

Relatório Final PIBIC

**Sistema algorítmico de rede neural artificial
para previsão da localização de ônibus de linha
em quedas de sinal**

VÍNICIUS DE OLIVEIRA CIVALI

CAMPINAS, SP
Setembro de 2021

1 Sumário

2 RESUMO.....	3
3 INTRODUÇÃO.....	4
4. ANÁLISE BIBLIOGRÁFICA.....	5
5. OBJETIVOS	7
6. METODOLOGIA	8
6.1. BASE DE DADOS	8
6.2. ALGORITMO DE PREDIÇÃO.....	9
6.3 ALGORITMO DE FILTRAGENS	9
6.3.1 <i>Algoritmo de classificação</i>	10
6.4 ALGORITMO DE MINERAÇÃO.....	17
6.5 ALGORITMO DE ERRO.....	18
7. CONCLUSÕES	18
8. CRONOGRAMA	19
9. BIBLIOGRAFIA.....	20

Lista de Figuras

Figura 1: Exemplo de uma das rotas dos circulares internos da Unicamp.....	5
Figura 2: Topologia de Varfis e Versino.....	6
Figura 3: Fluxograma representativo da metodologia geral	8
Blocos vermelhos representam entrada de dados, enquanto a verde saída.	
Figura 4: Fluxograma representativo do algoritmo de predição	9
Nt: número de amostras no vetor de treinamento da inteligência artificial, a definir; Tx: instantes “data/hora” da recepção da linha de dados, “[$t_0, t_{-1}, t_{-2}, \dots$]”; M: Quantidade de neurônios na primeira camada da rede, a definir;	
Figura 5: Dados originais	10
ts: Data e horário usado como index das linhas; receive_delay: diferença de tempo entre a medição e o tempo de chegada no servidor (usado para o futuro “ts_arrival”, tempo de chegada no servidor); ts_offset_gps: Diferença entre o horário do dispositivo IoT e o horário do GPS; latitude; longitude; altitude; speed: Velocidade do ônibus [km/h]; device_id: Número que identifica o ônibus a qual a linha de dados pertence; month;	
Figura 6: Dados filtrados.....	12
Remoção dos campos desnecessários; Inclusão de um campo exclusivo para o “index”; Cálculo do “ts_arrival” a partir do “ts_gps” e “receive_delay”, o qual foi removido após utilizado; “inRoute”: Campo para classificar a linha de dados na rota diurna ou noturna;	
Figura 7: Dados da tabela ponto-a-ponto da rota diurna	12
Rota diurna: “Index”: Endereçamento das linhas; “Unnamed: 0”: Endereçamento já advindo com os dados. Descartado na sequência; “lat”: latitude; “lon”: Longitude; “Distancia”: Distância entre o ponto em questão, a linha, e o ponto inicial, primeira linha.	
Figura 8: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto.....	12
Plot do gráfico dos pontos GPS advindos da Tabela ponto-a-ponto da rota diurna	
Figura 9: Dados da tabela ponto-a-ponto da rota noturna.....	13
Rota noturna: “Index”: Endereçamento das linhas; “Unnamed: 0”: Endereçamento já advindo com os dados. Descartado na sequência; “lat”: latitude; “lon”: Longitude; “Distancia”: Distância entre o ponto em questão, a linha, e o ponto inicial, primeira linha.	
Figura 10: Gráfico da rota noturna advindo dos dados GPS da tabela ponto-a-ponto.....	13
Plot do gráfico dos pontos GPS advindos da Tabela ponto-a-ponto da rota diurna.	
Figura 11: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto em azul; Classificação feita dos dados reais sobre a tabela em amarelo; Janela de tamanho 6.....	13
A linha reta no gráfico se deve ao uso de apenas 5000 primeiros pontos nesta parcela, desta forma, a partir deste endereço os valores são todos zeros, o que cria está reta entre o ponto 4999, latitude e longitude, e a coordenada (0,0).	
Figura 12: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto em azul; Classificação feita dos dados reais sobre a tabela em amarelo; Janela de tamanho 5.....	14
A linha reta no gráfico se deve ao uso de apenas 5000 primeiros pontos nesta parcela, desta forma, a partir deste endereço os valores são todos zeros, o que cria está reta entre o ponto 4999, latitude e longitude, e a coordenada (0,0).	
Figura 13: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto em azul; Classificação feita dos dados reais sobre a tabela em amarelo; Janela de tamanho 4.....	14

A linha reta no gráfico se deve ao uso de apenas 5000 primeiros pontos nesta parcela, desta forma, a partir deste endereço os valores são todos zeros, o que cria está reta entre o ponto 4999, latitude e longitude, e a coordenada (0,0).

[Figura 14: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto em azul; Classificação feita dos dados reais sobre a tabela em amarelo; Janela de tamanho 3..... 15](#)

A linha reta no gráfico se deve ao uso de apenas 5000 primeiros pontos nesta parcela, desta forma, a partir deste endereço os valores são todos zeros, o que cria está reta entre o ponto 4999, latitude e longitude, e a coordenada (0,0).

[Figura 15: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto em azul; Classificação feita dos dados reais sobre a tabela em amarelo; Janela de tamanho 2..... 15](#)

A linha reta no gráfico se deve ao uso de apenas 5000 primeiros pontos nesta parcela, desta forma, a partir deste endereço os valores são todos zeros, o que cria está reta entre o ponto 4999, latitude e longitude, e a coordenada (0,0).

[Figura 16: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto em azul; Azul: Classificação do ponto 4999 com janela de tamanho 2; Vermelho: Ponto 4999; Laranja \(sob o vermelho\): Classificação do ponto 4999 com janela de tamanho 3 16](#)

Logo atrás a estrela vermelha se encontra a laranja.

[Figura 17: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto em azul; Classificação feita dos dados reais sobre a tabela em amarelo; Janela de tamanho 15..... 16](#)

A linha reta no gráfico se deve ao uso de apenas 5000 primeiros pontos nesta parcela, desta forma, a partir deste endereço os valores são todos zeros, o que cria está reta entre o ponto 4999, latitude e longitude, e a coordenada (0,0).

[Figura 18: Gráfico da rota noturna advindo dos dados GPS da tabela ponto-a-ponto após o algoritmo de classificação..... 17](#)

Após os ajustes do algoritmo de classificação os dados de saída do algoritmo de filtragem ficam desta maneira.

[Figura 19: Exemplo das últimas posições de um vetor de treino 18](#)

Nesta imagem consegue-se notar as linhas de dados perdidas, NaNs, completas de zeros. Levando em consideração a medição das coordenadas GPS a cada 1 segundo, aproximadamente, feito pelo dispositivo do ônibus, constata-se a acurácia do algoritmo de mineração.

Lista de Tabelas

[Tabela 1: Cronograma de atividades atualizado 19](#)

Compleitude: Quantidade em porcentagem estimada de conclusão da etapa.

2 Resumo

Este projeto se insere no contexto do projeto Laboratório Vivo de Mobilidade Elétrica da Unicamp, executado em parceria com a CPFL, no âmbito do projeto P&D da ANEEL PD-00063-3043/2018. Devido às suas grandes dimensões, o campus de Barão Geraldo da Universidade Estadual de Campinas (Unicamp) possui um sistema de transporte interno composto por seis ônibus convencionais a diesel e um ônibus elétrico adquirido no âmbito do projeto distribuídos em cinco rotas circulares. Estes ônibus são monitorados em tempo real por meio de dispositivos equipados com sistema de posicionamento global (GPS), os quais enviam informações em tempo real para um servidor central. Atualmente, a comunicação entre os ônibus e o servidor central é feita via rede de telefonia celular e rede Wi-Fi disponíveis no campus.

