Representação Probabilística de Trajetórias Veiculares como Entrada para Redes Neurais Artificiais

Bianca Lahm Gomes [1], **Kame Haung Zhu** [1]

[1]Centro de Tecnologias Aplicadas – Itaipu Parquetec – Foz do Iguaçu, PR – Brasil

bianca.gomes@itaipuparquetec.org.br, kame.zhu@itaipuparquetec.org.br

Abstract. *This work proposes a probabilistic approach to represent vehicular trajectories using data from license plate recognition (LPR) cameras, aiming at real-time processing. It is based on the Hierarchical Pattern Bayes (HPB) model to generate matrices of trajectory typicality and temporal density. These matrices were evaluated using the Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN) algorithm, resulting in a 75% increase in the Silhouette score and a 31.1% reduction in the Davies-Bouldin index compared to the use of raw coordinates. These results indicate a more structured representation, enabling the application of supervised models for real-time anomaly detection.*

Resumo. *Este trabalho propõe uma abordagem probabilística para representar trajetórias veiculares a partir de dados de câmeras de reconhecimento de placas (LPR), visando um processamento em tempo real. Baseia-se no modelo Hierarchical Pattern Bayes (HPB) para gerar matrizes de tipicidade e densidade temporal dos trajetos. Essas matrizes foram avaliadas com o algoritmo Hierarchical Density-Based Spatial Clustering of Applications with Noise (HDBSCAN), resultando em um aumento de 75% no índice de Silhouette e uma redução de 31,1% no índice de Davies-Bouldin, em comparação ao uso de coordenadas brutas. Isso indica uma representação mais estruturada e viabiliza a aplicação em modelos supervisionados para detecção de anomalias em tempo real.*

1. Introdução

Câmeras LPRs têm ampliado a geração de dados de tráfego em tempo real, viabilizando aplicações em segurança pública e mobilidade urbana [Sun et al. 2020, Mao et al. 2024]. No entanto, muitas abordagens ainda representam trajetórias apenas por dados espaciais e temporais, sem considerar aspectos probabilísticos do comportamento veicular. Modelos com *autoencoders* ou *Long Short-Term Memory* (LSTMs) comprimem trajetórias, mas carecem de interpretabilidade [Peralta et al. 2023]. Métodos estatísticos, como Mahalanobis e *Isolation Forest*, não escalam bem e assumem distribuições restritivas [Liu et al. 2008]. Redes Bayesianas Dinâmicas (DBNs) modelam dependências temporais, mas com alto custo computacional [Ghahramani 1998]. Já análises de trajetos de forma agregada por região, limitam a granularidade e uso em tempo real [Cruz et al. 2018]. Dessa forma, este trabalho propõe uma representação probabilística individual baseada no modelo HPB, adaptado para estimar a tipicidade de trajetos e densidade temporal por meio de uma estrutura hierárquica, gerando vetores informativos para inferência. A avaliação, com dados reais anonimizados de LPRs fornecidos por órgãos públicos, compararam a representação com e sem o modelo proposto, visando reduzir falsos positivos e capturar melhor as variações reais nos comportamentos veiculares.

2. Trabalhos Relacionados

Para aplicar modelos de detecção de anomalias em trajetórias veiculares, os *autoencoders* e variantes baseadas em LSTMs têm sido amplamente explorados para aprender representações comprimidas e detectar desvios por meio da reconstrução dos dados [Peralta et al. 2023]. Abordagens baseadas em estatística clássica, como a utilização da distância de Mahalanobis, e métodos, como o *Isolation Forest*, possuem limitações decorrentes da suposição de normalidade dos dados e apresentam escalabilidade reduzida em cenários de alta dimensionalidade [Liu et al. 2008]. Modelos probabilísticos como as DBNs oferecem mecanismos para modelar dependências temporais, porém apresentam alto custo computacional em cenários com muitos estados e atributos [Ghahramani 1998]. Em [Cruz et al. 2018] foi proposto uma metodologia para detecção e classificação de anomalias em dados de mobilidade urbana, agregados espacial e temporalmente, com base em séries espaço-temporais derivadas de trajetórias de ônibus no Rio de Janeiro. A abordagem se baseia na identificação de anomalias em regiões predefinidas e na extração de padrões frequentes utilizando o algoritmo a priori, com foco em eventos sistêmicos e comportamento agregado.

