

Previsão de localização e tempo de chegada para rotas de ônibus

Álvaro de Carvalho Alves¹

¹Programa de Engenharia de Sistemas e Computação
COPPE - Universidade Federal do Rio de Janeiro

22/06/2025

1 Carga de dados

O processo consiste em um pipeline automatizado para extrair, transformar e carregar dados de GPS de ônibus, a partir de arquivos JSON para um banco de dados otimizado. O fluxo é executado pelo script '**fase1_carga_dados.py**' em quatro etapas principais:

1. **Extração e Filtragem Inicial:** O sistema primeiro localiza todos os arquivos de dados em um diretório especificado. De imediato, realiza uma pré-filtragem, selecionando apenas os arquivos cujos nomes correspondem ao horário de operação dos ônibus. Isso otimiza o processo, evitando a leitura de dados irrelevantes.
2. **Transformação e Limpeza:** Os arquivos selecionados são processados em paralelo para maior eficiência. Cada registro de GPS dentro dos arquivos passa por um processo de validação e limpeza para garantir a qualidade dos dados:
 - A data e hora são padronizadas para um fuso horário único.
 - Registros fora do horário de operação são descartados.
 - As coordenadas geográficas são validadas para assegurar que pertencem à área de interesse (a região metropolitana do Rio de Janeiro).
 - Velocidades consideradas irreais ou impossíveis são removidas.
 - Qualquer registro com informações essenciais faltando é descartado.
3. **Preparação e Carga no Banco de Dados:** Os dados já limpos e validados são carregados em grandes lotes para o banco de dados,

um método muito mais rápido do que inserir registro por registro. O banco de dados é previamente estruturado para alta performance, com os dados sendo automaticamente organizados em partições diárias e utilizando índices geoespaciais, o que acelera significativamente futuras consultas geográficas e por data.

4. **Garantia de Integridade:** Durante a carga, o sistema automaticamente converte as coordenadas de latitude e longitude em um formato de ponto geográfico. Além disso, o processo é desenhado para não inserir dados duplicados, mesmo que seja executado múltiplas vezes sobre os mesmos arquivos, garantindo a consistência e a integridade da base de dados final.

2 Análise exploratória

Este processo realiza uma análise exploratória focada nos dados de GPS de uma linha de ônibus específica em um dia determinado, com o objetivo de extrair informações sobre as rotas realizadas e os pontos que fogem do padrão. O fluxo é executado pelo script '*fase2_analise_exploratoria.py*' e dividido em três fases sequenciais:

1. **Extração e Preparação dos Dados:** A análise começa com a conexão ao banco de dados, onde os dados de GPS já foram previamente limpos e armazenados. O processo não utiliza a base de dados inteira; em vez disso, ele extrai um subconjunto de dados altamente específico: todos os registros de uma única linha de ônibus ('LINHA_ALVO') durante um único dia ('DIA_ALVO'). Essa abordagem focada permite uma análise detalhada e comparável do desempenho da linha em um dia de operação.
2. **Engenharia de Atributos:** Com os dados brutos em mãos (latitude, longitude, tempo), o processo enriquece o conjunto de dados calculando novas métricas (atributos) que são mais informativas para a análise. Para cada veículo, os dados são ordenados cronologicamente para reconstruir sua trajetória. A partir daí, são calculados:
 - O intervalo de tempo entre cada ponto de GPS consecutivo.
 - A distância percorrida entre esses mesmos pontos.
 - A velocidade do veículo, derivada da relação entre a distância percorrida e o tempo gasto.Esses novos atributos são fundamentais para entender a dinâmica do deslocamento dos ônibus.
3. **Visualização de Dados:** Utilizando os dados enriquecidos, o processo gera uma série de visualizações e análises para explorar diferentes fa-

