Mobilidade urbana sustentável: plataforma inteligente de monitorização

Conference Paper · September 2021

CITATIONS
0

READ
1

3 authors, including:

Matilde Pós-de-Mina Pato
Instituto Politécnico de Lisboa
16 PUBLICATIONS **158** CITATIONS
SEE PROFILE

Nuno Datia
Instituto Politécnico de Lisboa
26 PUBLICATIONS **44** CITATIONS
SEE PROFILE

Mobilidade urbana sustentável: plataforma inteligente de monitorização *

João Vaz[1], Nuno Datia[1,2], and M.P.M. Pato[1,3]

[1] FIT - Future Internet Technologies, ISEL - Instituto Superior de Engenharia de Lisboa, Instituto Politécnico de Lisboa
a41920@alunos.isel.pt, {nuno.datia, matilde.pato}@isel.pt
[2] NOVALINCS, Faculdade de Ciências e Tecnologia, Universidade Nova de Lisboa
[3] LaSIGE, Faculdade de Ciências, Universidade de Lisboa

Resumo O parque automóvel circulante em Portugal tem sofrido um crescimento constante, quer em número de veículos, quer na idade média dos veículos. Os congestionamentos de trânsito, com particular incidência nos centros urbanos, e.g. a cidade de Lisboa, resultam em impactos negativos na vida dos cidadãos, onde se incluem problemas de saúde, económicos, sociais e ambientais. Com recurso a diferentes formas de sensorização é possível estudar, compreender e prever fluxos de tráfego em zonas de interesse nos centros urbanos. A partir de modelos de aprendizagem automática, neste trabalho é possível criar e utilizar modelos preditivos de indicadores de tráfego, para diferentes momentos e pontos de interesse na cidade de Lisboa. Os resultados preliminares mostraram que, com o algoritmo XGBoost, é possível prever o tempo de atraso causado por um congestionamento com erros a variar, aproximadamente, entre os 2 e os 3 minutos, verificando-se ainda que a partir da fusão de dados de tráfego, de meteorologia e sociais estes são melhores. Estes modelos podem ser integrados com a plataforma de gestão integrada de Lisboa (PGIL) e contribuir para as tomadas de decisão relativas à mobilidade. São, assim, mais uma ferramenta que permite antecipar futuros congestionamentos e melhorar o planeamento e gestão urbana para que seja possível reduzir os congestionamentos e mitigar os seus consequentes impactos.

Keywords: Intelligent Monitorization Systems · Computational Modelling and Simulation · Predictive Analytics · Visualization

1 Introdução

O parque automóvel circulante em Portugal está em constante crescimento, ultrapassando os $6,2$ milhões de veículos [8]. Dado que os avanços nos sistemas

* Gostaríamos de agradecer à equipa do CGIUL o apoio prestado ao abrigo do Protocolo "Dados ao Serviço de Lisboa". Este trabalho é apoiado pela NOVA LINCS (UIDB / 04516/2020) e pelo LASIGE (UIDB / 00408/2020) com o apoio financeiro da FCT- Fundação para a Ciência e a Tecnologia, através de fundos nacionais.

de comunicação e nas tecnologias da informação potenciam a concentração de pessoas nos centros urbanos [5], é nestes locais que mais se fazem sentir os congestionamentos. Os impactos destes na vida das pessoas são negativos a vários níveis [13,11], destacando-se: (1) *a nível económico*, com o aumento no consumo de combustíveis e maior desgaste do automóvel, bem como mais tempo em deslocações; (2) *a nível social*, através do aumento dos níveis de stress e ansiedade, podendo conduzir a comportamentos agressivos; (3) *ao nível da saúde*, principalmente com impactos induzidos pelo aumento da poluição sonora e do ar, não esquecendo os níveis de ansiedade referidos anteriormente. O congestionamento de tráfego é um problema à escala mundial, sendo um desafio resolvê-lo [10]. É pois de enorme interesse monitorizar, analisar e tentar estimar fluxos de tráfego, para desenvolver políticas de mobilidade que tentem mitigar os congestionamentos e os problemas por estes causados. A Câmara Municipal de Lisboa (CML) está, desde 2017, a utilizar uma plataforma de dados urbanos, denominada de PGIL, para gerir a cidade [14]. Esta plataforma integra várias fontes de dados, internas e externas à CML, recolhidos em tempo real, e disponibiliza *dashboards* temáticos para apoiar as tomadas de decisão de diferentes intervenientes camarários (e.g. Polícia Municipal). No entanto, a PGIL ainda não disponibiliza alguma informação relevante sobre poluição atmosférica e tráfego automóvel, quer pela inexistência ou não integração dos dados [15], quer pela falta de modelos preditivos que permitam um melhor suporte à decisão. Nesse sentido, a CML, através do Laboratório de Dados Urbanos de Lisboa (LxDataLab) [1], propôs um desafio que pretende estudar o tráfego na cidade de Lisboa [2].