A prática tem evidenciado problemas de intermitência e falhas na comunicação dos dados que atrapalham a confiabilidade dos serviços prestados aos usuários do sistema de transporte. Para contornar tais problemas, este projeto propõe o desenvolvimento de um algoritmo baseado em redes neurais artificiais, para estimar a posição dos ônibus quando ocorrerem falhas no sistema de comunicação a fim de fornecer dados de posicionamento dos ônibus mais precisa a um aplicativo de auxílio de mobilidade ao usuário, também no escopo do laboratório vivo. Os algoritmos desenvolvidos devem, portanto, estimar as posições futuras dos ônibus, levando em conta informações extraídas do comportamento históricos dos ônibus e de medidas recentes.

Visto isto, em posse de uma base de dados dos ônibus internos da UNICAMP, consistindo em diversas medições de latitude, longitude, velocidade, tempo de medição, tempo de chegada dos dados ao servidor, entre outros mensurados pelo mesmo laboratório associado a este trabalho, conseguiu-se alcançar o tratamento da base dados, com o algoritmo de filtragem e classificação.

Quanto a síntese dos resultados, há dois principais que englobam os menores, sendo eles os dados pós filtragem/classificação, o que se trata de um arquivo CSV de saída que contém os dados úteis já existentes com os campos desnecessários descartados, assim como as repetições, enquanto outros campos calculados a partir dos úteis, organizados de uma forma inteligente e útil para os passos que se seguirão. E sequencialmente, os vetores de treino, base para o treinamento da rede neural a ser escolhida, concatenados em um CSV de saída, já salientando em quais posições há dados faltantes, desta forma, mostrando-se serem etapas fundamentais.

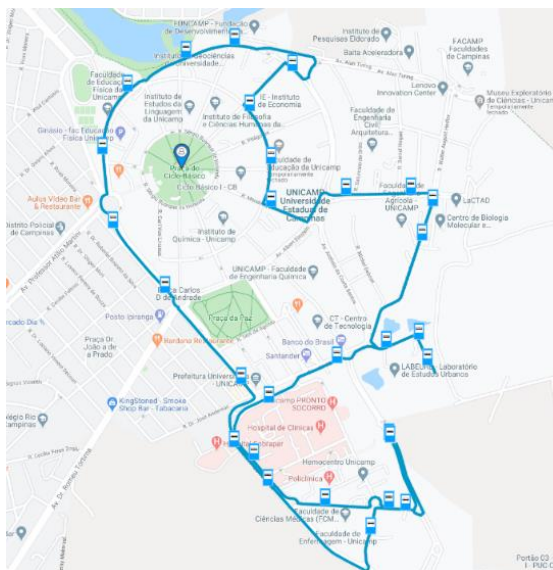
Consequentemente, devido à grande complexibilidade não esperada na filtragem/classificação da base de dados e preparação dos vetores de treino, a etapa de mineração, algumas parcelas não chegaram a serem desenvolvidas ainda, como a escolha e utilização em fato da rede neural e, o treinamento em si. Entretanto, visto a continuidade de desenvolvimento, essas etapas serão completas e terão suas devidas análises.

3 Introdução

Esta proposta se insere no contexto do projeto Laboratório Vivo de Mobilidade Elétrica para Transporte Coletivo na UNICAMP, executado pela Unicamp em parceria com a Agência Nacional de Energia Elétrica (Aneel) e a Companhia Paulista de Força e Luz (CPFL), em que se propõe, de forma inovadora, desenvolver um laboratório vivo de mobilidade elétrica no sistema de transporte público da UNICAMP para investigação de impactos técnicos, econômicos e ambientais da inserção de ônibus elétricos em frotas de transporte urbano de massa [1]. Para tal, é inserido um ônibus elétrico na frota de circulares internos da UNICAMP e serão desenvolvidas e implantadas tecnologias para monitoramento e metodologias para gestão do transporte público do campus, como exemplo da rota de um dos ônibus, segue-se a rota na [Figura 1](#).

Atualmente, os sistemas de monitoramento contendo funcionalidades como um Sistema de Posicionamento Global (GPS), estão em fase de ensaios de campo. Para que as grandezas monitoradas sejam enviados ao servidor de dados são adotadas as redes de telefonia celular e Wi-Fi disponíveis no campus. Entretanto, a prática tem mostrado que estas soluções de comunicação servem muito bem nos casos onde a comunicação em tempo real não é requerida. Nos casos em que a comunicação precisa ser feita em tempo real, para atender as necessidades de operação da frota, comunicando, por exemplo, a localização instantânea dos ônibus, problemas de intermitência e falhas na comunicação dos dados atrapalham a confiabilidade dos serviços prestados aos usuários do sistema de transporte. Nesse contexto, a solução precisaria passar pela concepção e implantação de um novo sistema de comunicação mais robusto e, portanto, mais caro, ou pelo desenvolvimento de algoritmos para previsão da posição dos ônibus na presença de falhas no sistema de comunicação.

Uma revisão de literatura mostra que a aplicação de técnicas de filtragem e mineração aos dados obtidos do GPS, permite a extração de informações como velocidade e aceleração dos veículos em diferentes trechos das vias ao longo do dia. Assim, com o uso de redes neurais artificiais (RNAs) alimentados por informações históricas e recentes extraídas da localização informada pelo GPS é possível estimar com boa aproximação o deslocamento futuro de veículos. Tecnologias utilizadas hoje em dia, tanto por aplicativos, quanto tecnologias de controle internas das empresas, são restritas a conexão contínua, e em caso de queda momentânea, desenformam o usuário.



[Figura 1: Exemplo de uma das rotas dos circulares internos da Unicamp](#)

4. Análise bibliográfica

Em vista do desafio proposto, a pesquisa teórica e sólida para confecção do projeto foi realizada concorrente a discussões e utilização de dados e artigos já confeccionados pelo mesmo laboratório ao qual este projeto se associa (Laboratório Vivo de Mobilidade Elétrica para Transporte Coletivo na UNICAMP [1]). Portanto, como base teórica foram utilizados estudos constantes nas referências a seguir:

[1] Projeto do mesmo laboratório ao qual esta pesquisa se vincula. Os dados base foram captados durante o desenvolvimento desde projeto, trabalhando com os dados captados do ônibus elétrico.

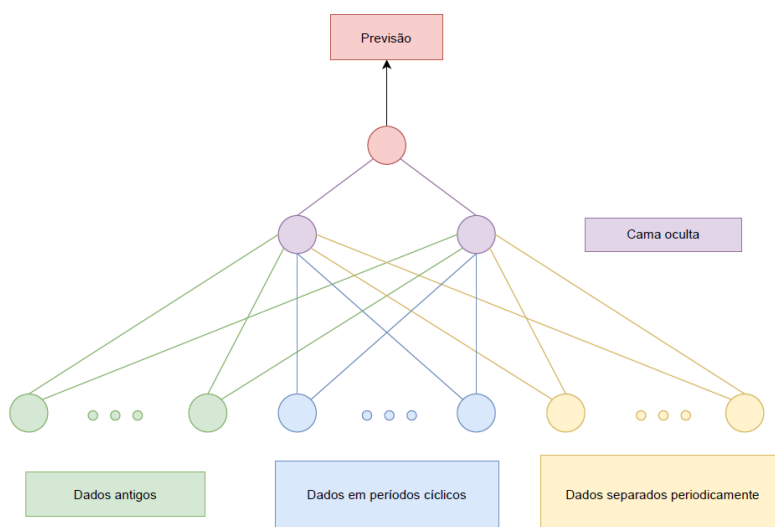
[2] Pesquisa realizada na Unicamp, a qual criou o Circulino. Esta solução inclui a coleta de dados de geolocalização dos veículos de transporte público internos do campus e a geração de informações úteis para a comunidade universitária. Anteriormente seria uma das fontes de dados para este projeto, mas foi optado somente pela utilização dos dados do ônibus elétrico, o qual as pesquisas do laboratório em que o trabalho se vincula circunda.