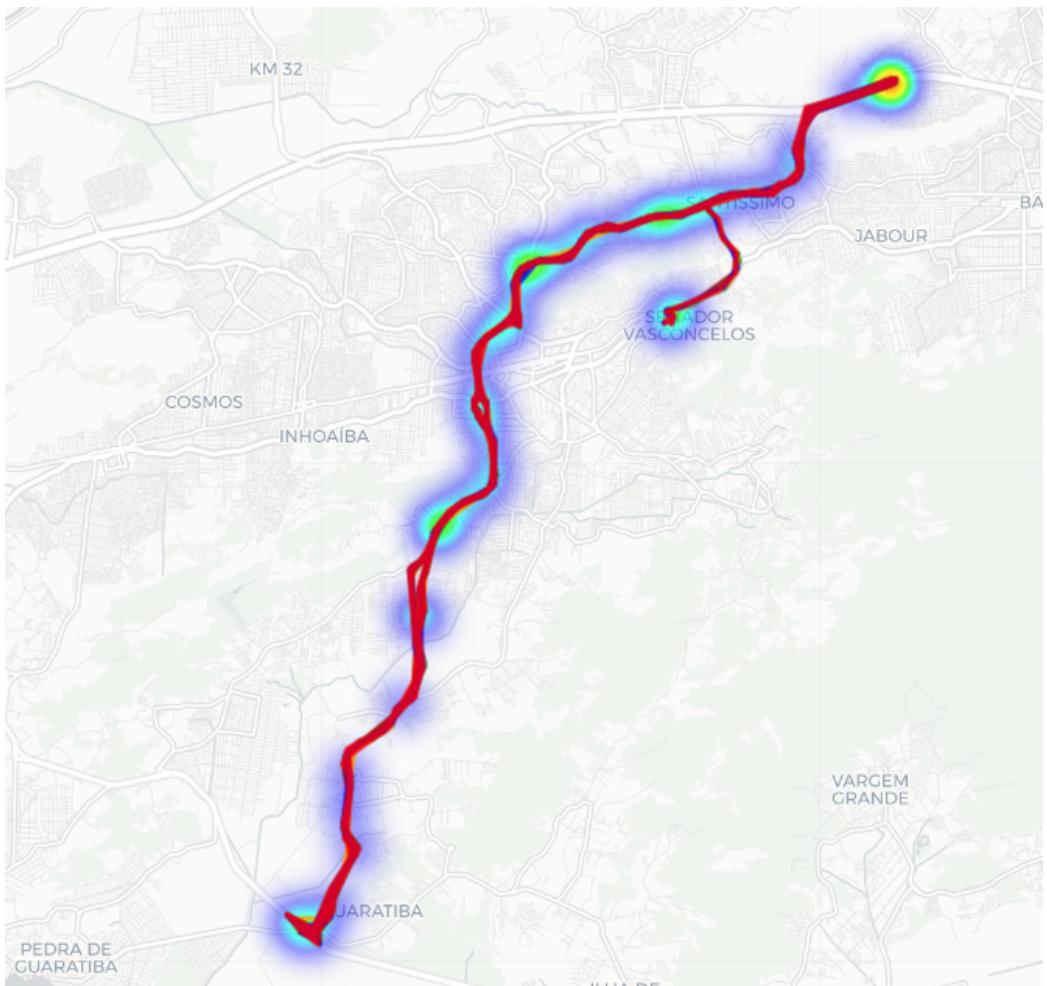


Figura 1: Mapa de calor dos pontos do 853.

cetas da operação da linha:

- **Análise Geográfica de Trajetórias e Congestionamento:** É gerado um mapa interativo que plota a rota de cada veículo da linha. Sobreposto a essas rotas, um "mapa de calor"(heatmap) destaca visualmente as áreas geográficas onde os ônibus se movem com baixa velocidade, apontando para possíveis pontos de congestionamento.
- **Análise Estatística das Métricas:** São criados gráficos (histogramas e boxplots) para visualizar a distribuição das velocidades, a frequência dos sinais de GPS (tempo entre registros) e a latência dos dados. Isso oferece um panorama quantitativo do comportamento geral da linha e da qualidade do sinal.

Análise Estatística Básica da Linha 853

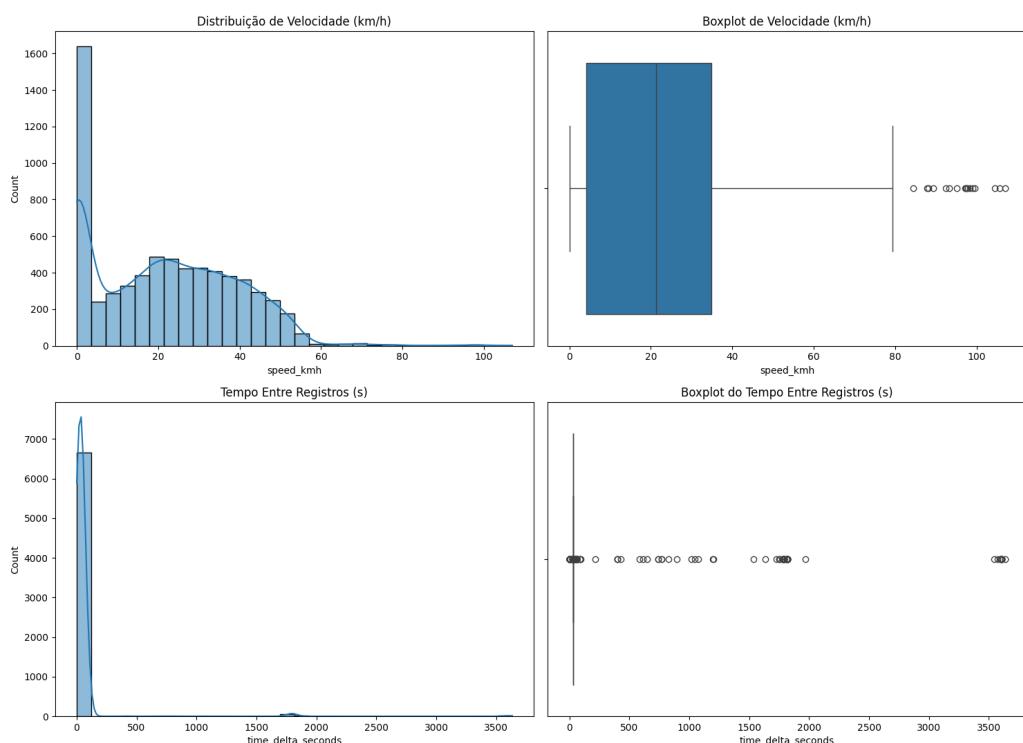


Figura 2: A visualização velocidade mostra a predominância de muitos pontos nos quais a velocidade é zero. A grande maioria dos dados possuem velocidade menor que 60 Km/h.