A aplicação de técnicas de aprendizagem automática poderá contribuir, não só para a identificação de pontos críticos, mas também na previsão da evolução de congestionamentos, abrindo assim caminho para a criação de soluções de aplicações que visam mitigar esses congestionamentos. Estas técnicas possibilitam previsões que desempenham um papel fundamental na resolução de problemas atuais, influenciando a vida das pessoas, qualquer que seja a cidade do mundo, e em múltiplos contextos, revelando-se como um aspecto fundamental na construção de uma cidade inteligente [3]. Antever uma determinada situação no tráfego permite estimar múltiplas variáveis e, consequentemente, formular planos para atingir as metas programadas de avanço no bem-estar social, tais como a melhoria da gestão do tempo no planeamento de viagens ou a decisão da eventual vantagem de alteração do trajeto. As previsões são hoje entendidas como o ponto de partida de qualquer método de planeamento de uma atividade futura [4], permitindo antever dificuldades e problemas, antecipar soluções e consequentemente reduzir custos, principalmente a nível temporal.

Neste trabalho são analisados, construídos e aplicados modelos preditivos através da utilização de algoritmos de mineração de dados e de aprendizagem automática sobre um conjunto de dados reportado pelo Waze e pelo IPMA (ver Figura 1). O conjunto contém dados relativos a congestionamentos de tráfego na cidade de Lisboa e dados de metereologia, respetivamente. Pretende-se, deste modo, criar modelos capazes de prever congestionamentos e assim dar oportunidade à construção e/ou ao aprimoramento de eventuais planeamentos urbanos.

No entanto, há incertezas nos dados que podem dificultar a criação destes modelos. Assim, os contributos deste trabalho são: (1) Pré-processamento alternativo de valores com incerteza associados ao atraso nos congestionamentos (*delay*); (2) Desempenho dos modelo preditivos usados para estimar o atraso provocado pelos congestionamento, tendo em conta diferentes pré-processamentos.

O artigo encontra-se estruturado da seguinte forma: A secção 2 apresenta uma análise de trabalhos relacionados cujas soluções são semelhantes a este. Na secção 3 segue-se a formulação do problema e a arquitetura da solução adoptada. A secção 4, descreve as decisões tomadas ao nível da transformação e mineração de dados. De seguida, na secção 5 são apresentados e discutidos os resultados da avaliação dos modelos produzidos. Terminamos com a secção 6 que formula as conclusões do trabalho, sendo igualmente apontadas direcções para trabalho futuro.

2 Trabalho relacionado

Neste capítulo é apresentada uma visão geral de trabalhos que, por usarem conjuntos de dados, nomeadamente de meteorologia e de trânsito, bem como algoritmos para a construção de modelos de previsão assim como métricas de avaliação para esses modelos, se relacionam com o problema relatado neste trabalho.

Ni et al. [12] apresenta a previsão do fluxo de tráfego na Califórnia perante eventos que atraem massas. A análise de previsão de tráfego, em determinadas vias, utiliza o fluxo de tráfego por hora, detetado por sistemas de vigilância e por dados provenientes da rede social Twitter.Para a construção do conjunto de dados foi necessário escolher (a) um conjunto de palavras adequadas ao evento em questão, com as quais se obteve o número de *tweets* por hora, (b) o número de diferentes utilizadores que enviaram esses *tweets*, (c) o número de palavras definidas como apropriadas presentes nesses *tweets*, (d) o número de *tweets* que mencionam outro utilizador e, (e) o número de *tweets* que contêm *links*. Os dois conjuntos de dados são cruzados tendo em conta as principais vias de acesso ao local onde irá ocorrer o evento, tendo sido possível obter um conjunto de dados final mais rico. Com base nesse conjunto foram gerados modelos de previsão para essas vias, com uma diminuição de cerca de 7% e 24% do Erro Médio Percentual Absoluto (do inglês, Mean Absolute Percentage Error) (MAPE) e da Raiz do Erro Quadrático Médio (do inglês, Root Mean Squared Error) (RMSE), respectivamente, em relação a um conjunto de dados sem características provenientes de redes sociais. Os modelos de aprendizagem automática utilizados foram construídos através das técnicas de regressão como o Autoregressive Model (ARM), Neural Networks (NN), Support Vector Regression (SVR) e K-Nearest Neighbor (KNN). Os resultados, para os diferentes modelos de regressão aplicados aos fluxos de tráfego provenientes de quatro sensores colocados nas principais vias de acesso a dois espaços destinados a atrações públicas, (*Oracle Arena* e *O.co Coliseum*), permitiram inferir que o modelo gerado através da técnica SVR apresenta os melhores resultados entre as quatro técnicas de regressão experimentadas, obtendo-se os menores valores quer de MAPE, quer de RMSE, independente-

mente do conjunto de dados utilizado. Os autores também concluíram que os valores de MAPE não apresentam resultados tão consistentes como os valores de RMSE.