[3] Este trabalho realiza uma discussão da capacidade de predição, em séries temporais, feita pelas redes neurais artificiais, RNAs, baseando em duas metodologias propostas em [4], com procedimentos salientados em [5], [6]. Desta forma, se mostra importante, visto que é de elevada relevância no contexto desta pesquisa, que trabalha sobre uma base de séries temporais que em estatística, econometria, matemática aplicada e processamento de sinais, é uma coleção de observações feitas sequencialmente ao longo do tempo.

Também se mostra importante por abordar a discussão da escolha do modelamento das RNAs de acordo com as necessidades dos projetos, levando a questionamentos como o tamanho ótimo para rede de treinamento, realizando exemplos empíricos para o desenvolvimento destas discussões. Visto a incerteza no modelo e tamanho da rede neural desta pesquisa, este artigo é e será importante durante todo o projeto.

O artigo propõe e discute duas propostas de modelagens para redes neurais artificiais. A primeira proposta, em **Varfis e Versino (1990)** [4], foca na previsão de séries econômicas, como salientado no artigo e, busca uma modelagem da entrada da rede seguindo uma ideia de captação de diferentes períodos e organizações da série temporal. É um método que propõem em primeira instância a separação dos dados em busca de melhores resultados. Segue-se desta forma:

- 1º) Valores passados da série, mais antigos;
- 2º) Valores passados referentes ao mesmo período de tempo, ciclicidade, visando captar tendências ou ciclos da série;
- 3º) Grupo representativo, grande extensão de dados divididos simetricamente, para captação da sazonalidade da série.



[Figura 2: Topologia de Varfis e Versino](#)

Para determinação da quantidade de neurônios da camada oculta é realizado uma discussão em torno de **Caudill (1990)** [5], que apoia a escolha deste parâmetro no “bom senso”, ou seja, não tão grande para ocorrer memorização de dados, perda de generalização, nem tão pequena, para não necessidade de exaustivas interações para bons resultados.

Não só, mas também, **Rumelhart e Weigend (1990)** [6], segue na ideia de que a melhor rede é a mais generalista sobre os dados em que se trabalha. Assim, propõem uma extensão do método de gradiente descendente, ou seja, fazer uma especiação ponderada das importâncias de cada ligação Inter neuronais, resumidamente, iniciar-se com uma rede grande com um custo associado a cada conexão, custo este decisivo na exclusão de um determinado peso ou conexão. Desta forma, sendo o custo a soma de dois termos, a soma dos quadrados dos erros e, um custo para cada peso contido. Logo, se um patamar de desempenho é alcançado com pouco peso em um grupo de ligações e pesos, estes vão ao longo do tempo perdendo sua importância e deixando de participarem da RNA.

O artigo seguindo a discussão salienta as características “Não observáveis da Série”, como tendência, ciclo, sazonalidade, irregular, etc. E em posse destas, sugere a utilização em uma fase preliminar à RNA utilizando as observações **Rumelhart e Weigend (1990)** [6], buscando assim

estabelecer o número de neurônios da camada oculta. E frisa a ideia a ser seguida nessa busca de dimensão, reduzir o número de neurônios sem perda de capacidade preditiva.

A validações são feitas em torno da aplicação das metodologias citadas em torno de duas séries temporais conhecidas, um índice pluviométrico e índice mensal de passageiros de linhas aéreas. Em ambos os casos a aplicação da metodologia supracitada para a modelagem da RNA se mostrou muito positiva e, desta forma, será usada futuramente.

[7] A análise em que este estudo se debruça é pautada já em um contexto mais próximo ao deste, visto que se trata da discussão do desenvolvimento de três metodologias com utilização de redes neurais de três camadas com inputs de uma coleção de 925.368 pontos de dados de latitude e longitude.

Neste trabalho o enfoque fica em torno da previsão do tempo de viagem de ônibus em suas respectivas rotas e não para predição de posição do ônibus, sem discutir perdas de sinal GPS. Entretanto trabalha com dados similares ao nosso e com utilização de redes neurais.

[8] Método de predição de posição de veicular baseado em grade. É um método que utiliza de uma “grade” de sensores de detecção, o qual os nós são outros carros e a previsão de movimento é feita com o vetor velocidade no instante e a angulação do carro.

Teoricamente é mais simples, e viável de se implementar.

Mas serviria apenas para grandes centros, em que se tem um grande fluxo de carros, ou seja, necessidade de muitos nós

[9] Este método é similar ao [8], entretanto o dado final é uma área prevista para a localização do carro, não uma coordenada. Este inclusive destaca que o método anterior se mostra impreciso.

Tem o mesmo mecanismo de decomposição do vetor velocidade para previsão da posição veicular, entretanto resulta em uma área de probabilidade para a predição. Também considera a utilização de veículos estacionários como nós da “grade”. Mas novamente, é uma solução melhor desenvolvida para centros de grande fluxo.

[10] Trata-se de um algoritmo heurístico, inteligência artificial e otimização matemática, para prever a flutuação da velocidade, assim prevendo o movimento do veículo, entretanto considerando curvas de movimentos de tendência, segundo em qual estrada se encontra e o congestionamento da mesma, considerando também movimentos aleatórios com base probabilística.

5. Objetivos

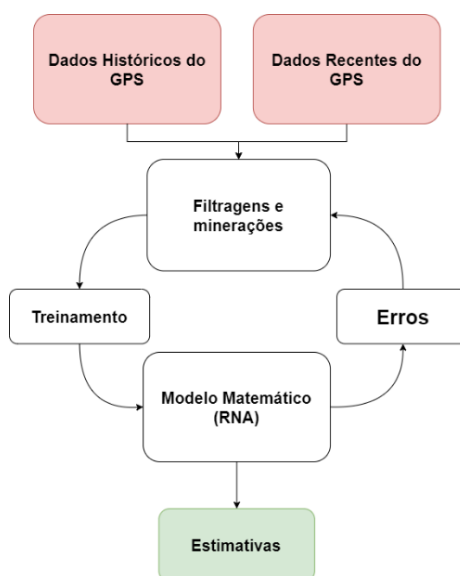
Neste projeto serão desenvolvidos algoritmos capazes de estimar a localização instantânea dos ônibus quando ocorrerem falhas no sistema de comunicação que conecta os dispositivos GPS instalados nos ônibus ao servidor de dados do sistema de transporte. Para tal, serão usadas informações extraídas do comportamento históricos dos ônibus [1] e conhecimentos obtidos dos dados de GPS mais recentes, tendo como base os próprios circulares do campus da universidade em que o projeto se desenvolve. Com isso, tanto as experiências dos usuários, como a gestão do transporte no campus, devem ser beneficiadas pela disponibilidade de dados de localização mais precisos, incluindo momentos de instabilidade e quedas.

6. Metodologia

A literatura sobre técnicas de previsão é vasta e indica uma grande quantidade e variabilidade de métodos promissores que podem ser utilizados de acordo com a quantidade e a qualidade dos dados disponíveis, a natureza das grandezas a serem previstas e os horizontes de previsão. Assim, por exemplo, há métodos adequados para previsões de longo, curto e curtíssimos prazos. No contexto de previsão das posições futuras de veículos, a literatura também é extensa, bastante atualizada e rica [7] ... [10], o que torna este projeto desafiador.

A técnica matemática a ser adotada neste projeto ainda não está definida, porém a literatura indica que as chamadas Redes Neurais Artificiais (RNA) [3]-[6], são uma ferramenta poderosa para a realização de previsão de séries temporais, capazes de realizar prognósticos com o mesmo nível de precisão ou, em alguns casos, até maior que o modelo e séries de tempo estrutural. Redes Neurais fuzzy [7], bem como técnicas clássicas de identificação de sistemas (ARIMA, SARIMA, etc.) se mostram, em primeira análise, promissoras.

