Universidade Federal do Maranhão
Departamento de Informática - CCET
Programa Institucional de Bolsas de Iniciação Científica

Relatório

Aluno: Paulo Gabriel Borralho Gomes
Professor orientador: Prof. Dr. Alexandre César Muniz de Oliveira

Agosto
2021

Universidade Federal do Maranhão
Departamento de Informática - CCET
Programa Institucional de Bolsas de Iniciação Científica

Relatório

Relatório de Plano de Trabalho de Pesquisa apresentado ao Programa Institucional de Bolsas de Iniciação Científica do Curso Ciência da Computação da Universidade Federal do Maranhão.

Aluno: Paulo Gabriel Borralho Gomes

Professor orientador: Prof. Dr. Alexandre César Muniz de Oliveira

Setembro
2021

Conteúdo

1 Resumo

A geração de dataset é de fundamental importância para trabalhos que envolvam aprendizagem e modelagem de dados. Porém, no tocante a dados de covid-19 de comportamento de usuários em transporte público, há ainda poucos datasets com dados históricos.

Existem vários desafios para geração de datasets a partir de dados históricos. Os mesmos possuem ruídos, inconsistências e falhas que inviabilizam seu uso direto, por consequência é necessário fazer uma filtragem dos dados.

Neste trabalho, formatou-se um dataset de fluxo de passageiros no sistema de metrô de Londres a partir de registro histórico de viagens de usuários não identificados, apontando estações de entrada e saída, e respectivos tempos.

Para tanto, foram desenvolvidos programas para geração de percursos a partir de grafos de representação do sistema do metro de Londres, bem como programas para interpolar os *timestamps* de todas as viagens em estações intermediárias.

Alguns subsets foram apresentados para ilustrar o processo de validação por inspeção que foi realizado neste trabalho. Os programas desenvolvidos possibilitam gerar vários tipos de recortes de dados, dependendo de necessidades específicas, tais como viagens ocorridas em dias específicos, partindo de estações, em determinados horários, apresentando ainda as estações intermediárias com seus respectivos tempos de passagem previstos.

Dataset de viagens podem ser usados em aplicações de aprendizado de máquina para predição de risco de infecção uma vez que pode-se prever a quantidade de pessoas que compartilham um dado espaço físico (estação) em um dado momento (dia e hora).

Palavras-chaves: Mineração de dados, Transporte público, Covid-1.

2 Introdução

A COVID-19, assim como a SARS, faz parte de um modelo tradicional de causalidade das doenças transmissíveis, chamada de tríade epidemiológica (Figura 1), onde o ambiente é o conjunto de todos os fatores que mantém relações interativas com o agente etiológico e o hospedeiro, sem se confundir com os mesmo, o agente embora, de modo geral, se considere que cada doença infecciosa tem seu agente etiológico específico, deve-se ter claro que não há um único agente da doença, e o hospedeiro é aquele onde a doença se desenvolverá e terá oportunidade de se manifestar clinicamente.

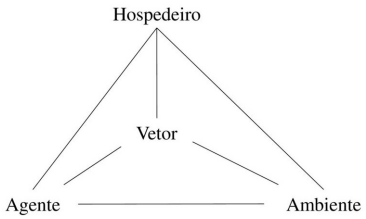

Figura 1: Tríade epidemiológica

Mais de 100 países do mundo estiveram em quarentena para enfrentar a COVID-19, doença que foi relatada pela Organização Mundial da Saúde (OMS) em 31 de dezembro de 2019, em Wuhan, China. A pandemia afeta não somente economicamente os países mas moralmente também. A humanidade já enfrentou outras epidemias e pandemias, como a SARS em 2003 que afetou e a varíola em no início de 1500 que tinha 50% de taxa de mortalidade. A COVID-19 conta com mais de 220 milhões de casos e 4,5 milhões de mortes no mundo até o presente momento, conforme mostra a Figura 2 [1].

3 Justificativa

O raciocínio lógico e outras técnicas de modelagem têm sido usadas para identificar a disseminação epidemiológica desde 1854, quando o médico *John Snow* conseguiu identificar a origem da contaminação de cólera no centro de Londres [2].