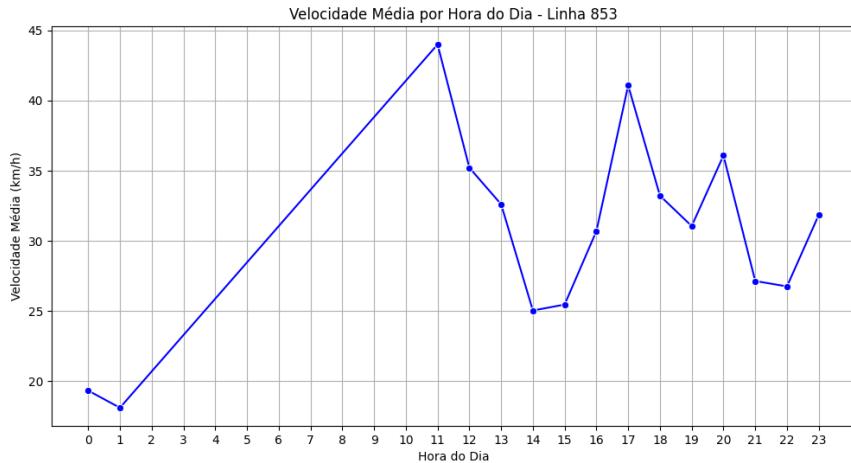


Figura 3: Enter Caption

- Análise Temporal do Desempenho: A performance da linha é analisada ao longo do dia. Um gráfico mostra como a velocidade média dos ônibus varia hora a hora, permitindo identificar facilmente os períodos de pico e de menor movimento.
- Análise Comparativa entre Veículos: O desempenho de cada veículo individual da mesma linha é comparado. Essa análise permite verificar a consistência da operação e identificar se algum veículo específico apresenta um comportamento atípico (por exemplo, sendo consistentemente mais lento que os demais).

O resultado final do processo é um conjunto de arquivos (um mapa interativo e várias imagens de gráficos) que, juntos, fornecem algumas informações da linha de ônibus analisada.

3 Detecção dos pontos finais

O método caracteriza-se pelo tratamento de outliers e pela geração de artefatos que permitem a rastreabilidade e verificação dos resultados intermediários. A arquitetura do processo foi projetada para execução paralela, visando a escalabilidade para a análise de um grande volume de linhas. A parte principal é executada pelo script *fase3_identificacao_terminais.py* e consiste das seguintes etapas.

1. **Discretização do Espaço Operacional e Mapeamento da Grade de Atividade:** A etapa inicial consiste na criação de uma gre-

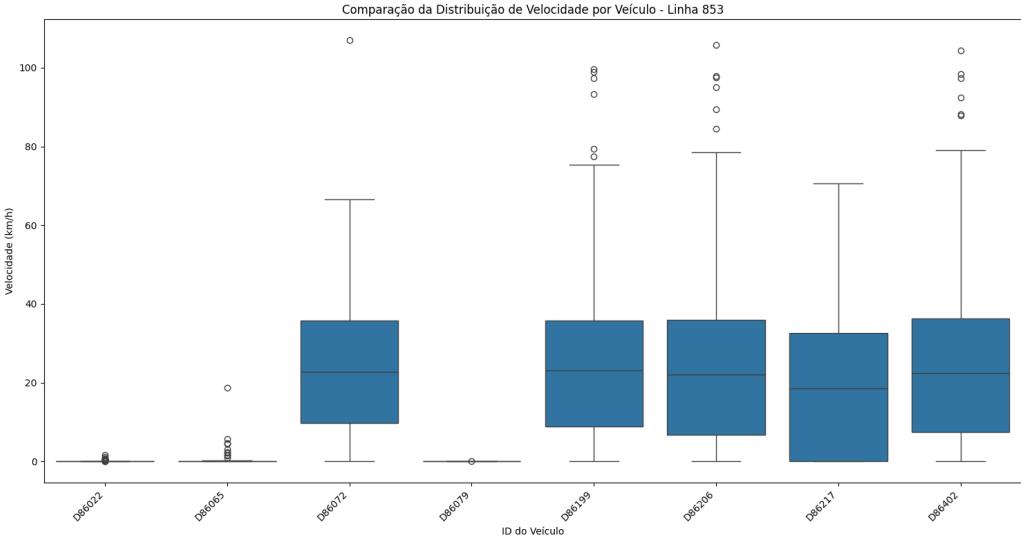


Figura 4: A figura mostra um comportamento atípico da ordem D86217. Também mostra que algumas ordens devem ter passado o dia na garagem.

lha (*grid*) de células com resolução predefinida. Subsequentemente, realiza-se o mapeamento de todas as células da grelha que apresentam atividade para uma determinada linha, com base em um agregado de dados temporais de duas semanas. Este procedimento resulta em uma representação completa do espaço de operação, que inclui tanto o trajeto canônico da rota quanto desvios e localizações operacionais secundárias, como garagens.

2. **Definição do Corredor Operacional por Análise de Conectividade:** Nesta etapa, emprega-se o algoritmo de clusterização baseado em densidade *DBSCAN* sobre o conjunto de células ativas da grelha. O objetivo é a identificação do maior componente espacialmente conectado de células, o qual é definido como o corredor operacional principal. Componentes desconexos, que representam ilhas de atividade geograficamente isoladas do corredor principal, são classificados como *outliers* e excluídos das fases posteriores. Algumas linhas separadas por túneis (como a 315 ou a 483) acabam tendo problemas na hora da clusterização, quebradas em duas rotas separadas. Para lidar com isso, sempre são selecionados os dois maiores clusters válidos.
3. **Identificação de Zonas de Alta Permanência por Análise Temporal:** Concomitantemente, executa-se uma análise temporal sobre os registros de veículos com velocidade nula ou próxima de zero. O objetivo é identificar "eventos de paragem longa", definidos como períodos