Apesar de a RNA a ser utilizada na previsão ainda não estar definida, a metodologia geral de desenvolvimento está clara e é ilustrada na [Figura 3](#).



[Figura 3: Fluxograma representativo da metodologia geral](#)

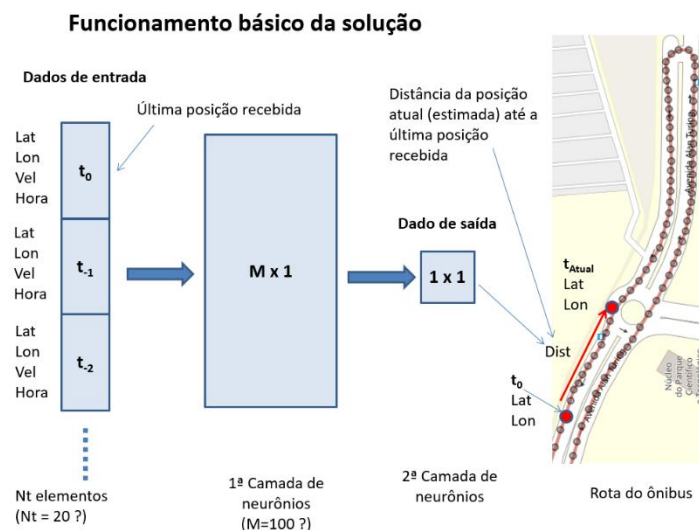
6.1. Base de dados

Os dados necessários para esta empreitada são de fácil aquisição, visto que é feita através dos próprios dados GPS, que na maioria das vezes já é implementado em veículos de transporte público, porém trabalhosa devido a extensão de tempo necessária para criação de uma base de dados sólida. Desta maneira, os dados deste trabalho já foram captados desta forma por trabalhos de pesquisa realizados anteriormente na universidade [1], assim, devido a isto, teve-se uma rapidez no desenvolvimento inicial.

6.2. Algoritmo de predição

O preditor de posição do ônibus, algoritmo principal do trabalho, o qual gera as estimativas, tem característica inerentes que reforçam o uso de métodos de redes neurais. Estas características se expressam no caráter misto de predição de deslocamento e classificatória em que encaramos o desafio, visto a grande quantia de dados de posição e velocidade dos ônibus ponto-a-ponto, com diferença de segundos e, uma tabela, também ponto-a-ponto, das posições durante as rotas. Desta forma a predição classificaria o ônibus em um ponto representativo de deslocamento mais à frente na tabela de posições da rota, logo, redes neurais se mostra bem viável, visto que problemas com esse contexto tem ótima assimilação com elas.

Em virtude disto, a rede neural recebe dados de posicionamentos de instantes anteriores à perda de sinal, produz como resultado um deslocamento aproximado, o qual representa a distância percorrida pelo ônibus entre a última posição recebida e a posição prevista. Com posse da tabela ponto-a-ponto de posições das rotas, utilizando o último dado recebido, e com o deslocamento predito, classifica-se na tabela onde o ônibus está aproximadamente, sendo feito o algoritmo, a grosso modo, desta forma.



[Figura 4: Fluxograma representativo do algoritmo de predição](#)

É válido salientar que a tabela ponto-a-ponto dos dados GPS das rotas foi desenvolvido no projeto Laboratório Vivo de Mobilidade Elétrica para Transporte Coletivo na UNICAMP.

6.3 Algoritmo de filtragens

Consequentemente, há necessidade do desenvolvimento de um algoritmo para geração dos dados que serão usados nas etapas subsequentes. Este algoritmo realiza:

1. Leitura dos dados base;
2. Remoção do campo de milissegundos;

3. Seleção dos dados GPS apenas do ônibus elétrico;
4. Remoção de dados duplicados baseado no tempo de medida dos mesmos (campo “ts_gps”);
5. Descarte dos campos desnecessários, da fonte de dados, para este projeto;
6. Ordenação segundo o tempo de chegada dos dados ao servidor “ts_arrival”;
7. Remoção das 24 linhas iniciais de dados por questões de qualidade;

A fim de salientar a necessidade e importância deste passo temos que, os dados originais, antes da filtragem, apresentavam 3.126.626 linhas, cada uma com 8 campos (colunas), após as filtrações passaram para 643.417 linhas com 4 campos. Desta maneira, faz com que se tenha apenas os dados úteis, sem os demais que na verdade atrapalhavam o desenvolver da RNA.

Não obstante, fez-se necessário a mudança da fonte de dados durante o processo, visto que a fonte primeira, após alguns meses de uso, mostrou-se ter alguns comportamentos não esperados e que para contorná-los gastar-se-ia muito tempo. Desta forma optou-se por tentar outra fonte de dados, do mesmo laboratório, o que deu bons resultados.

Consequentemente a remoção dos campos desnecessários é feita só no passo 5 devido a utilização dos mesmos nos passos intermediários anteriores.

Inicialmente este algoritmo realizava apenas estas filtrações iniciais para que, então, fornecesse os dados apurados ao algoritmo de mineração. Entretanto, após a finalização do mesmo, foi acrescentado a este uma nova parcela de algoritmo, cujo nome explicativo é “Algoritmo de classificação”, discutido na seção que se segue.

ts	receive_delay	ts_offset_gps	latitude	longitude	altitude	speed	device_id	month
2021-01-04 07:46:00	1096214	0	-22.828278	-47.06644	nan	28.3356	1000612	2021-01-01
2021-01-04 07:46:01	845	0	-22.82788	-47.060844	nan	0.0	1000617	2021-01-01
2021-01-04 07:46:01	1094699	0	-22.828224	-47.06649	nan	30.7432	1000612	2021-01-01
2021-01-04 07:46:02	2377	0	-22.82788	-47.060844	nan	0.1852	1000617	2021-01-01
2021-01-04 07:46:02	1093287	0	-22.828163	-47.066547	nan	32.7804	1000612	2021-01-01

[Figura 5: Dados originais](#)

Nota-se que a [Figura 5](#) há a existência de mais campos e, também, no campo “device_id” há dispositivos diferentes que 1000612, ou seja, aos dados dos demais ônibus, além do elétrico.

6.3.1 Algoritmo de classificação

Após a confecção do algoritmo de filtragem, adveio a necessidade da parcela classificatória. É válido lembrar que ambos ocorrem juntos, mas veem de funções distintas, a filtragem é um processo anterior, mais primordial, enquanto a classificação tem sua usabilidade mais ao fim do ciclo. Seguindo, esta parcela realiza a alocação de cada ponto da base de dados filtrada à tabela ponto-a-ponto da rota. Ademais faz-se a distinção dentre rota diurna e noturna do ônibus elétrico, para a classificação.

Desta forma, esta parcela fica encarregada de:

1. Criação de um campo “inRoute” nos dados filtrados pelo algoritmo anterior. Este campo serve para classificar cada linha de dados em uma das rotas;
2. Ler os dados GPS, pré-coletados, ponto-a-ponto da rota diurna e noturna;
3. Extrair os dados de “Latitude” e “Longitude” das rotas;
4. Extrair, dos dados filtrados no algoritmo anterior, um vetor contendo os campos de “Latitude”, “Longitude”, “ts_gps_time” (apenas HH:MM:SS do “ts_gps”) e o campo “inRoute” criado no paço 1;
5. Enquadrar cada ponto, linha de dados filtrados, preenchendo o campo “inRoute” com “d” para rota diurna e “n” para noturna baseado no horário de coleta do mesmo ($07:00:00 < d < 17:45:00$ && $17:45:00 \leq n < 23:59:99$);
6. Calcular a distância de cada ponto, cada linha dos dados filtrados, em relação a todos os pontos da rota especificada em “inRoute”, ou seja, diurna ou noturna de acordo com a classificação adequada;
7. Criar uma máscara de dados em forma de janela de dados, sendo analisado o gap a partir do endereço da classificação do ponto anterior (index da rota adequada associado ao ponto corrente anterior) mais um diferencial (atualmente de 6 linhas de dados para frente da última classificação).
8. Associar ao ponto corrente o endereço da linha de dados da rota apropriada, entre os quais que estão na janela do paço anterior, que apresenta a menor distância. Associar o index identificado à linha do dado corrente.