[1] https://www.worldometers.info/coronavirus/

[2] https://summitsaude.estadao.com.br/tecnologia/como-o-modelo-matematico-faz-previsao-do-avanco-do-coronavirus/

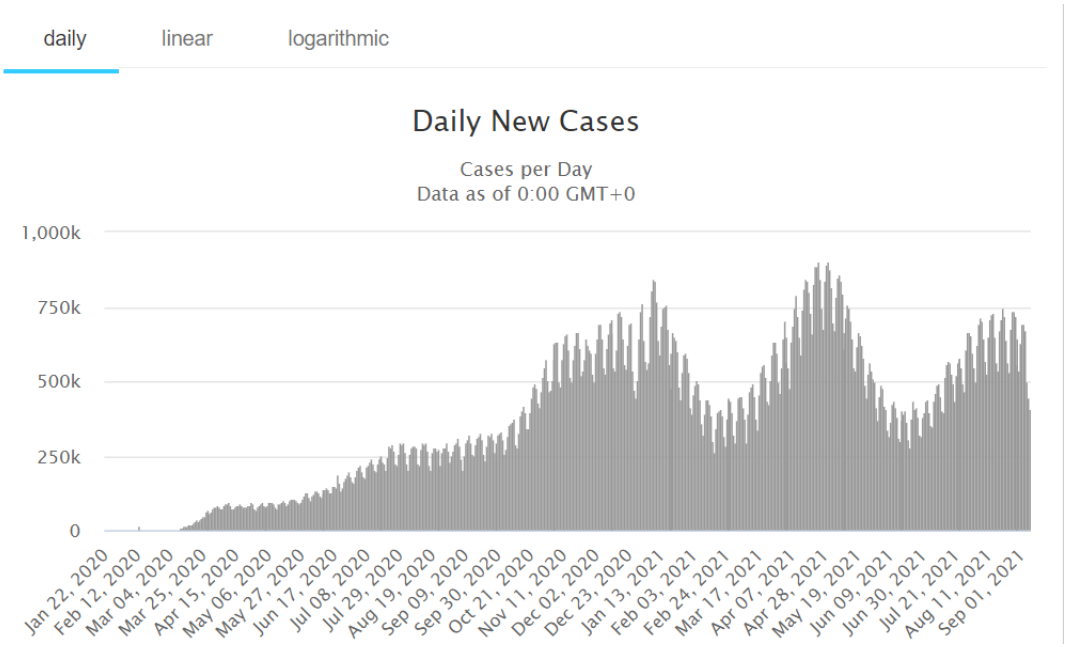

Figura 2: Gráfico de novos casos diários desde janeiro de 2020

No Brasil, algoritmos de regressão foram usados, por exemplo, para predizer o número de casos de dengue treinando algoritmos de aprendizagem de máquina com dados históricos [3].

No tocante à pandemia, é provável que dados históricos de tempos anteriores sejam mais representativos para modelos realísticos de um futuro próximo, o qual interessa às autoridades compotentes modelar e predizer.

A geração de dataset é de fundamental importância para trabalhos que envolvam aprendizagem e modelagem de dados. Porém, no tocante a dados de covid-19 de comportamento de usuários em transporte público, há ainda poucos datasets com dados históricos.

Em (Küpper and Seyfried, 2020), foi realizada uma análise detalhada de como os pedestres usam o espaço usando novos métodos de medição e avaliação. Foram analisadas as trajetórias de passageiros nas plataformas de Berna e Zurique, na Suíça. Os passageiros no embarque e desembarque apresentam comportamentos distintos, considerando as trajetórias, tempos de espera e velocidade média, proporcionando novas medidas para a ocupação do espaço físico. A análise mostrou que é necessário filtrar os dados para se chegar a uma avaliação realista do nível de serviço (Küpper and Seyfried, 2020).

[3] bdm.unb.br/handle/10483/21569

Métodos e modelos têm sido usados para traduzir grandes volumes de dados de várias fontes em novos conhecimentos e percepções que podem ser usados para melhorar o planejamento e as operações em transportes públicos. Em (Luo, 2020) são examinados a ocupação a bordo de veículos, dimensionalidade em fluxos de passageiros em grande escala, agrupamentos de paradas para a construção de matrizes OD (origem-destino) zona a zona e análise de acessibilidade das redes de serviços (Luo, 2020).

Existem vários desafios para geração de datasets a partir de dados históricos. Os mesmos possuem ruídos, inconsistências e falhas que inviabilizam seu uso direto, por consequência é necessário fazer uma filtragem dos dados (Bhatnagar (2016)).

4 Objetivos

4.1 Objetivo Geral

O objetivo deste trabalho é formatar um dataset de fluxo de usuários de um sistema de transporte público por meio de metodologia de filtragem e tratamento sobre dados brutos disponíveis em repositório na Web. Entende-se por fluxo de usuários como sendo dados de registro histórico (logs) de viagens de um conjunto grande de usuários não identificados indicando estações de entrada e saída, bem como itinerário de viagem.