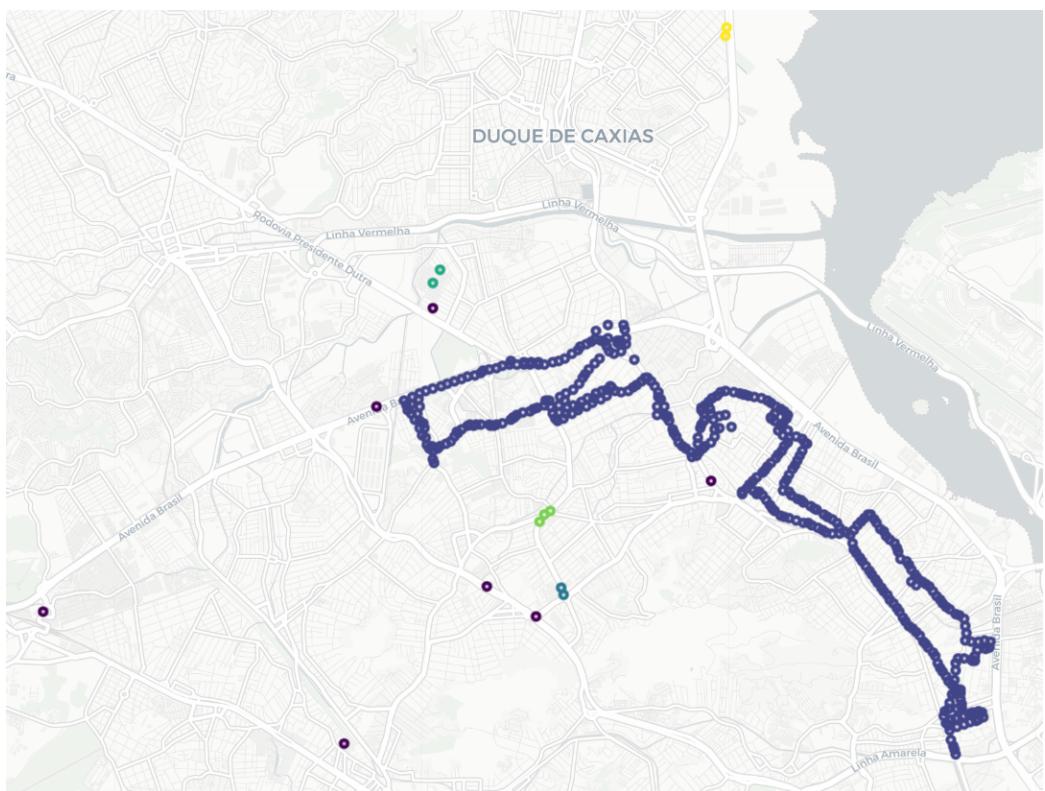


Figura 5: Resultados da clusterização da linha 905. É possível identificar alguns clusters isolados em verde e amarelo e alguns outliers em preto.

contínuos de permanência de um veículo em uma mesma célula da grelha, dentro de um intervalo de tempo parametrizado (e.g., entre 5 e 80 minutos). As células que apresentam uma alta frequência de tais eventos são catalogadas como "zonas de alta permanência", indicando um comportamento compatível com o de um ponto terminal.

4. **Interseção das Análises Espacial e Temporal para Filtragem de Candidatos:** A etapa subsequente consiste na interseção dos resultados das análises espacial e temporal. Somente as "zonas de alta permanência" que estão geograficamente contidas no "corredor operacional principal" (definido na etapa 2) são retidas. Dessa forma, zonas de alta permanência que não pertencem ao corredor, como garagens ou pátios de manutenção, são sistematicamente excluídas do conjunto de potenciais terminais.
5. **Clusterização e Seleção de Par Terminal por Heurística de Otimização:** As zonas de alta permanência filtradas são submetidas a um segundo processo de clusterização para consolidar pontos próximos em candidatos a terminal únicos. Para a seleção do par final, emprega-se uma *heuristic*a de otimização. Uma função objetivo é calculada para cada par de candidatos, ponderando a frequência de eventos de paragem em ambos e sua separação espacial. A separação é mensurada pela distância euclidiana ao quadrado, para que se dê mais peso a distância entre os pontos que a frequência. O par que maximiza a função objetivo é selecionado.
6. **Execução e Geração de Artefatos para Validação:** Para otimizar o tempo de processamento sobre múltiplas linhas, o fluxo de trabalho é implementado em um modelo de computação paralela. Ao final, o sistema persiste as coordenadas do par de terminais identificado em uma base de dados estruturada. Adicionalmente, são gerados artefatos visuais (mapas interativos) para validação. Estes mapas contêm camadas independentes para cada fase do processo, incluindo:
 - **Camada de Análise de Conectividade:** Exibe todos os *clusters* de células da grelha, permitindo a inspeção das áreas classificadas como *outliers*.
 - **Camada de Resultados:** Apresenta o corredor operacional final, todos os candidatos a terminal e o par selecionado.

Ao final da etapa, nem todas as linhas tiveram bons resultados. Entre as 48 linhas analisadas, 9 não tiveram terminais classificados. A linha 852, por exemplo, teve problema com o método proposto, pois é uma linha circular e só possui um terminal. Já a linha 100 tinha muitas rotas alternativas, provavelmente provenientes de ônibus que não trocaram de linha no sistema, mas estavam fazendo outro serviço. A linha 917 sequer teve um candidato a