Em contraste com abordagens anteriores, este trabalho propõe uma representação probabilística de trajetórias veiculares em nível individual, incorporando tipicidade e densidade temporal do trajeto por meio de uma estrutura hierárquica. Fundamentado no modelo HPB, o modelo lida com dados esparsos, suaviza estimativas e proporciona uma modelagem mais precisa e sensível ao comportamento veicular [Filho and Wainer 2008]. A abordagem supera limitações de escalabilidade, granularidade e interpretabilidade, viabilizando a detecção não supervisionada de anomalias em tempo real, o que é um aspecto essencial para aplicações em segurança pública e mobilidade urbana.

3. Metodologia

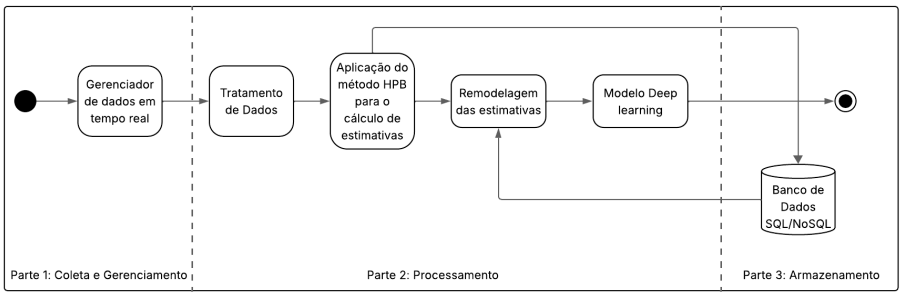

Figura 1. *Pipeline* de inferência representando o fluxo de dados da modelagem de trajetórias veiculares em tempo real.

O *pipeline* ilustrado na Figura 1 é estruturado em três blocos:

- **Coleta e Gerenciamento:** As passagens de veículos extraídas por câmeras LPR são transmitidas em tempo real via *Apache Kafka* e pré-processadas com *Apache Spark*, garantindo escalabilidade e eficiência na transmissão entre os módulos apresentados no *pipeline*;
- **Processamento:** As estimativas $P(TJ_i)$ e $P(H_{(i+1 \to i)})$, obtidas com o modelo HPB, mapeiam o comportamento veicular. Os dados são remodelados em vetores de quatro dimensões e processados por um modelo de rede neural.

- **Armazenamento:** Os módulos interagem com bancos SQL e NoSQL para armazenar estimativas, inferências e frequências empíricas. Adotou-se o *Apache Cassandra* pela escalabilidade e desempenho em consultas.

3.1. O cálculo de estimativas das passagens $P(TJ_i)$ e $P(H_{(i+1\to i)})$

O método HPB utiliza uma abordagem hierárquica para suavizar estimativas de probabilidade condicional, com desempenho robusto mesmo em contextos com dados esparsos. Um trajeto é definido como uma sequência ordenada de passagens por pontos de câmera, sendo o último ponto registrado aquele que desencadeia a inferência. Para refletir essa estrutura, as dependências entre os pontos são modeladas na ordem inversa da sequência, ou seja, do ponto mais recente n até o primeiro. Essa escolha se justifica pelo fato de que, na prática, o interesse recai sobre a detecção de anomalias no último ponto observado — o mais atual e informativo sobre a movimentação do veículo. Assim, a tipicidade é recalculada continuamente com base nesse ponto final, que representa o instante mais recente em que o veículo foi detectado. Denota-se a probabilidade geral do trajeto como $P(TJ)$, enquanto cada ponto da sequência é representado por TJ_i. A fórmula da probabilidade total do trajeto é dada por:

$$P(TJ) = \prod_{i=n}^{0} P(TJ_i \mid TJ_{(n\to i+1)}) \cdot P(H_{(i+1\to i)}) \tag{1}$$

onde $H_{(i+1\to i)}$ representa o tempo decorrido entre os pontos TJ_{i+1} e TJ_i, incorporando uma componente temporal ao modelo. O termo $P(TJ_i \mid TJ_{n\to i+1})$ simboliza a probabilidade a priori calculada recursivamente e seguindo o produtório $i = n$ até 0.

