na página 35.

GAO, J. *Fundamentals Of Spatial Analysis And Modelling*. 1. ed. Boca Raton: CRC, 2022. Citado na página 20.

GIANNOTI, F. et al. Trajectory pattern mining. In: *Proceedings of the Knowledge Discovery and Data Mining Conference (KDD'07).* [S.l.]: Association for Computing Machinery, 2007. p. 330–339. Acesso em: 12 aug. 2007. Citado na página 19.

GOLD, C. Tessellations in gis: Part i—putting it all together. *Geo-spatial Information Science*, v. 19, n. 1, p. 9–25, 2016. Citado 2 vezes nas páginas 19 e 20.

GRASER, A. et al. Deep learning from trajectory data: a review of neural networks and the trajectory data representations to train them. In: . Workshop on Big Mobility Data Analysis BMDA2023 in conjunction with EDBT/ICDT 2023, 2023. Disponível em: <https://ceur-ws.org/Vol-3379/BMDA_2023_paper_7556.pdf>. Acesso em: 28 dez. 2024. Citado na página 15.

GREWAL, J. K.; KRZYWINSKI, M.; ALTMAN, N. Markov models — hidden markov models. *Nature*, v. 16, p. 793–796, 2019. Citado 2 vezes nas páginas 23 e 26.

GREWAL, J. K.; KRZYWINSKI, M.; ALTMAN, N. Markov models—markov chains. *Nature*, v. 16, p. 661–664, 2019. Citado na página 23.

HORNSBY, K. S.; COLE, S. Modeling moving geospatial objects from an event-based perspective. *Transactions in GIS*, v. 11, n. 4, p. 555–573, 2007. Disponível em: <https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1467-9671.2007.01060.x>. Citado na página 32.

JAVIDROOZI, V.; SHAH, H.; FELDMAN, G. Urban computing and smart cities: Towards changing city processes by applying enterprise systems integration practices. *IEEE Access*, v. 7, p. 108023–108034, 2019. Citado na página 15.

JUNIOR, J. B. F.; DUTRA, J. F.; NETO, F. D. N. Evaluation of trajectory and destination prediction models: a systematic classification and analysis of methodologies and recent results. *Journal of Internet Services and Applications*, v. 15, n. 1, p. 474–484, 2024. Citado 5 vezes nas páginas 15, 30, 33, 35 e 38.

JURAFSKY, D.; MARTIN, J. H. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition*. 3rd. ed. Draft version, 2024. Third edition draft, acessado em 2025. Disponível em: <https://web.stanford.edu/~jurafsky/slp3/>. Citado na página 16.

KELLEHER, J. D.; NAMEE, B. M.; D'ARCY, A. *Fundamentals of Machine Learning for Predictive Data Analytics: algorithms, worked examples, and case studies*. USA e England: MIT Press, 2020. Citado 2 vezes nas páginas 21 e 31.

KITCHENHAM, B.; CHARTERS, S. *Guidelines for performing systematic literature reviews in software engineering.* 2007. Citado na página 33.

LASSOUED, Y. et al. A hidden markov model for route and destination prediction. In: *2017 IEEE 20th International Conference On Intelligent Transportation Systems (ITSC)*. [S.l.: s.n.], 2017. p. 1–6. Citado 4 vezes nas páginas 17, 35, 37 e 41.

LI, K. et al. Route search and planning: A survey. *Big Data Research*, v. 26, 2021. Citado 2 vezes nas páginas 15 e 19.

LING, C.; LV, M.; GENCAI, C. A system for destination and future route prediction based on trajectory mining. *Pervasive and Mobile Computing*, v. 6, p. 657–676, 2010. Citado na página 36.

LING, C. et al. A personal route prediction system based on trajectory data mining. *Information Sciences*, v. 181, p. 1264–1284, 2010. Citado na página 36.

LIU, L. et al. Contextualized spatial–temporal network for taxi origin-destination demand prediction. *IEEE Transactions on Intelligent Transportation Systems.*, v. 20, p. 3875–3887, 2019. Citado na página 22.

MovingPandas Contributors. *MovingPandas: A Python library for movement data analysis and visualization*. [S.l.], 2024. Versão 0.17.1. Disponível em: <https://movingpandas.org>. Acesso em: 2024. Citado 2 vezes nas páginas 45 e 46.