Desta forma, esse pré-trabalho facilitará a interpretação do resultado da saída da rede neural no futuro, visto que se pode classificar o resultado de saída da RN na rota respectiva ao horário (na tabela ponto-a-ponto de cada rota), e verificar a acurácia da aproximação, visto que com o passo 4 do algoritmo de mineração se sabe a localidade real do ônibus no momento em que se quer fazer a previsão da localização.

Portanto, é algo que será usado no futuro próximo, nos algoritmos de treinamento e na interpretação da saída da RNA. Não só, mas também, isto viabiliza a implementação de duas RNs, uma para rota diurna e outra para noturna, o que é uma abordagem que perpassa vários possíveis erros nos tratamentos de diferenciação entre rotas.

A seguir, consegue-se checar a alteração da filtragem, comparando a [Figura 5](#) e [Figura 6](#).

Index	latitude	longitude	speed	ts_gps	ts_arrival	inRoute	ts_gps_time
0	-22.816105	-47.072548	0.0	2021-01-04 07:01:12	2021-01-04 07:01:43.360000	0	07:01:12
1	-22.816105	-47.072548	0.0	2021-01-04 07:01:10	2021-01-04 07:01:43.361000	0	07:01:10
2	-22.816105	-47.072548	0.0	2021-01-04 07:01:42	2021-01-04 07:01:43.361000	0	07:01:42
3	-22.816105	-47.072548	0.0	2021-01-04 07:01:09	2021-01-04 07:01:43.485000	0	07:01:09
4	-22.816105	-47.072548	0.0	2021-01-04 07:01:08	2021-01-04 07:01:43.580000	0	07:01:08
5	-22.816103	-47.07256	0.0	2021-01-04 07:01:07	2021-01-04 07:01:44.142000	0	07:01:07

Figura 6: Dados filtrados

A fim de elucidar e construir um respaldo da tabela ponto-a-ponto de sinal GPS das rotas diurna e noturna, plotou-se em um gráfico, em que o eixo “x” enquadrou a latitude e o “y” a longitude. Assim, segue a representação gráfica das duas rotas do ônibus elétrico, [Figura 8](#) e [Figura 10](#).

Portanto, fica mais compreensível a visualização do que se trata a dita tabela. Nota-se, de acordo com o início da tabela ponto-a-ponto de ambas as rotas, [Figura 7](#) e [Figura 9](#), que apresenta uma boa qualidade de dados e que, futuramente, estes gráficos servirão de métrica para qualidade do algoritmo de classificação ao longo de sua conclusão.

Index	Unnamed: 0	lat	lon	Distancia
0	0	-22.816280364990230822286321199499070644378662109375	-47.072643280029296875	0
1	1	-22.816152572631839490213678800500929355621337890625	-47.07263946533203125	14.215239877493266007491001801099628210067749023438
2	2	-22.816051483154300427713678800500929355621337890625	-47.072650909423828125	25.51690988771753865194114041514694690704345703125
3	3	-22.815921783447269177713678800500929355621337890625	-47.07263946533203125	39.98647971756972907542149187065660953521728515625
4	4	-22.815792083740230822286321199499070644378662109375	-47.07262420654296875	54.4929797386528917968462337739765644073486328125
5	5	-22.815677642822269177713678800500929355621337890625	-47.072605133056640625	67.3675177159652349700991180725395679473876953125

Figura 7: Tabela ponto-a-ponto rota diurna

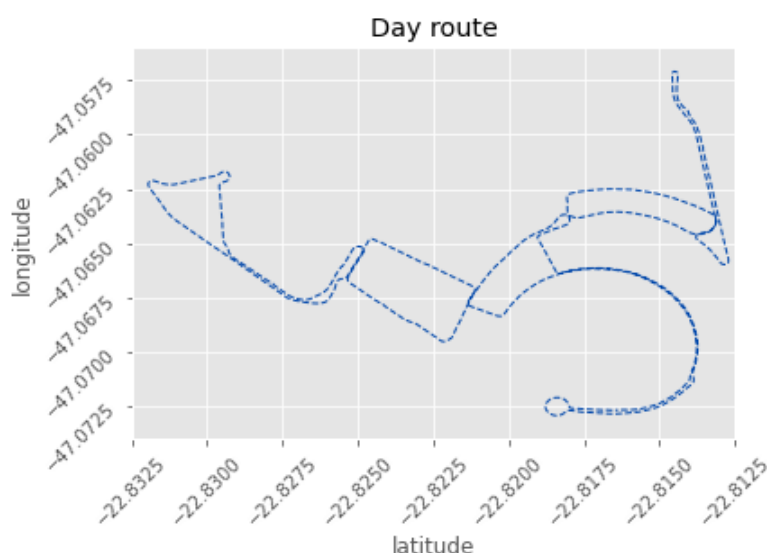


Figura 8: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto

Index	Unnamed: 0	lat	lon	Distancia
0	0	-22.81628036000000037120116758160293102264404296875	-47.07264328000000119800461106933653354644775390625	0
1	1	-22.8161525699999998550993041135370731353759765625	-47.072639469999985959948389790952205657958984375	14.214964501615666492284617561381310224533081054688
2	2	-22.816049580000001384405550197698175907135009765625	-47.07265091000000012400050763972103595733642578125	25.726799890473511567279274458996951580047607421875
3	3	-22.8159179700000009916038834489881992340087890625	-47.072639469999985959948389790952205657958984375	40.408062132854638548451475799083709716796875
4	4	-22.81579208000000136280505103059113025665283203125	-47.0726242100000007440030458383262157440185546875	54.493498938226508698789984919130802154541015625
5	5	-22.815677640000000536701918463222682476043701171875	-47.0726051299999994625977061641216278076171875	67.3680374182082459810771979391574859619140625

Figura 9: Tabela ponto-a-ponto rota noturna

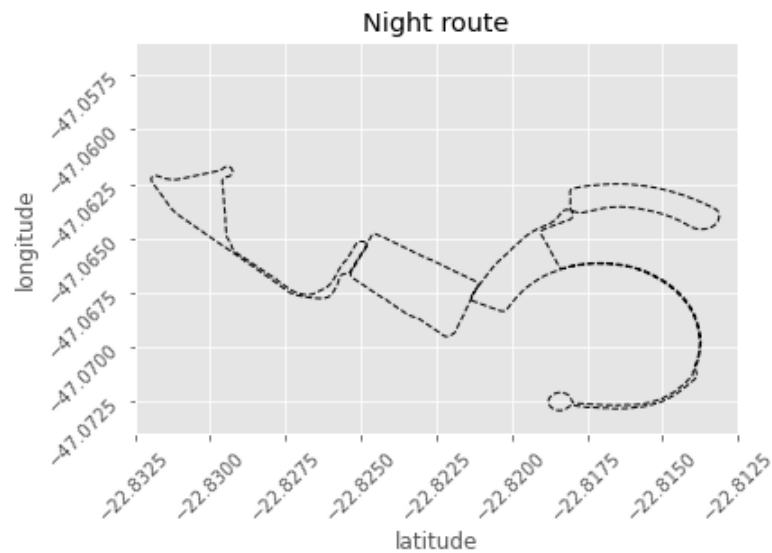


Figura 10: Gráfico da rota noturna advindo dos dados GPS da tabela ponto-a-ponto

Além do mais, é possível notar nos gráficos, [Figura 11](#) ~ [Figura 15](#), a evolução do algoritmo de classificação, onde foi testado janelas de tamanho 6, 5, 4, 3 e 2 linhas respectivamente para os 5000 primeiros pontos.