3.1.1. O cálculo das estimativas dos pontos no trajeto $P(TJ_i)$

Baseado na hierarquia proposta pelo HPB, adotou-se uma estimativa recursiva da probabilidade condicional de cada ponto do trajeto, com suavização por um prior baseada em níveis superiores da hierarquia contextual. A estimativa para $P(TJ_i \mid TJ_{(n-j\to i+1)})$ é definida por:

$$P(TJ_i \mid TJ_{(n-j\to i+1)}) = \frac{NTJ_{(n-j\to i)} + S \cdot P(TJ_i \mid TJ_{(n-j-1\to i+1)})}{NTJ_{(n-j\to i+1)} + S} \tag{2}$$

A probabilidade de um ponto TJ_i, dado um histórico $TJ_{(n-j\to i+1)}$, é estimada recursivamente a partir da probabilidade associada a um histórico posterior mais curto $TJ_{(n-j-1\to i+1)}$, combinada com duas frequências observadas: $NTJ_{(n-j\to i)}$, que representa a contagem de veículos que percorreram o trecho de $n-j$ até i, e $NTJ_{(n-j\to i+1)}$, referente ao trecho de $n-j$ até $i+1$. Essas informações são fornecidas por um módulo de monitoramento contínuo das passagens, responsável por contabilizar o fluxo de veículos entre trechos e subtrechos monitorados. S é o parâmetro de suavização do modelo HPB. A recursão prossegue até que o histórico atinja seu comprimento mínimo, isto é, quando $n - j$ é igual a i, conhecido como nível base. Quando a recursão atinge o nível base, utiliza-se a probabilidade marginal:

$$P(TJ_i) = \frac{NT_i + \frac{S}{|TJ|}}{NT + S} \tag{3}$$

em que $|TJ|$ representa o número total de pontos de monitoramento, NT_i é a frequência total observada nos pontos de passagens TJ_i, e NT é a frequência total de passagens por quaisquer pontos.

3.1.2. O cálculo das estimativas sobre a densidade temporal $P(H_{(i+1\to i)})$

Propôs-se uma estimativa da densidade de probabilidade do tempo entre pontos, considerando a direção temporal (*sinal*) e estimativas empíricas e paramétricas. A variável *sinal* é introduzida para tratar inconsistências temporais que ocorrem quando, por eventuais atrasos nos sistemas de transmissão ou ingestão de dados, uma passagem mais recente de um veículo é processada antes de uma passagem mais antiga. Como os pontos do trajeto são processados em tempo real, essa inversão pode resultar em intervalos de tempo negativos entre passagens consecutivas, indicando uma violação da ordem cronológica esperada. A probabilidade de um tempo H em um trecho de dois pontos é dada por:

$$P(H_{(i+1\to i)}) = P(sinal_H) \cdot P(H \mid sinal_H) \tag{4}$$

Na Equação 5, a probabilidade $P(sinal_H)$ é suavizada via Bayes, onde $N_{sinal,trecho}$ é o número de vezes que o sinal H ocorreu em um trecho específico, N_{trecho} é o total de observações no trecho, e S é um parâmetro de suavização. Já $P_{todos}(sinal_H)$, definido pela Equação 6, representa a frequência relativa global do sinal, com o N_{sinal} indicando o número total de ocorrências do sinal no conjunto completo de dados e N o total de observações globais.

$$P(sinal_H) = \frac{N_{sinal,trecho} + S \cdot P_{todos}(sinal_H)}{N_{trecho} + S} \tag{5}$$

$$P_{todos}(sinal_H) = \frac{N_{sinal} + \frac{S}{2}}{N + S} \tag{6}$$

Na Equação 7, a probabilidade condicional $P(H \mid sinal_H)$ combina modelos não paramétricos e paramétricos, sendo uma média ponderada entre $P(H \mid KDE)$ (obtida via *Kernel Density Estimation*) e uma estimativa de fundo $P_{back,trecho}$. O peso dado a cada componente depende de $\sqrt{N_{trecho}}$, onde N_{trecho} é o número de observações no trecho, e S é o parâmetro de suavização.

$$P(H \mid sinal_H) = \frac{\sqrt{N_{trecho}} \cdot P(H \mid \text{KDE}) + S \cdot P_{back,trecho}}{\sqrt{N_{trecho}} + S} \tag{7}$$

Na Equação 8, a estimativa $P_{back,trecho}$ é suavizada entre uma regressão logística local $P(H \mid logistic)$ e uma genérica $P_{back} = P_{erro}(H \mid reglin)$ baseada em regressão linear. N_{trecho} representa o número de observações no trecho e S o parâmetro de suavização que controla a influência relativa entre as estimativas específicas e genéricas.