Uma vez de posse desses dados, pode-se aplicar aprendizado de máquina para identificar a probabilidade e o risco de infecção a partir de parâmetros como estações de entrada e saída, dia da semana e horário. Esses cálculos devem levar e consideração a quantidade de pessoas que compartilham um dado espaço físico (estação) em um dado momento (dia e hora).

4.2 Objetivo Específicos

Para tanto, são esperados os seguintes objetivos específicos que contém algum tipo de produto de pesquisa a ser entregue:

1. Geração de um dataset de fluxo de usuários livre de erros e inconsistências em formato CSV;
2. Pesquisa bibliográfica sobre técnicas de programação para integração de dados, aprendizagem de máquina e modelos de propragação de vírus;

3. Levantamento de repositórios contendo dados brutos na Web;
4. Implementação de programas para integração e tratamento de dados;
5. Realização de experimentos que validem o dataset proposto.

5 Metodologia

O trabalho iniciou-se com a procura de um dataset de fluxo de passageiros de um transporte público que contenha as informações do passageiro da sua estação inicial, estação final, horário inicial, horário final e dia da semana.

Londres é uma importante capital mundial em muitos setores, embora talvez mais notadamente em finanças. Com uma economia em expansão, as pessoas estão migrando para a cidade de todo o mundo, atraindo pelo próspero mercado de trabalho e altos padrões de vida.

O cartão TFL Oyster é uma interessante alternativa para obtenção de dados de usuários de um importante sistema de transporte multimodal, incluindo metro, ônibus, trens e até bicicletas. Dados do TFL Oyster são comumente usados em competições de algoritmos de aprendizagem de máquina para descrever o comportamento dos usuários [4].

O processo de filtragem consiste na retirada dos dados inconsistentes e registros com atributos faltantes do dataset. No presente trabalho, foram encontrados registros minimamente consistentes apenas para metrô. Meios de transporte como ônibus e bicicletas apresentaram muitas inconsistências e foram descartados.

Os dados por si não estão adequados para o tipo de dataset que se pretende gerar no âmbito deste trabalho. A topologia do sistema de metrô tem implicações nos dados a serem gerados. Uma vez que não existe um *tracking* completo do usuário, passando por todas as estações, é necessário presumir o percurso de cada pessoa a partir das estações de entrada e saída, bem como horários.

Técnicas de aprendizado de máquina podem ser usadas para criar modelos de comportamento de usuários, permitindo predizer o número de encontros entre usuários a partir dos parâmetros da viagem.

Como consequência pode-se calcular a probabilidade do passageiro se infectar com COVID-19 a partir das estações de origem e destino, bem como dia e hora.

[4] https://www.kaggle.com/benivitai/tfl-oyster-card-journeys-analysis

A distribuição de Poisson Magalhães and Lima (2011) é uma distribuição de probabilidade de variável aleatória discreta que expressa a probabilidade de uma série de eventos ocorrer num certo período de tempo se estes eventos ocorrem independentemente de quando ocorreu o último evento.

5.1 Descrição de atividades

Nesta seção, são descritas as principais atividades desenvolvidas durante esta pesquisa.

1. Levantamento dos principais repositórios de dados de transporte público de grandes cidades;
2. Análise de viagens de cartão TFL Oyster;
3. Desenvolvimento de programas para gerar percursos a partir de grafos de representação das estações de metro de Londres;
4. Desenvolvimento de programas para completar os dados de viagens, considerando os dados de usuários e grafo do sistema de estações;
5. Geração do dataset contendo todas as viagens, respectivos *timestamps* de usuários em cada estação;
6. Aplicação de técnicas de regressão para predição de itinerário no sistema de transporte de Londres;
7. Aplicação de técnicas de probabilidade para medir o risco de contaminação a partir da predição de comportamento.

Utilizou-se o dataset de 2019 em CSV da estação de metrô de Londres que utiliza *TFL oyster card*, um sistema de cartão de viagem eletrônico que armazena não somente as informações necessárias como também o tipo de viagem, tarifa de desconto etc.

Armazenou-se os dados filtrados em uma matriz na qual cada linha representa passageiros e colunas representam informações da sua viagem

Em seguida, utilizou-se o grafo representativo de todas as conexões da linha de metrô de Londres e, juntamente com os dados armazenados, fez-se um merge de cada linha da matriz, traçando a rota (estações intermediárias) mais curta entre a estação inicial e estação final.

O *timestamp* de cada usuário em cada estação intermediária foi presumido uniformemente a partir do tempo total de viagem, disponível no dataset original, ignorando variações na distância entre as estações e eventuais atrasos.