Zhang and Kabuka [17] exploram a previsão de tráfego tendo em conta os possíveis impactos que determinadas condições atmosféricas podem ter na fluidez do tráfego. O conjunto de dados utilizado contém dados de tráfego captados em tempo real por um conjunto de 39000 sensores espalhados pelas principais áreas metropolitanas de todo o estado da Califórnia, obtendo-se o número de veículos que passaram pelo sensor durante um determinado intervalo de tempo e local. Estes dados foram combinados com dados meteorológicos provenientes do repositório da *National Oceanic and Atmospheric Administration* (NOAA), dados estes que são reportados de hora em hora, ficando então com características como a precipitação, temperatura máxima e mínima. Tendo em conta que estes dados serão aplicados a uma rede neuronal, foi realizada uma normalização tanto das características de tráfego como meteorológicas, de acordo com a normalização *Min-Max*. O conjunto de dados foi dividido em conjunto de treino e conjunto de teste, na proporção contendo 90% e 10%, respectivamente. Após a aplicação dos dados em diferentes técnicas a precisão das previsões obtidas foi avaliada pelas métricas de avaliação de Erro Absoluto Médio (do inglês, Mean Absolute Error) (MAE), Erro Quadrático Médio (do inglês, Mean Squared Error) (MSE) e RMSE pelo que os melhores resultados foram obtidos pelo modelo construído através de *Deep GRU Recurrent Neural Network* (DGRNN), descrito posteriormente, com duas camadas escondidas com 500 neurónios cada uma, apresentando um MAE, MSE e RMSE de 7.9×10^{-3}, 3.76×10^{-5} e 1.9×10^{-3}, respectivamente. O modelo proposto apresenta sempre os melhores resultados, em comparação com os resultados obtidos para os restantes modelos construídos através de outras técnicas tais como redes neuronais simples, redes neuronais com *Long Short-Term Memory* (LSTM), SVR e Random Forest (RF), para as três métricas de avaliação utilizadas.

Neste trabalho pretende-se também tirar partido da utilização de conjuntos de dados externos aos de tráfego, nomeadamente, meteorológicos sendo que a utilização destes permite, por norma, a obtenção de melhores resultados. Ao nível de algoritmos, iremos explorar um outro algoritmo de regressão, bem como uma transformação da variável dependente. Na avaliação dos modelos são usadas um subconjunto das métricas apresentadas.

3 Formulação do Problema

Como mencionado no primeiro capitulo, o conjunto de dados a utilizar é proveniente do Waze e reporta informação relativa a congestionamentos de tráfego na cidade de Lisboa, intitulados por *Jams*. Contudo pelo que foi analisado na literatura, verifica-se que os modelos preditivos apresentam melhores resultados quando construídos com outras características relevantes, pelo que se decidiu utilizar também um conjunto de dados, denominado de *Weather*, proveniente do IPMA que reporta informação relativa a dados meteorológicos,. As carac-

Tabela 1. Descrição das características presentes no conjunto de dados *Jams*.

Nome do campo	Descrição
country	Código representativo do país segundo a norma ISO 3166-1 [9]
city	Nome da cidade ou estado
level	Nível de congestionamento de tráfego (0 = via completamente livre, 5 = via completamente congestionada)
length	Comprimento do congestionamento em metros
turn_type	Tipo de curva
type	Tipo de zona de recolhimento de tráfego
uuid	Identificador único de congestionamento
end_node	Saída mais próxima do final do congestionamento
speed	Velocidade média do tráfego em metros por segundo
road_type	Tipo de estrada
delay	Tempo de atraso do tráfego em comparação com a via completamente livre em segundos (-1 = via completamente congestionada)
street	Nome da rua
pub_millis	Data da publicação em Unix time
bbox	Conjunto de coordenadas que representam o local do congestionamento

terísticas presentes nestes conjuntos de dados podem ser observadas nas Tabelas 1 e 2, respectivamente. Tendo como objetivo a previsão de congestionamentos, a característica a prever (variável dependente) será o `delay`, presente no conjunto de dados *Jams*. Uma análise desta característica permitiu verificar que, de semana para semana, os seus valores máximos apresentam algumas diferenças podendo estas indicar, por vezes, a presença de outliers. Verificou-se ainda que existem cerca de 50% de entradas com o valor -1, valor este que representa uma via completamente congestionada. Dado que não existem atrasos negativos, este valor gera incerteza quanto ao tempo real de atraso do congestionamento, assim como problemas ao nível da utilização de modelos de previsão.