$$P_{back,trecho} = \frac{N_{trecho} \cdot P(H \mid \text{logistic}) + S \cdot P_{back}}{N_{trecho} + S}, \quad P_{back} = P_{erro}(H \mid \textit{reglin}) \quad (8)$$

3.2. Remodelagem das estimativas

Após o cálculo das estimativas em tempo real, cada trajeto de veículo é submetido à fase de remodelagem, que prepara os dados para entrada em um modelo de *deep learning*, com o objetivo de inferir a probabilidade de associação do veículo a comportamentos ilícitos.

Como os trajetos variam em extensão, os dados são padronizados em matrizes de dimensão fixa e manipulados em larga escala com PySpark. Adotou-se uma matriz de 20×10 com 4 dimensões, correspondentes aos atributos (TJ_i, $H_{i \to i+1}$, latitude, longitude). Para representar o trajeto completo, este é segmentado em 10 níveis hierárquicos (0 a 9). O nível 0 armazena os 20 primeiros pontos brutos, enquanto os níveis seguintes agrupam os dados de forma exponencial: no nível 1, cada uma das 20 posições da matriz representa a média aritmética de 2 pontos consecutivos (2^1); no nível 2, de 4 pontos (2^2); e assim por diante, até o nível 9, em que cada posição representa um agrupamento de 512 pontos. Isso resulta em uma representação de até 20.460 pontos do trajeto. Conforme o veículo avança e novos pontos são registrados, a matriz é reestruturada em tempo real, permitindo atualização contínua da representação.

3.3. Modelo de *Deep Learning* Não Supervisionado

O *pipeline* foi avaliado com dados sintéticos em um cenário exclusivamente para testes de carga. Nessa etapa, o sistema demonstrou capacidade de processar cerca de 4 milhões de registros por dia (20 mil registros de passagem por *batch*, com 200 *batches* por hora) em um servidor dedicado. Para validar a representação probabilística proposta, aplicou-se aprendizado não supervisionado por meio de *autoencoders*, seguido de um algoritmo de agrupamento chamado HDBSCAN, este método permite representações compactas, robustas a ruído e sem a necessidade de definir o número de *clusters* previamente [Campello et al. 2013]. Característica importante para trajetos urbanos com grande variabilidade comportamental.

Os dados reais anonimizados utilizados incluíam aproximadamente 82 mil veículos e mais de 300 mil passagens. Os testes consideraram dois conjuntos de entrada: um *baseline* com latitude, longitude e *timestamps*, e outro com vetores contendo coordenadas e estimativas de tipicidade $P(TJ)$ e densidade temporal $P(H)$, obtidas a partir do método HPB. Todos os dados foram normalizados com *Standard Scaler*. O *autoencoder* denso utilizado possui codificador com camadas de 64, 32 e 16 neurônios (ReLU, *batch normalization*, *dropout* de 0.3) e decodificador com ativação sigmoide. O treinamento foi feito com otimizador Adam (lr=0,001), MSE como função de perda, por 50 épocas e um *batch size* de 32. Os vetores codificados (dimensão 16) foram reduzidos com *Uniform Manifold Approximation and Projection* (UMAP), utilizando inicialização espectral *init=spectral* e controle da aleatoriedade por meio de *random state*=42. Os *embeddings* gerados foram agrupados com HDBSCAN, configurado com no mínimo 500 amostras por *cluster*.

4. Resultados

A Figura 2 apresenta gráficos de dispersão dos agrupamentos com e sem as estimativas $P(TJ)$ e $P(H)$. Observa-se que o uso das estimativas resulta em *clusters* mais bem

definidos e com menor ruído, enquanto sua ausência acarreta sobreposição entre grupos e maior proporção de *outliers*.