Todos os registros foram salvos um conjunto processado de 7 datasets, cada um referente a um dia da semana.

Em seguida, foram desenvolvidos uma série de programs para leitura dos datasets processado e, de acordo com a estação e hora desejada, calculavam o número de pessoas em cada estação, minuto a minuto, totalizando por hora como mostra a Figura 3.

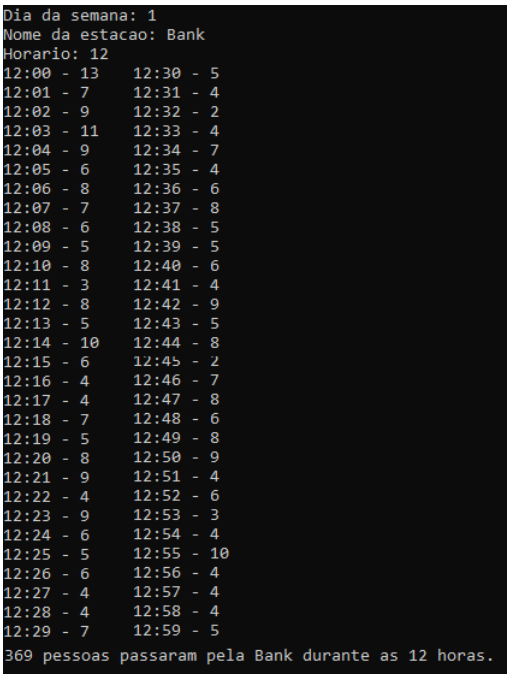

Figura 3: Resultados do programa

5.2 Atividades não executadas

Os experimentos de validação do dataset foram parcialmente realizados por meio de inspeção de algumas amostras de viagens.

Todavia, as etapas de aprendizagem de máquina e distribuição de Poisson (etapas 6 e 7), ainda estão sendo iniciadas por meio da plataforma MATLAB, que é um software interativo de alta performance voltado para o cálculo numérico, atualmente disponível para a comunidade

acadêmica da UFMA [5].

6 Resultados

Nesta seção, alguns subsets são apresentados para ilustrar o processo de validação por inspeção que foi realizado neste Projeto. Os programas desenvolvidos possibilitam gerar vários tipos de recortes de dados, dependendo de necessidades específicas.

As tabelas de 1 a 4 mostram exemplos de viagens ocorridas em um dia de domingo, partindo de uma dada estação, a partir das 10h00, apresentando ainda as estações intermediárias com seus respectivos tempos de passagem previstos, incluindo sempre Victoria, estação de grande importância do centro de Londres. Esse subset das viagens que passam por estação de Victoria está disponível em repositório na Web [6].

Estações	Horário
Earls Court	10:00
Gloucester Road	10:04
South Kensingto	10:08
Sloane Square	10:12
Victoria	10:16
Pimlico	10:21

Tabela 1: Viagem entre Earls Court e Pimlico, passando por Victoria

Estações	Horário
Clapham North	10:00
Stockwell	10:03
Vauxhall	10:06
Pimlico	10:09
Victoria	10:12
Green Park	10:17

Tabela 2: Viagem entre Clapham North e Green Park, passando por Victoria

Estações	Horário
Victoria	10:01
Pimlico	10:04
Vauxhall	10:07
Stockwell	10:11

Tabela 3: Viagem de 10 minutos de Victoria a Stockwell

Estações	Horário
St James's Park	10:02
Victoria	10:06
Green Park	10:10
Hyde Park Corner	10:14
Knightsbridge	10:21

Tabela 4: De St James's Park a Knightsbridge, passando por Victoria, em menos de 20 minutos

[5] https://www.mathworks.com/academia/tah-portal/ufma-universidade-federal-do-maranhao-31545745.html

[6] https://drive.google.com/drive/folders/1UDswrwydrsl5Bdp5urgvJyH5Uc_16LwP?usp=sharing

As figuras de 4 a 8 mostram um a quantidade de pessoas na estação de Victoria, acumuladas em intervalos de tempo específicos, nos dias de Segunda-feira, Quarta-feira, Sexta-feira e Domingo.