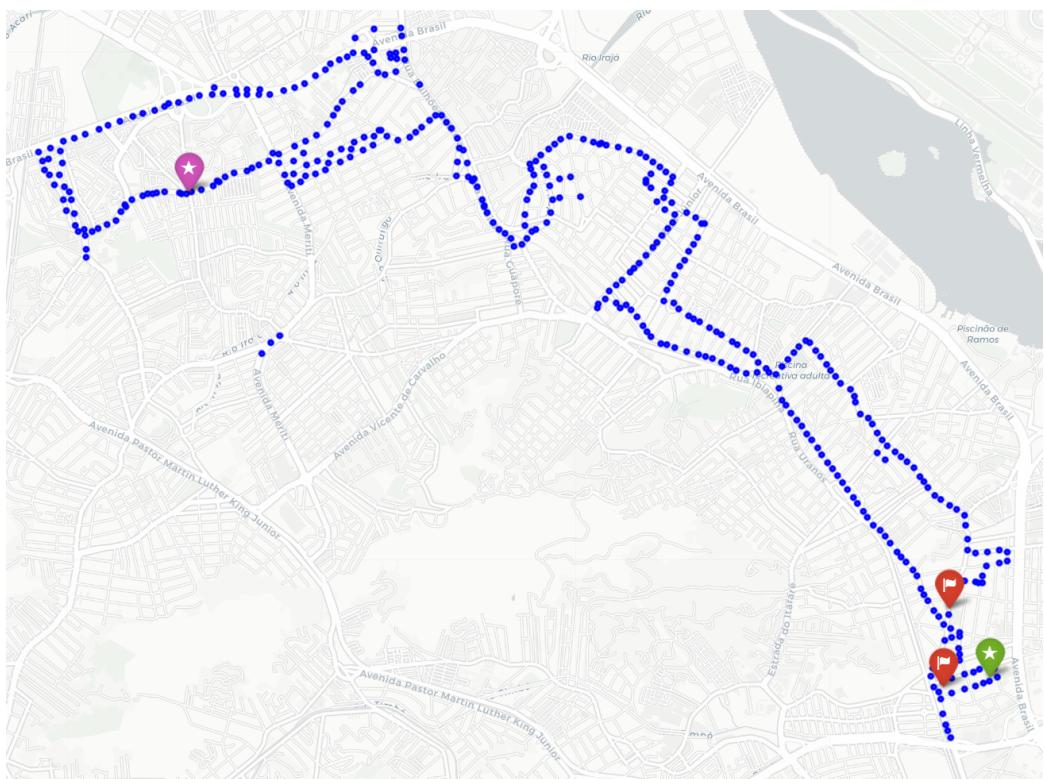


Figura 6: A figura mostra os resultados da classificação dos terminais para a linha 905, em verde e rosa. Em vermelho estão outros candidatos a terminais.

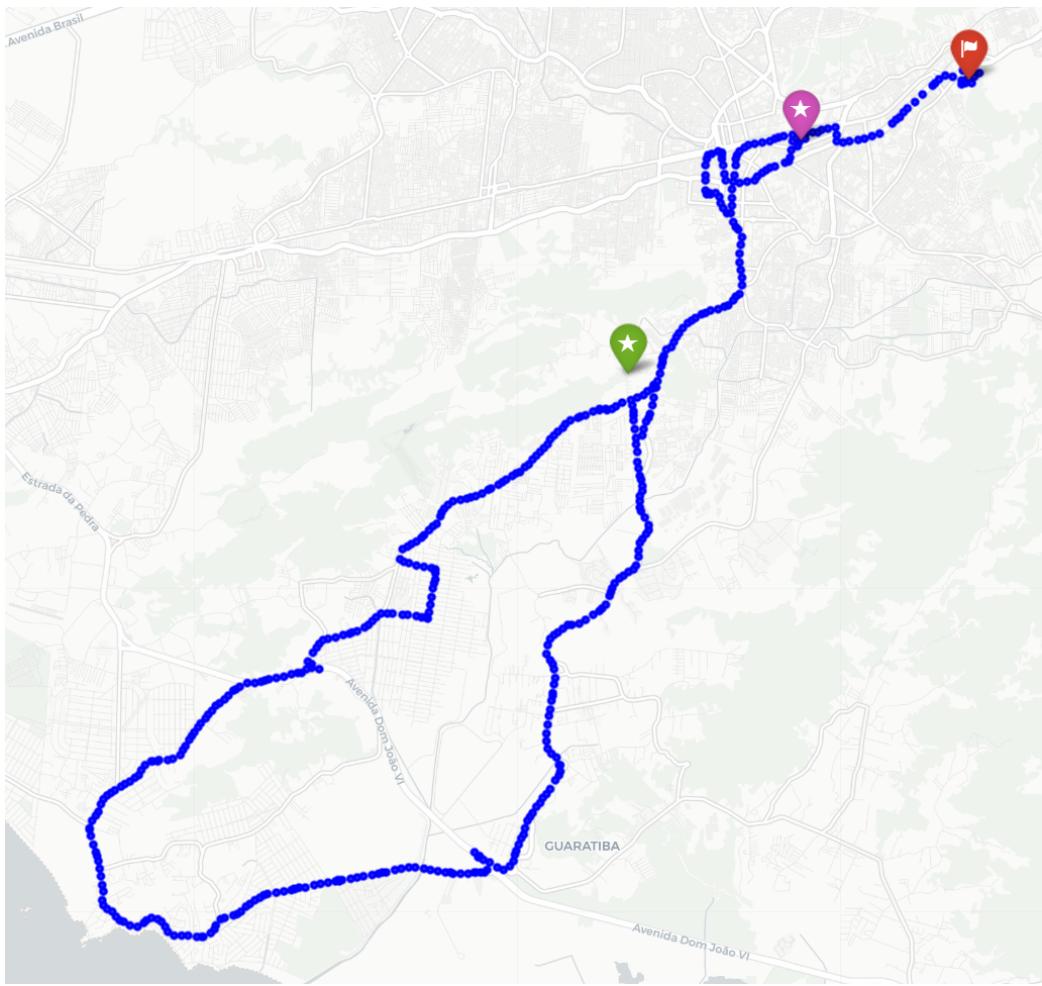


Figura 7: Resultado da classificação de terminais para a linha 852.

terminal identificado na região de Padre Miguel, seu verdadeiro ponto final. Para estes casos, os terminais foram adicionados manualmente, através do script *fase3b_insercao_manual_terminais.py*.