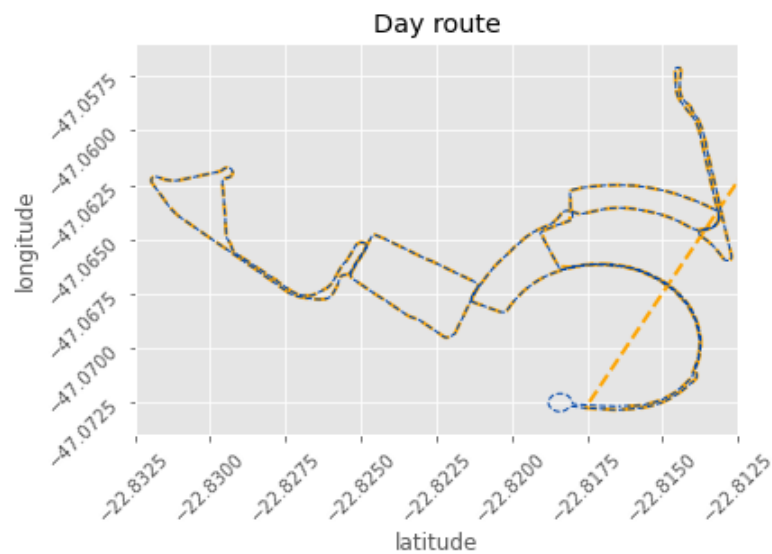


Figura 11: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto em azul; Classificação feita dos dados reais sobre a tabela em amarelo; Janela de tamanho 6

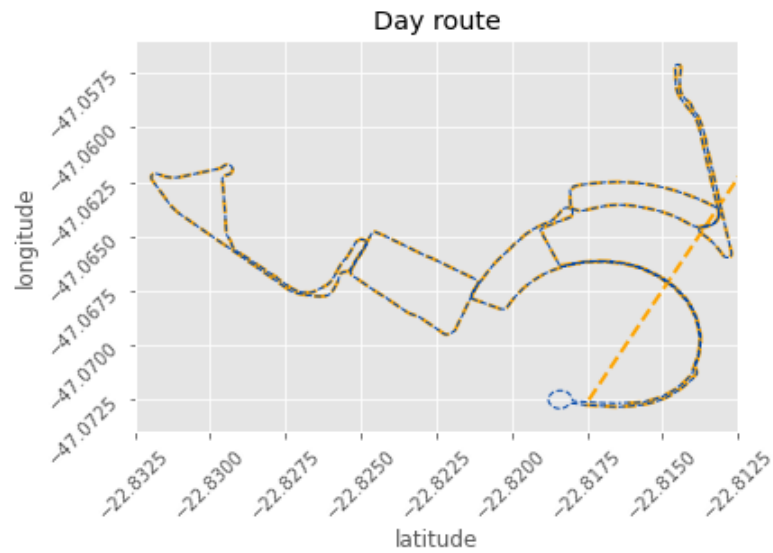


Figura 12: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto em azul; Classificação feita dos dados reais sobre a tabela em amarelo; Janela de tamanho 5

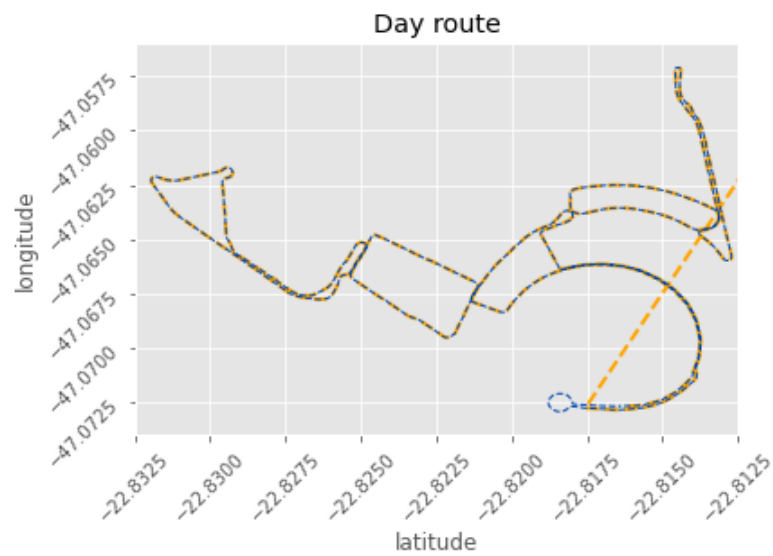


Figura 13: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto em azul; Classificação feita dos dados reais sobre a tabela em amarelo; Janela de tamanho 4

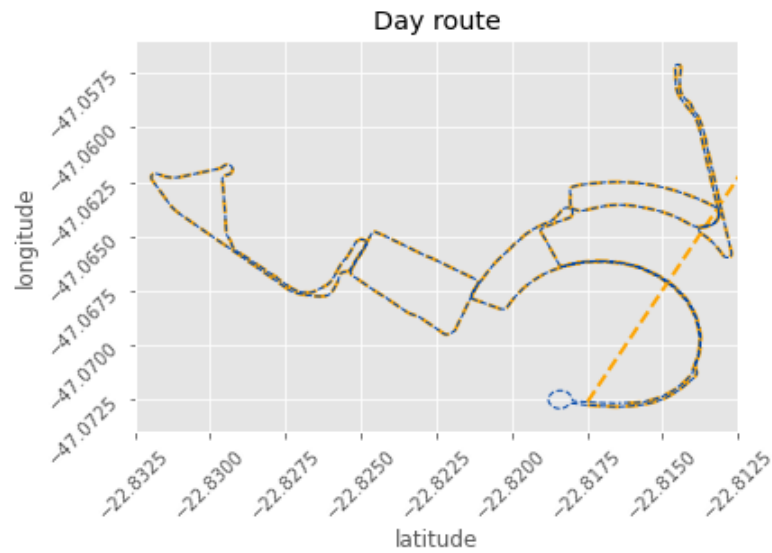


Figura 14: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto em azul; Classificação feita dos dados reais sobre a tabela em amarelo; Janela de tamanho 3

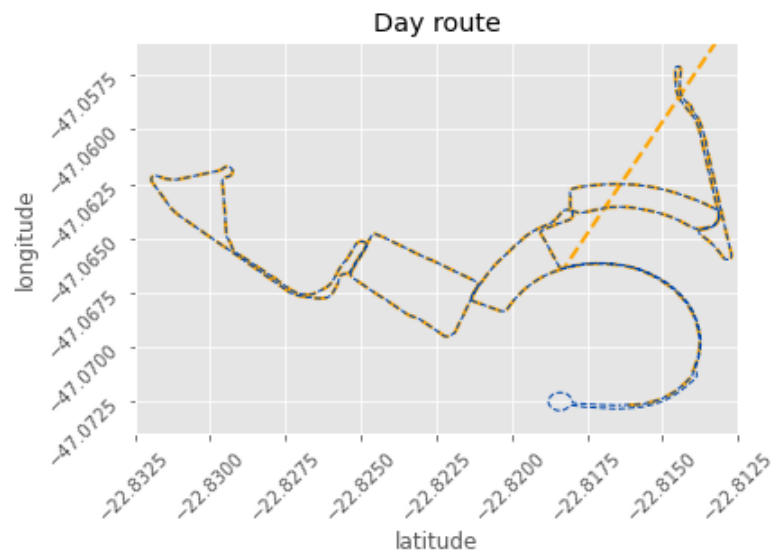


Figura 15: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto em azul; Classificação feita dos dados reais sobre a tabela em amarelo; Janela de tamanho 2

Como se nota, com exceção da janela de tamanho 2, [Figura 15](#), as demais tiveram resultados aproximados. Em vista disto, para confirmar o resultado foi plotado o gráfico da [Figura 16](#).