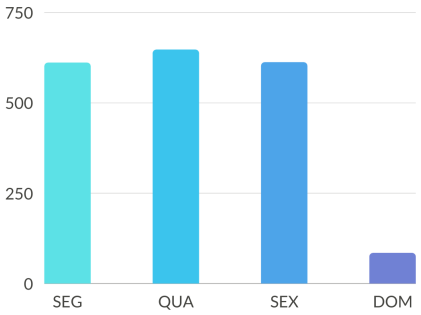

Figura 4: Quatro horas

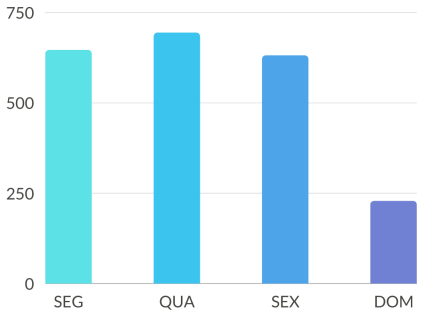

Figura 5: Seis horas

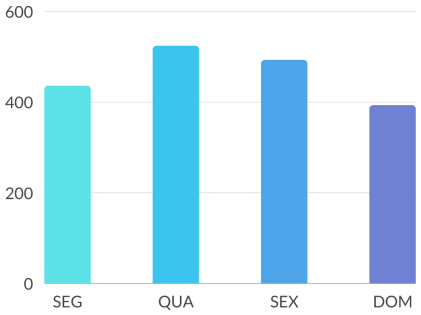

Figura 6: Oito horas

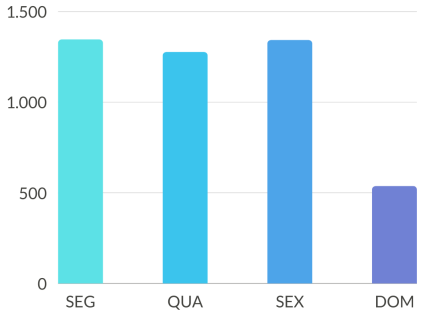

Figura 7: Dez horas

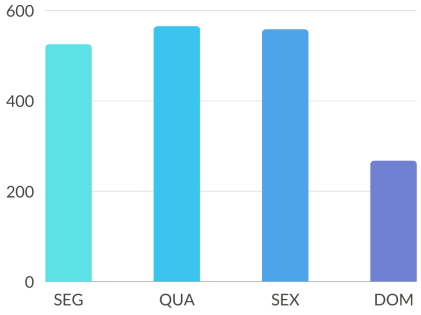

Figura 8: Doze horas

7 Conclusão

A geração de dataset é de fundamental importância para trabalhos que envolvam aprendizagem e modelagem de dados. Porém, no tocante a dados de covid-19 de comportamento de usuários em transporte público, há ainda poucos datasets com dados históricos.

Existem vários desafios para geração de datasets a partir de dados históricos. Os mesmos possuem ruídos, inconsistências e falhas que inviabilizam seu uso direto, por consequência é necessário fazer uma filtragem dos dados.

Neste trabalho, formatou-se um dataset de fluxo de passageiros no sistema de metrô de Londres a partir de registro histórico de viagens de usuários não identificados, apontando estações de entrada e saída, e respectivos tempos.

Para tanto, foram esenvolvidos programas para geração de percursos a partir de grafos de representação do sistema do metro de Londres, bem como programas para interpolar os *timestamps* de todas as viagens em estações intermediárias.

Dataset de viagens podem ser usados em aplicações de aprendizado de máquina para predição de risco de infecção uma vez que pode-se prever a quantidade de pessoas que compartilham um dado espaço físico (estação) em um dado momento (dia e hora).

Como trabalho futuro, estão previstas atividades para adequação do dataset para ser usado em experimentos de aprendizagem de máquina objetivando o cálculo de parâmetros de risco de contaminação por doença infecciosa a partir de viagens em dias específicos da semana, considerando ainda a densidade de probabilidade de usuários compartilharem o mesmo espaço fechado durante suas viagens.

Ainda como trabalho futuro, pretende-se investigar registros de viagens em grandes cidades brasileiras (via autoridade de transporte municipal ou dados de telefonia móvel) para emprego da mesma metodoligia desenvolvida neste trabalho.

Referências

Bhatnagar, V. (2016). *Collaborative filtering using data mining and analysis*. IGI Global.

Küpper, M. and Seyfried, A. (2020). Analysis of space usage on train station platforms based on trajectory data. *Sustainability*, 12(20).

Luo, D. (2020). *Data-driven Analysis and Modeling of Passenger Flows and Service Networks for Public Transport Systems*. PhD thesis, Delft University of Technology. TRAIL Thesis Series no. T2020/2, the Netherlands Research School TRAIL.

Magalhães, M. N. and Lima, A. C. P. d. (2011). *Noções de probabilidade e estatística*. EDUSP.