4 Construção dos trajetos mais comuns

O processo foi desenvolvido com foco na robustez, por meio de uma estratégia de filtragem espacial sequencial e análise comportamental, e na auditabilidade, por meio da geração de artefatos visuais para cada etapa de processamento. A implementação contempla a execução em paralelo para otimizar a análise de múltiplas linhas. Ele é orquestrado pelo script *fase4_ro-*

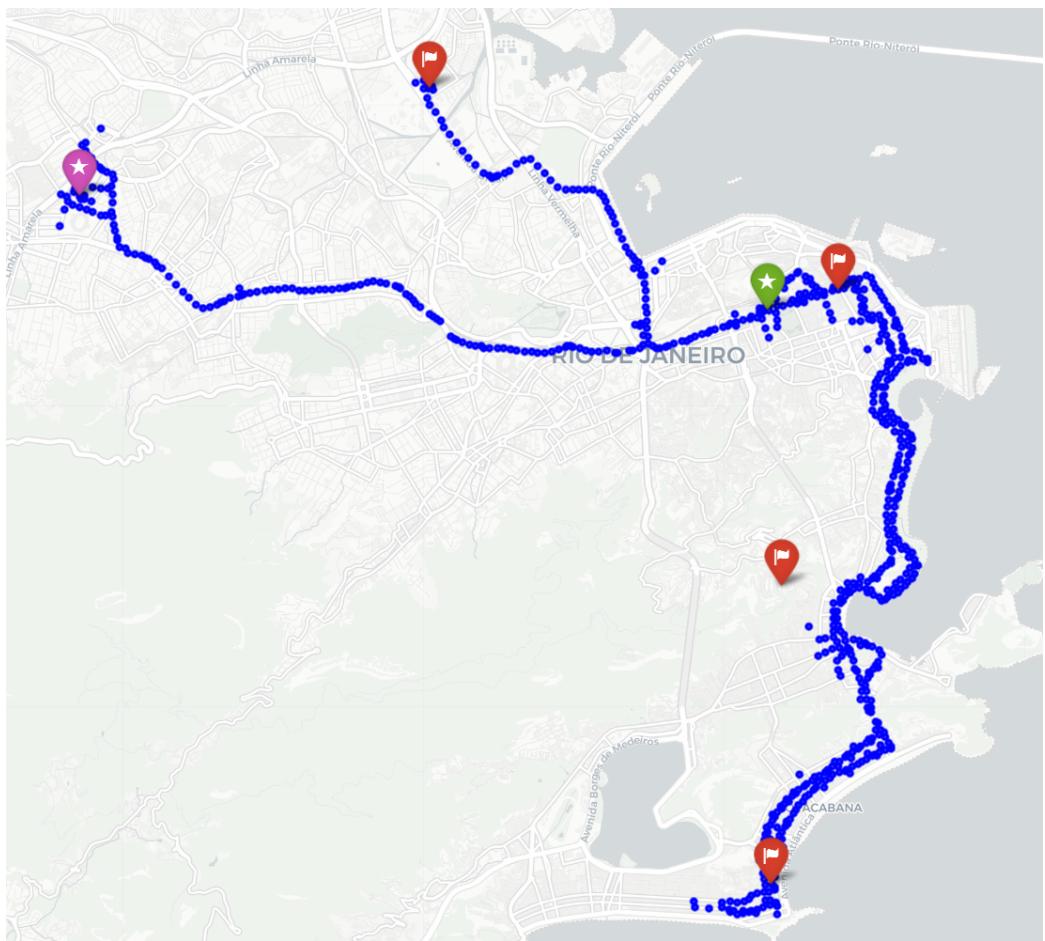


Figura 8: Resultado da classificação de terminais para a linha 100.

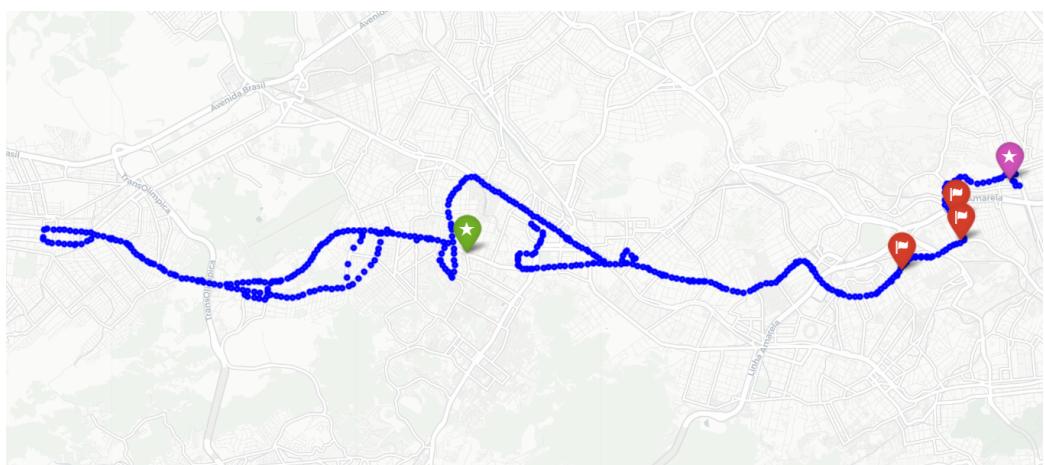


Figura 9: Resultado da classificação de terminais para a linha 917.

tas_ida_volta.py.