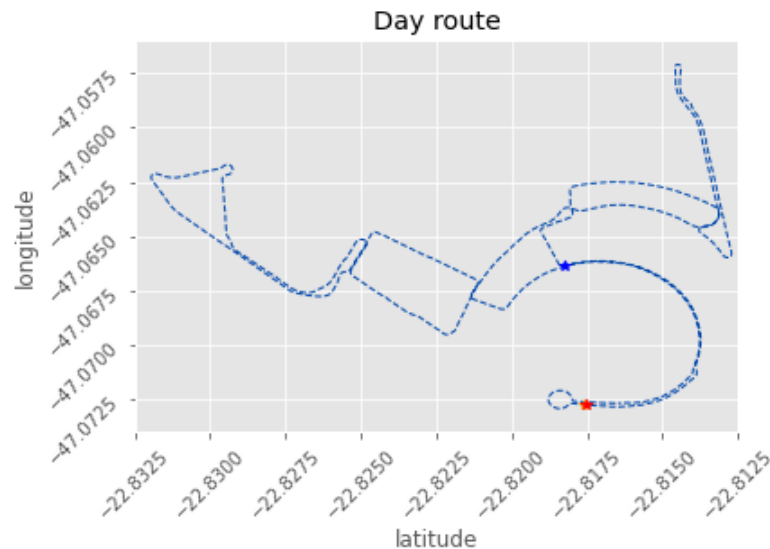


Figura 16: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto em azul; Azul: Classificação do ponto 4999 com janela de tamanho 2; Vermelho: Ponto 4999; Laranja (sob o vermelho): Classificação do ponto 4999 com janela de tamanho 3

Consequentemente, tem-se em vermelho o último ponto, ou seja, de endereço 4999 dos dados reais; em azul, o dado da rota diurna o qual o algoritmo com janela 2 classificou o ponto real 4999; em laranja, justaposto ao vermelho, tem-se o dado da rota diurna o qual o algoritmo com janela 2 classificou o ponto real 4999. Desta maneira, confirmou-se que as janelas de tamanho 6 ~ 3, [Figura 11](#) ~ [Figura 14](#), eram mais adequadas.

Ademais, também foi testado os limites superiores da janela de dados, entretanto, quanto maior o tamanho menos ajustado fica os dados reais a rota, o que se verifica na [Figura 17](#) com uma janela de tamanho 15, o qual apresenta cortes durante o trajeto.

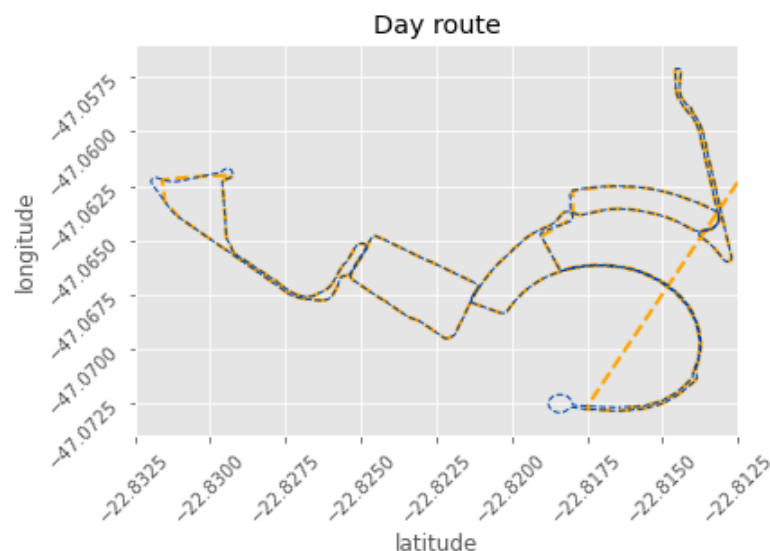


Figura 17: Gráfico da rota diurna advindo dos dados GPS da tabela ponto-a-ponto em azul; Classificação feita dos dados reais sobre a tabela em amarelo; Janela de tamanho 15

Por fim se tem um exemplo dos dados base para o algoritmo de mineração na Figura 18:

Index	latitude	longitude	speed	ts_gps	ts_arrival	inRoute	ts_gps_time	representationPointRoute
0	-22.816105	-47.072548	0.1852	2021-01-04 07:01:27	2021-01-04 07:01:39.917000	d	07:01:27	1
1	-22.816105	-47.072548	0.0	2021-01-04 07:01:28	2021-01-04 07:01:39.726000	d	07:01:28	1
2	-22.816105	-47.072548	0.0	2021-01-04 07:01:30	2021-01-04 07:01:38.739000	d	07:01:30	1
3	-22.816105	-47.072548	0.0	2021-01-04 07:01:31	2021-01-04 07:01:38.665000	d	07:01:31	1
4	-22.816105	-47.072548	0.0	2021-01-04 07:01:32	2021-01-04 07:01:38.170000	d	07:01:32	1
5	-22.816105	-47.072548	0.1852	2021-01-04 07:01:33	2021-01-04 07:01:37.808000	d	07:01:33	1

Figura 18: Gráfico da rota noturna advindo dos dados GPS da tabela ponto-a-ponto após o algoritmo de classificação

É interessante ressaltar o “por que” do uso no plot de gráficos do método o qual cria linhas entre os pontos, e não apenas plota o ponto em si. Esta escolha foi feita para acompanhar o sequenciamento das classificações, pois caso ocorresse saltos grande dentro da classificação apareceriam linhas ligando pontos que não fariam sentido, ou seja, classificações não sequenciais, parecido com alguns pontos no gráfico da [Figura 15](#).

Em virtude do fato dos gaps de dados por perdas afins, este método passara por melhoras ao decorrer da próxima etapa para o melhoramento do tamanho da janela a fim de abarcar grandes intervalos sem dados.

6.4 Algoritmo de Mineração

Por sua vez, o algoritmo de mineração, o mais robusto até então, desenvolve uma série de papéis intermediários ao treinamento, os quais serão discutidos e explicados a seguir.

A função desta parcela do código é definir os vetores que virão a alimentar o treinamento da rede neural para realizar de fato as predições, desta maneira ele fica encarregado por:

1. Aleatoriamente selecionar um instante. Este representa o momento em que um usuário gostaria de acessar a posição do ônibus, mas eventualmente tenha ocorrido atraso no envio da informação do GPS;
2. Procurar o instante selecionado nos dados. Mais precisamente no campo do tempo de chegada dos dados no servidor “ts_arrival”, lembrando que até então os dados estão ordenados segundo o próprio “ts_arrival”;
3. Destacar em um vetor os N dados anteriores ao “atraso da informação”;
4. Procurar o instante sorteado no campo do tempo de medição “ts_gps”. Desta vez, primeiro os dados devem ser ordenados de acordo com o “ts_gps”, nesta parcela tenta-se identificar onde o ônibus está de fato no momento sorteado, ou seja, quando o usuário está tentando saber sua posição.
5. Realizar a inclusão dos dados faltantes “NaNs”, no vetor destacado de N posições, seja por falha de comunicação entre o dispositivo GPS e o servidor durante o percurso, ou pelos dados faltantes entre o último dado recebido no servidor e o instante sorteado (relembrando, que representa o local em que o ônibus se encontrava no momento no momento de queda).

Assim, sucintamente se explana o algoritmo de mineração. Não obstante, como o algoritmo de classificação já pontuou os locais na tabela ponto-a-ponto das rotas, consegue-se interpretar em que local o ônibus de fato está no momento da queda do sinal, ponto destacado no paço 4, e futuramente, entender e avaliar a predição de saída da rede neural.