OH, G.; LEBLANC, D. J.; PENG, H. Vehicle energy dataset (ved), a large-scale dataset for vehicle energy consumption research. *IEEE TRANSACTIONS ON INTELLIGENT TRANSPORTATION SYSTEMS*, v. 23, n. 4, 2022. Citado na página 65.

PyMove Development Team. *PyMove: A Python library for processing and visualization of trajectories and other spatial-temporal data*. [S.l.], 2024. Versão 3.0.0. Disponível em: <https://pymove.readthedocs.io/>. Acesso em: 2024. Citado na página 46.

QIAO, S. et al. A self-adaptive parameter selection trajectory prediction approach via hidden markov models. *IEEE Transactions on Intelligent Transportation Systems*, v. 16, n. 1, p. 284–296, 2015. Citado na página 37.

RABINER, L. R. A tutorial on hidden Markov models and selected applications in speech recognition. *Proc. IEEE*, IEEE, v. 77, n. 2, p. 257–286, 1989. Citado 2 vezes nas páginas 26 e 28.

Scikit-mobility Team. *Scikit-mobility: A Python library for the analysis, generation and risk assessment of mobility data*. [S.l.], 2024. Versão 1.3.0. Disponível em: <https://scikit-mobility.github.io/scikit-mobility/>. Acesso em: 2024. Citado na página 46.

SHEN, G. et al. Spatio-Temporal Interactive Graph Convolution Network for Vehicle Trajectory Prediction. *Internet of Things 24*, 2023. Citado na página 34.

SILVA, C. L. da; PETRY, L. M.; BOGORNY, V. A survey and comparison of trajectory classification methods. In: . 2019 8th Brazilian Conference on Intelligent Systems (BRACIS), 2019. Disponível em: <https://www.researchgate.net/publication/337791304_A_Survey_and_Comparison_of_Trajectory_Classification_Methods>. Acesso em: 28 dez. 2024. Citado na página 19.

SILVA, T. H. et al. Urban computing leveraging location-based social network data: A survey. *Association for Computing Machinery*, v. 52, n. 1, p. 17–39, 2019. Citado na página 19.

SIRQUEIRA, T. F. M. et al. *Application of Statistical Methods in Software Engineering: Theory and Practice*. 2020. Disponível em: <https://arxiv.org/abs/2006.15624>. Citado na página 67.

SPACCAPIETRA, S. et al. A conceptual view on trajectories. *Data Knowledge Engineering*, v. 65, n. 1, p. 126–146, 2008. ISSN 0169-023X. Including Special Section: Privacy Aspects of Data Mining Workshop (2006) - Five invited and extended papers. Disponível em: <https://www.sciencedirect.com/science/article/pii/S0169023X07002078>. Citado na página 19.

SUN, Y. et al. Streaming trajectory segmentation based on stay-point detection. *Database Systems for Advanced Applications. DASFAA 2024. Lecture Notes in Computer Science, vol 14850. Springer, Singapore*, v. 14850, 2024. Citado 2 vezes nas páginas 29 e 46.

TANIMURA, C. *SQL Para Análise de Dados: Técnicas Avançadas Para Transformar Dados em Insights*. 1. ed. Rio de Janeiro: Novatec Editora, 2022. Acesso em: 22 jul 2022. Citado na página 19.

WEINTRIT, A. The concept of time in navigation. transnav the international journal on marine navigation and safety of sea transportation. *TransNav*, v. 11, p. 23–33, 2017. Citado na página 22.

ZENG, Q.; WANG, J.; HE, K. Improving destination prediction via ensemble of trajectory movement separation and adaptive clustering. *IEEE Access*, v. 8, p. 28142–28154, 2020. Citado na página 21.

ZHANG, W. et al. Trajectory prediction with recurrent neural networks for predictive resource allocation. In: *2018 14TH IEEE International Conference On Signal Processing (ICSP)*. [S.l.: s.n.], 2018. Citado na página 34.

ZHOU, Y. et al. An efficient destination prediction approach based on future trajectory prediction and transition matrix optimization. *IEEE Transactions on Knowledge and Data Engineering*, v. 32, n. 2, p. 203–217, 2020. Citado na página 36.