4.1 Definição do Corredor Operacional por Análise de Conectividade Espacial

A primeira fase da metodologia consiste na determinação precisa da área de operação de cada linha. Para tal, o espaço geográfico é inicialmente discretizado em uma grelha (*grid*) regular. As células da grelha que concentram uma densidade de registros de GPS acima de um limiar predefinido são selecionadas, formando um "esqueleto" que representa a pegada geográfica completa da linha. Sobre este esqueleto, aplica-se um algoritmo de clusterização baseado em densidade (*DBSCAN*). Esta etapa é fundamental para isolar o maior componente de células espacialmente conectadas, que é então definido como o "corredor operacional principal". Componentes menores e geograficamente isolados, tipicamente correspondentes a garagens ou desvios operacionais, são classificados como *outliers* e expurgados. A partir deste corredor validado, uma geometria poligonal contínua (*geofence*) é gerada, delimitando a área de operação canônica da linha.

4.2 Segmentação Temporal, Filtragem Espacial e Classificação de Viagens

A segunda fase visa identificar e classificar segmentos de viagem individuais. Utilizando funções de janela em consultas *SQL*, os registros de GPS de cada veículo são segmentados em "viagens" distintas, com base em interrupções temporais que excedam um limiar máximo estabelecido (e.g., 30 minutos). Subsequentemente, uma filtragem espacial rigorosa é aplicada: apenas as viagens cujos registros se encontram majoritariamente contidos dentro do *geofence* do corredor, definido na etapa anterior, são consideradas válidas para a construção da rota. Por fim, cada viagem validada é classificada em um dos dois sentidos ('ida' ou 'volta') por meio da análise da ordem temporal em que o veículo passa mais próximo dos pontos terminais previamente identificados para aquela linha.

4.3 Síntese e Refinamento da Geometria da Rota

A etapa final consiste na construção da geometria da rota canônica para cada sentido. Inicialmente, todos os registros de GPS pertencentes a viagens validadas e classificadas para um mesmo sentido são agregados. Sobre este



Figura 10: Rota clusterizada com geofence identificado para a linha 108.

conjunto de dados consolidado, o processo de discretização em grelha é re-aplicado para gerar um esqueleto denso e representativo para aquele sentido específico. A ordenação topológica dos pontos deste esqueleto é resolvida através da teoria dos grafos. Um grafo não direcionado é construído, onde os centróides das células da grelha são os nós, e as arestas conectam nós geograficamente próximos. O trajeto canônico é então determinado pelo cálculo do caminho mais curto (*shortest path*) entre os nós correspondentes aos terminais de início e fim. Por fim, a sequência de pontos resultante é submetida a um pós-processamento que inclui: i) suavização (*smoothing*), por meio de um filtro de média móvel para reduzir o ruído de alta frequência; e ii) simplificação, utilizando o algoritmo de *Douglas – Peucker* para remover vértices redundantes, preservando a forma essencial da rota.

4.4 Persistência e Geração de Artefatos de Validação

As geometrias finais, representadas como objetos do tipo *LineString* para cada linha e sentido, são persistidas em uma tabela dedicada no banco de dados. Para garantir a transparência e a auditabilidade do processo, são gerados mapas interativos. Estes artefatos visuais contêm camadas de dados independentes e que podem ser ativadas seletivamente, permitindo a inspeção detalhada do esqueleto bruto, do corredor principal após a remoção de *outliers*, do *geofence* poligonal, dos registros de GPS utilizados e da rota canônica final.

5 Engenharia de Atributos

O objetivo central dessa etapa é transformar registros de GPS brutos em um conjunto de dados estruturado e enriquecido, culminando na construção de um modelo de velocidades médias segmentado por dimensões espaciais e temporais. A metodologia foi desenhada para ser escalável e robusta, empregando processamento paralelo e técnicas de inserção de dados em massa.

5.1 Preparação do Ambiente e Estruturação dos Dados de Destino

A fase inicial compreende a preparação da infraestrutura de dados. Duas tabelas relacionais são criadas no sistema de gerenciamento de banco de dados, caso não existam:

- *feature_table*: Tabela destinada a armazenar cada registro de GPS individual, enriquecido com os atributos gerados. A estrutura inclui