Index	latitude	longitude	speed	ts_gps	ts_arrival
0	0	0	0	0000-00-00 00:00:00.000000+00:00	0000-00-00 00:00:00.000000+00:00
1	-22.816255...	-47.072662...	0	2021-01-22 17:39:07.000	2021-01-22 17:39:08.811
2	0	0	0	0000-00-00 00:00:00.000000+00:00	0000-00-00 00:00:00.000000+00:00
3	-22.816255...	-47.072662...	0	2021-01-22 17:39:05.000	2021-01-22 17:39:07.134
4	0	0	0	0000-00-00 00:00:00.000000+00:00	0000-00-00 00:00:00.000000+00:00
5	-22.816255...	-47.072662...	0	2021-01-22 17:39:03.000	2021-01-22 17:39:04.916

[Figura 19: Exemplo das últimas posições de um vetor de treino](#)

Assim, foi escolhido como representação dos “NaNs” uma linha com valores nulos, amenos provisoriamente.

6.5 Algoritmo de erro

Da mesma forma, o algoritmo de erro, o qual será base para o ajuste fino da RNA, terá sua análise feita sobre essa diferença entre o local em que o ônibus está e o local previsto, o que salienta a importância do algoritmo de classificação novamente.

Este algoritmo fica a cargo de ajustar os parâmetros das RNs através de um método, a fim de melhorar a localização prevista para o ônibus. Este método, por sua vez, ainda está em análise, visto que o que se seguirá será a construção da RN em si antes deste. Mas como visto na seção anterior, já está preparado o background para que se tenha o ajuste automático.

7. Conclusões

Por fim, este trabalho vem apresentando diversos resultados promissores além de proporcionar ao autor diversas oportunidades de aprendizado, seja através da prática, com análises minuciosas que apenas a mesma concebe, ou pela realização de cursos, estudos e pesquisas com o acompanhamento de diversos orientandos do laboratório de pesquisa associado.

Se tratando dos resultados em si, foi realizado a filtragem adequada dos dados adquiridos durante as viagens dos ônibus, sendo feito a organização, retirada dos dados repetidos. Também foi feita a classificação dos pontos GPS em relação à rota ponto-a-ponto que define a rota, assim como, a construção dos vetores de treino a partir de sorteios aleatórios na tal base filtrada. É válido salientar que esta pesquisa aplicada tem diversos sub resultados que só são perceptíveis em nível de código e debug, desta maneira, esses resultados de fato levaram tempo a serem conquistados, mas o tempo necessário.

Subsequentemente, será feita a escolha e treinamento da rede neural, mas cabe salientar que este desenvolvimento se perpetuará, sendo continuado mesmo após a vigência, e não só, mas também, em se tratando de inteligência artificial, sempre a parte mais trabalhosa é no preparo dos dados e treinamento, partes a qual estão completamente ou quase finalizadas como supracitado.

8. Cronograma

Em relação ao cronograma, muitos passos tiveram suas datas mantidas e concluídas, como os passos iniciais de revisão bibliográfica, entretanto, como dito, a preparação da base de dados e criação da base para o treino da rede neural levou mais tempo que o planejado, principalmente por se tratar de dados GPS, os quais são precisos em nível macro, entretanto no micro, instáveis e difíceis de se esperar uma linearidade na qualidade.

Desta forma, os passos faltantes, a parcela da predição em si, com a escolha da rede neural e, sequencialmente, o treinamento, não foram completados. O “Estudo e análise qualitativa dos métodos promissores” teve seu desenvolvimento feito em início, em conjunto da “Revisão aprofundada da literatura” e cursos de inteligência artificial e redes neurais realizados, entretanto, devido à necessidade do foco na parte de preparatória para o treino em si, a escolha factual da RNA foi postergada para o momento que se tivesse desenvolvido bem a base para a escolha e treinamento.

Em se tratando da “Implementação dos algoritmos para predição” e “Testes e validação de algoritmos para predição”, estas partes sempre estão em avanço devido o fato de serem o total, ou seja, partes maiores que apenas dão nomenclatura para o conjunto dos algoritmos e testes, desta forma, com o desenvolvimento das áreas já decorridas no texto, estas partes avançaram bem e continuarão.

Como último fato, a “Mineração dos dados para extração de informações para alimentação dos algoritmos” não está considerada totalmente completa por conta de alguns tratamentos de casos específicos a ainda serem implementados, como exemplo já citado, na parcela de classificação fazer um tratamento para grandes gaps, dados faltantes sequencias.

Início	Fim	Atividade	Completo (%)
agosto, 2020	setembro, 2020	Revisão aprofundada de literatura	100%
outubro, 2020	março, 2021	Mineração dos dados para extração de informações para alimentação dos algoritmos.	90%
abril, 2021	maio, 2021	Estudo e análise qualitativa dos métodos promissores	50%
maio, 2021	julho, 2021	Implementação dos algoritmos para predição.	70%
agosto, 2020	agosto, 2021	Testes e validação de algoritmos implementados.	70%
agosto, 2021	agosto, 2021	Confecção de relatórios finais.	100%

[Tabela 1: Cronograma de atividades atualizadas](#)

9. Bibliografia

- [1] Ugarte, L. F., Sarmiento, D. N., Mariotto, F. T., Lacusta, E., & de Almeida, M. C.. “Living Lab for Electric Mobility in the Public Transportation System of the University of Campinas”. In 2019 IEEE PES Innovative Smart Grid Technologies Conference-Latin America (ISGT Latin America, 2019, September). IEEE.
- [2] R. A Barbosa, R. P. Sousa, F. A. Oliveira, H. C. Oliveira, P. D. G. Luz, L. T. Manera. Circulino: An IoT solution applied in the university transport service. 4th Brazilian Technology Symposium (BTSym'18). Smart Innovation, Systems and Technologies, vol 140, 2018.
- [3] Luiz Gustavo L. Fernandes, Marcelo Savino Portugal, Philippe O. A. Navaux. “Previsão de Séries de Tempo: Redes Neurais Artificiais e Modelos Estruturais. Gmap.puvrs.br, 1995”. Disponível em: <<https://gmap.pucrs.br/files/publications/articles/5afca684be812ed8e2a901173d20e6b6.pdf>>. Acesso em: 30/04/2020
- [4] Varfis, A. Versino, C. (1990), Univariate Economic Time Series Forecasting, Cambridge University Press, Cambridge.
- [5] Caudill, M. (1990), Neural Network Primer, Miller Freeman Publications.
- [6] Rumelhart, D. E., Weigend, S. A. (1990), "Predicting the Future: A Connectionist Approach", Stanford PDP-90-01, PARC-SSL-90-20.
- [7] Fan Jiang. “Bus Transit Time Prediction using GPS Data with Artificial Neural Networks”.
- [8] Dr.-Ing. Christian Donn, Dr.-Ing. Maximilian Engler, Dr. rer. nat. Alexander Viehl, M.Sc. Rhea Valentina, M.Sc. Marc René Zofka, Dr.-Ing. Lutz Bersiner, Dipl.-Ing., Dipl.-Wirt. -Ing. Ralf Traub, Dipl.-Phys. Georg Müller, Dipl.-Met. Bernhard Mühr “Optimized Utilization of E-Vehicle Range Using Route-Based Online Weather Forecast Data”. Disponível em <<https://www.researchgate.net/publication/283048464>>. Acesso em: 16 de abril 2020.
- [9] Si-Ho Cha, Keun-Wang Lee, and Hyun-Seob Cho, “Grid-Based Predictive Geographical Routing for Inter-Vehicle Communication in Urban Areas”. Disponível em <<https://journals.sagepub.com/doi/10.1155/2012/819497>>. Acesso em: 16 de abril 2020
- [10] Haojing Huang and Shukui Zhang, “A Routing Algorithm Based on Dynamic Forecast of Vehicle Speed and Position in VANET”. Disponível em <<http://dx.doi.org/10.1155/2013/390795>>. Acesso em: 15 de abril 2020.