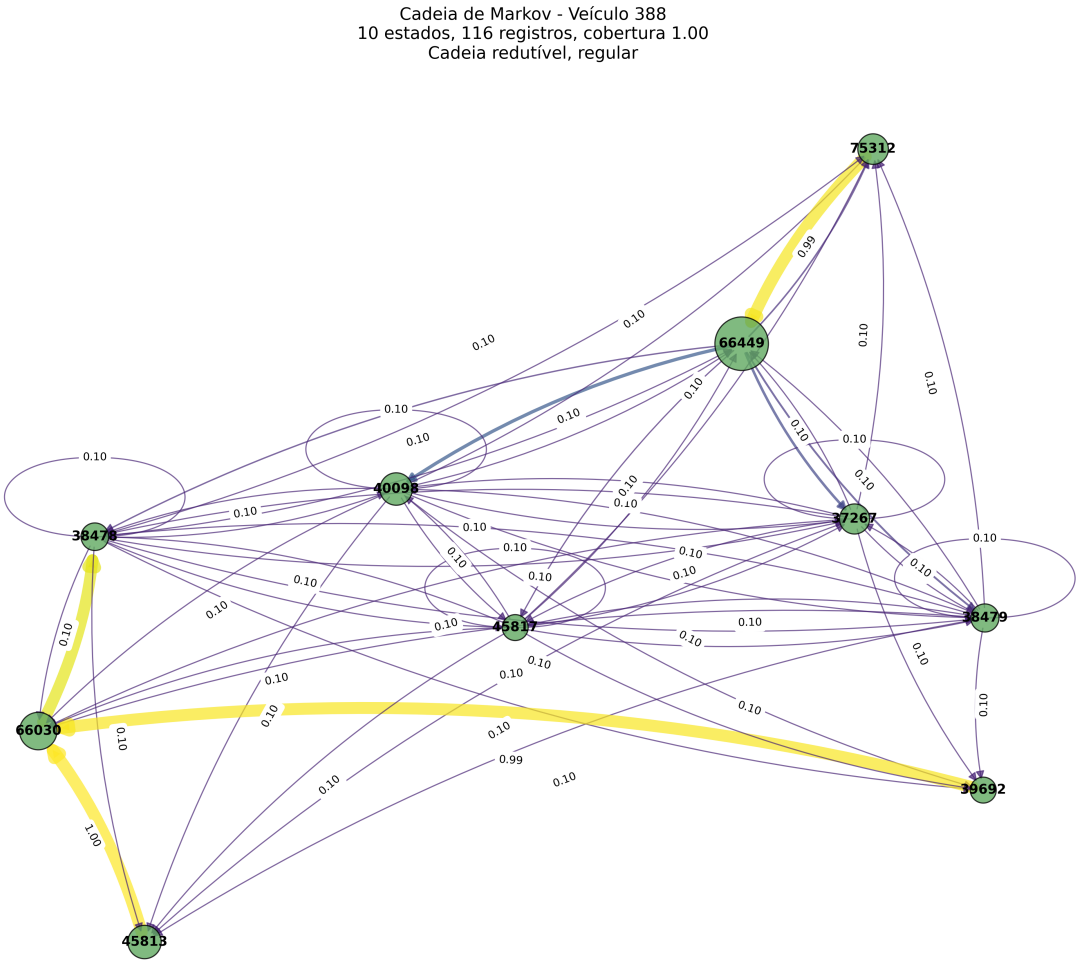
Apêndices

APÊNDICE A – GRAFOS - VEÍCULO 388 - CADEIAS DE MARKOV

Para o estudo de um dos casos, o do veículo 388, tem-se conforme na Figura 18, no qual se observam nos nós, maior proeminência para a grade de destino cujo rótulo é "66449". Nos arcos do grafo, as porcentagens, em valores decimais, das probabilidades de transição, de mudança de estado.

Observa-se que a Figura em si ilustra toda uma cadeia de relações cujos nós podem representar destinos de um usuário, enquanto os arcos (as linhas), seguem representando as trajetórias sobre os logradouros de Ann Arbor. Neste caso, é preciso lembrar que apenas fatores geográficos são considerados, sem nada contextual.

Figura 18 – Grafos com Cadeias de Markov.