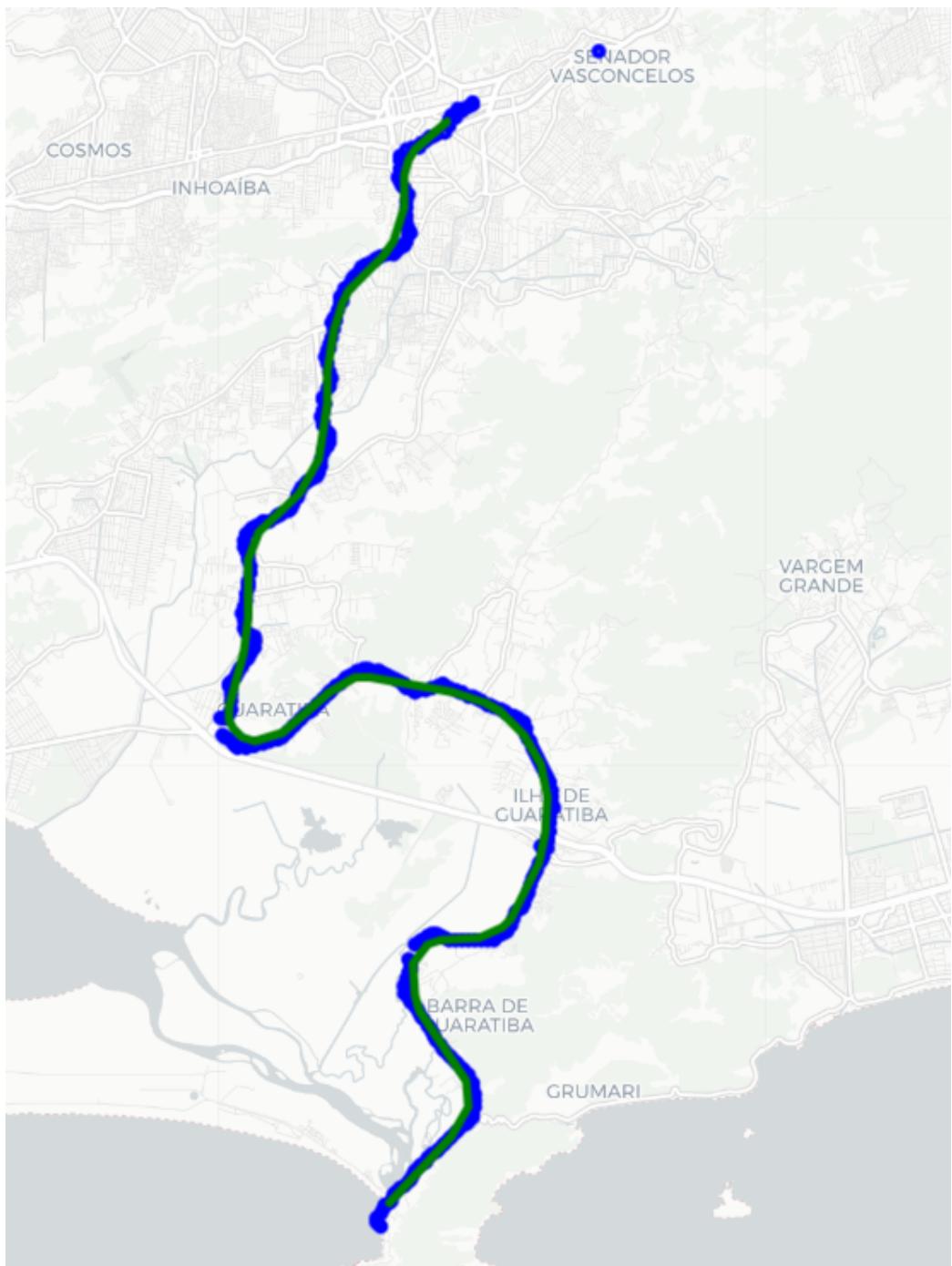


Figura 11: Resultado do traço suavizado para a linha 867.

o identificador do registro original, o identificador da rota canônica associada, o sentido da viagem, o progresso linear ao longo da rota e atributos temporais derivados.

- *velocity_model*: Tabela de agregação final, projetada para armazenar um modelo de velocidades médias. A chave primária é composta pela rota, por um segmento espacial da rota, pelo dia da semana e pela faixa horária.

5.2 Processamento Distribuído por Linha para Geração de Atributos

O núcleo do processo é executado em paralelo, onde cada linha de ônibus é processada por um *worker* independente. Este procedimento é subdividido nas seguintes etapas:

- **Carregamento das Rotas Canônicas**: Para cada linha, o sistema recupera as geometrias das rotas canônicas de ida e volta (objetos do tipo *LineString*), previamente calculadas e armazenadas. A existência e unicidade de ambas as rotas são validadas para assegurar a consistência.
- **Processamento em Lotes (Batch Processing)**: A extração dos registros de GPS brutos da base de dados é realizada em lotes de tamanho fixo (*paginação*). Esta abordagem previne o esgotamento de memória ao lidar com grandes volumes de dados.
- **Projeção Espacial e Associação de Sentido**: Cada registro de GPS em um lote é projetado sobre as geometrias das rotas canônicas de ida e volta. Para determinar o sentido da viagem, calcula-se a distância euclidiana do ponto de GPS a cada uma das duas rotas. O ponto é associado à rota de menor distância. O principal atributo espacial gerado é o *progresso_rota_m*, que consiste na projeção ortogonal do ponto sobre a geometria da rota associada. Esta operação transforma a coordenada bidimensional (*latitude*, *longitude*) em uma medida linear unidimensional, indicando a distância percorrida desde o início da rota.
- **Extração de Atributos Temporais**: A partir do carimbo de data/hora de cada registro (*data_hora_servidor*), são extraídos atributos temporais, como a hora do dia (0 – 23), o dia da semana (1 – 7) e um indicador booleano para fins de semana.
- **Persistência em Massa**: Os dados enriquecidos de cada lote são inseridos na tabela *feature_table* utilizando o comando *COPY* do PostgreSQL.

5.3 Agregação Final e Construção do Modelo de Velocidade

Após a conclusão do processamento de todas as linhas e a população completa da *feature_table*, uma única consulta de agregação *SQL* é executada para construir o modelo de velocidade. Esta consulta realiza as seguintes operações:

1. **Junção de Dados:** Une a *feature_table* com a tabela original de registros (*cleaned_gps_pings*) para acessar o dado de velocidade instantânea.
2. **Discretização Espaço-Temporal:** Agrupa os dados por múltiplas dimensões: Identificador da rota canônica (*id_rota_canonica*); Segmento espacial da rota, obtido pela discretização da variável *progresso_rota_m* em intervalos de 100 metros (*id_segmento*); Dia da semana (*dia_da_semana*). Faixa horária (*faixa_horaria*).
3. **Cálculo da Média:** Para cada um destes grupos multidimensionais, calcula-se a velocidade média e a contagem de registros.
4. **Persistência do Modelo:** Os resultados agregados são inseridos na tabela *velocity_model*. Esta tabela funciona como uma *Lookup Table* (LUT) pré-computada que descreve o perfil de velocidade histórico da rede de transporte, servindo como base para modelos preditivos, como por exemplo, a estimativa de tempo de chegada (*ETA*).