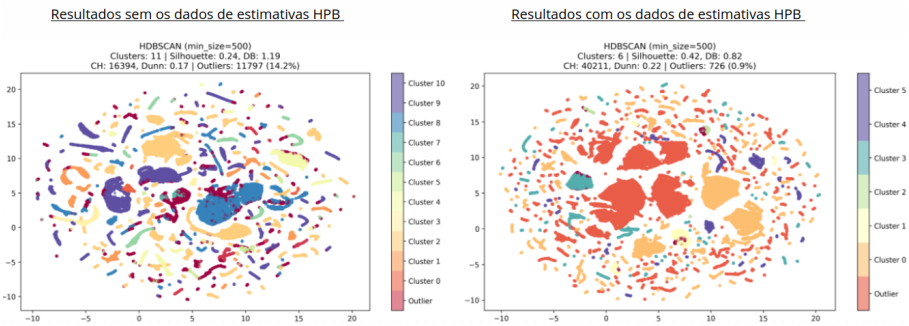

Figura 2. Gráficos de dispersão resultantes com e sem o método HPB, após aplicar o modelo não supervisionado HDBSCAN

A avaliação quantitativa da segmentação foi conduzida com base em quatro métricas clássicas de agrupamento. O índice de Silhouette avalia a coesão *intra-cluster* e a separação entre *clusters*, variando de –1 a 1, sendo desejáveis valores mais próximos de 1. A métrica Davies-Bouldin (DB) estima a sobreposição entre *clusters*, sendo melhor quanto menor o valor. O índice de Calinski-Harabasz (CH) corresponde à razão entre a variância entre e dentro dos grupos. E o índice de Dunn que mede a separabilidade relativa entre *clusters*, sendo mais elevado quando há maior distinção entre os grupos.

Tabela 1. Comparação das métricas de agrupamento com e sem uso do HPB

Métrica	Sem HPB	Com HPB
Silhouette	0.24	0.42 (+75%)
Davies-Bouldin	1.19	0.82 (–31,1%)
Calinski-Harabasz	16394	40211 (+145,2%)
Índice de Dunn	0.17	0.22 (+29,4%)
Outliers detectados	14,2%	0,9%

A Tabela 1 evidencia que a aplicação do HPB promoveu melhorias expressivas na qualidade das representações, com aumento de até 145% no índice de Calinski-Harabasz (CH), redução de 31,1% no índice de Davies-Bouldin (DB) e de 13,3 pontos percentuais na proporção de *outliers*, indicando representações mais compactas e menos ambíguas.

5. Conclusão

Este trabalho propôs uma abordagem de representação probabilística hierárquica para trajetórias veiculares, demonstrando que a estratégia adotada contribui para a melhoria na representação dos dados. Os principais diferenciais são a adaptabilidade a dados esparsos e a interpretabilidade dos vetores. Como limitação, destaca-se a ausência de rótulos confiáveis. Futuramente, pretende-se aplicar classificadores supervisionados e estender a metodologia a novos contextos urbanos, com o objetivo de confirmar automaticamente anomalias e melhorar a robustez da análise, incorporando um modelo supervisionado ao *pipeline* em tempo real.

Referências

Campello, R. J. G. B., Moulavi, D., and Sander, J. (2013). Density-based clustering based on hierarchical density estimates.

Cruz, A. B., Ferreira, J., Carvalho, D., Mendes, E., Pacitti, E., Coutinho, R., Porto, F., and Ogasawara, E. (2018). Detecção de anomalias frequentes no transporte rodoviário urbano. In *Anais do XXXIII Simpósio Brasileiro de Banco de Dados*, pages 271–276, Porto Alegre, RS, Brasil. SBC.

Filho, J. J. and Wainer, J. (2008). Hpb: A model for handling bn nodes with high cardinality parents. *Journal of Machine Learning Research*, 9(70):2141–2170.

Ghahramani, Z. (1998). *Learning dynamic Bayesian networks*, pages 168–197. Springer Berlin Heidelberg, Berlin, Heidelberg.

Liu, F. T., Ting, K. M., and Zhou, Z.-H. (2008). Isolation forest. In *2008 Eighth IEEE International Conference on Data Mining*, pages 413–422.

Mao, Y., Shi, Y., and Lu, B. (2024). Detecting urban traffic anomalies using traffic-monitoring data. *ISPRS International Journal of Geo-Information*, 13(10).

Peralta, B., Soria, R., Nicolis, O., Ruggeri, F., Caro, L., and Bronfman, A. (2023). Outlier vehicle trajectory detection using deep autoencoders in santiago, chile. *Sensors*, 23(3).

Sun, L., Chen, X., He, Z., and Miranda-Moreno, L. (2020). Routine pattern discovery and anomaly detection in individual travel behavior. *CoRR*, abs/2004.03481.