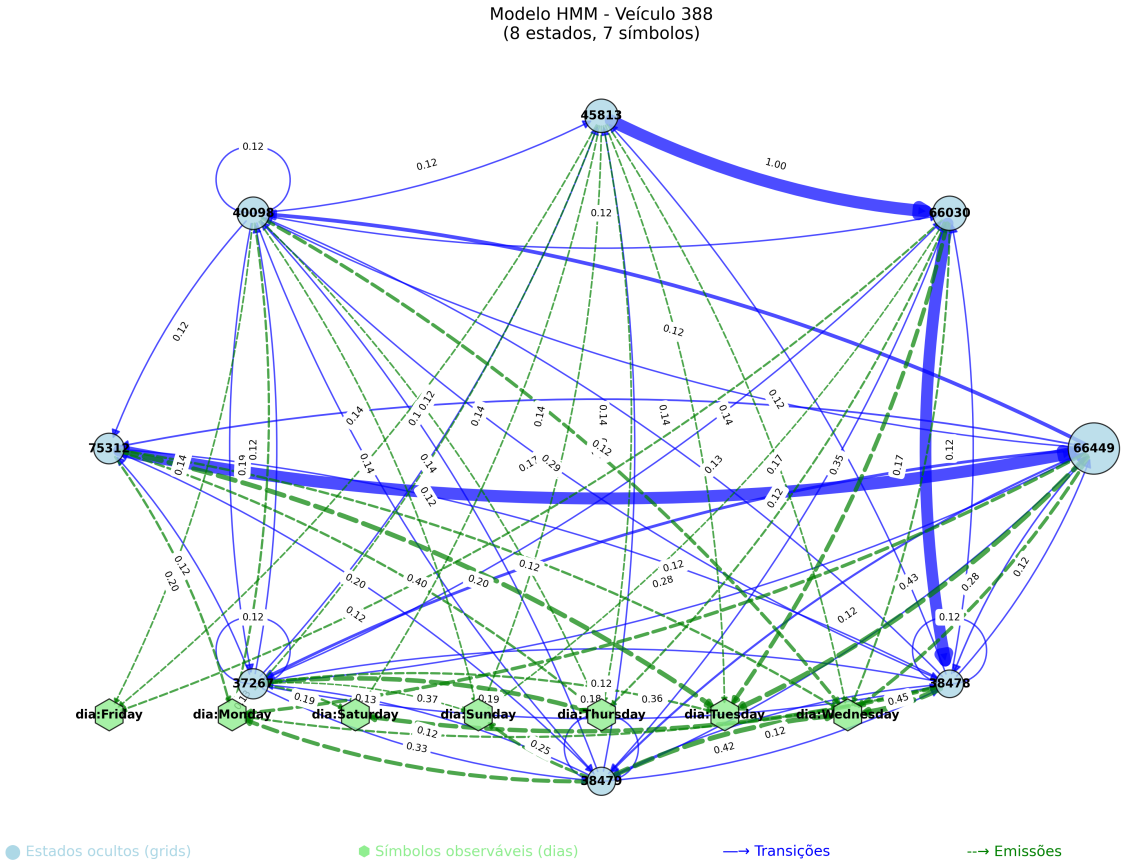
Fonte: Elaboração própria.

APÊNDICE B – GRAFOS - VEÍCULO 388 - HMM

Para o estudo de um dos casos, o do veículo 388, tem-se conforme na Figura 19, no qual se observam nos nós, maior proeminência para a grade de destino cujo rótulo é "66449". Nos arcos do grafo, as porcentagens, em valores decimais, das probabilidades de transição, de mudança de estado, além dos arcos tracejados, com o indicativo das mudanças de estados dos símbolos (dias da semana com rótulos de grades de origem).

É preciso considerar que essa relação se compõe considerando grade de origem e dia, para treinamento, gerando uma predição de rótulos de grades de destino - apenas, e a consequente relação com os símbolos para os dados de teste (rótulos de grades de destino).

Figura 19 – Grafos para HMM.



Fonte: Elaboração própria.

ENTREGA DA VERSÃO FINAL DE DISSERTAÇÃO

Eu, PROF. DR. FRANCISCO DANTAS NOBRE NETO, autorizo o aluno(a) JOÃO BATISTA FIRMINO JUNIOR a entregar a versão final da dissertação de mestrado, à secretaria do PPGTI, que foi por mim analisada e está de acordo com os apontamentos feitos pelos membros da banca de apresentação do referido aluno.

Prof. Dr. Francisco Dantas Nobre Neto
Orientador

João Pessoa, 30 de Junho de 2025.