3.1 Arquitetura geral da solução

A solução passa pelo desenvolvimento de um sistema capaz de recolher e armazenar dados de diversas fontes de informação, com o objetivo final de deles extrair conhecimento recorrendo a modelos de previsão e de visualização analítica.

A arquitectura da solução é apresentada na Figura 1 e divide-se nos seguintes componentes: (I) Fontes de Dados, onde estão representadas as fontes dos dados usado para criar o modelo preditivo; (II) Servidor, que será o responsável pela realização de pedidos HTTP às fontes de dados, assim como do

Tabela 2. Descrição das características presentes no conjunto de dados *Weather*.

Nome do campo	Descrição
YYYY-mm-ddThh:mi	Data e hora da observação
idEstacao	Id da estação meteorológica observada
intensidadeVentoKM	Intensidade do vento registada a 10 metros de altura em kilometros por hora
temperatura	Média da temperatura do ar registada a 1.5 metros de altura numa hora em graus centígrados
idDireccVento	Classe do rumo do vento ao rumo predominante do vento registado a 10 metros de altura. (0: sem rumo, 1 ou 9: "N", 2: "NE", 3: "E", 4: "SE", 5: "S", 6: "SW", 7: "W", 8: "NW")
precAcumulada	Valor acumulado da precipitação registada a 1.5 metros de altura numa hora em milímetros
intensidadeVento	Intensidade do vento registada a 10 metros de altura em metros por segundo
humidade	Média da humidade relativa do ar registada a 1.5 metros de altura num hora em percentagem
pressao	Média da pressão atmosférica reduzida ao nível médio do mar numa hora em hectopascal
radiacao	Radiação solar em quilojoule por metro quadrado

armazenamento dos respetivos dados recolhidos através desses mesmos pedidos; (III) Base de Dados, onde irão ficar armazenados os dados recolhidos; (IV) Pré-Processamento, que consiste no tratamento dos dados recolhidos para que estes possam ser integrados tanto na ferramenta de extração de conhecimento como na ferramenta de visualização gráfica; (V) Extração de Conhecimento, que irá permitir a construção e aplicação de modelos preditivos; e, por fim, (VI) Visualização, que permitirá a visualização gráfica dos dados diretamente integrados num mapa. Ambos os conjuntos de dados estão a ser recolhidos e armazenados de forma automática, sendo que os dados relacionados com a meteorologia estão a ser recolhidos a cada hora e os dados relacionados com o tráfego estão a ser

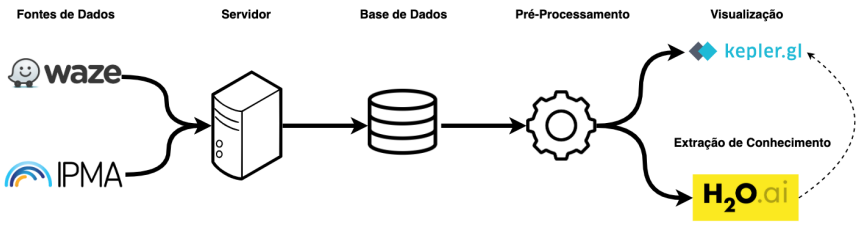

Figura 1. Arquitectura da solução

recolhidos de 5 em 5 min. A escolha da periodicidade está relacionada com o fato destes dados serem o principal objeto de estudo do trabalho, mas também pelas suas alterações frequentes.

4 Modelação

Após obtenção dos dados, é necessário um pré-processamento que consiste na limpeza dos conjuntos de dados de modo a não trabalharmos com valores irrelevantes, duplicados ou vazios/omissos.

Para o conjunto de dados *Weather*, como estão a ser recolhidos dados de três estações meteorológicas de Lisboa, o processamento consiste em agregar a informação dessas três estações para cada hora do dia. A agregação foi realizada através da extracção do valor médio, valor mínimo e valor máximo para cada uma das características do conjunto de dados. Nesta fase de escolha de características para a construção de modelos, decidiu-se manter apenas as características `intensidadeVentoKM`, `temperatura` e `precAcumulada` devido ao facto de se acreditar que estas podem ser as que mais influencia têm no tráfego.

Para o conjunto de dados *Jams*, as características que se mantiveram foram o `datetime`, `city`, `street`, `level`, `length`, `end_node`, `speed`, `road_type`, `delay` e `position`. Sendo que o `datetime` corresponde a uma adaptação da característica `pub_milis`, para facilitar a interpretação humana, e o `position` corresponde a um objecto GeoJSON que contêm as coordenadas exactas da localização da via onde ocorreu o congestionamento. As restantes características tais como `country`, `turn_type`, `uuid` e `bbox` foram removidas.

De seguida, é realizado um processamento que consiste em agregar conjuntos de linhas idênticas por um determinado período de tempo, tendo como objetivo garantir que para o período de tempo em análise, neste caso 30 minutos, não existem entradas repetidas pois isso poderia por em causa a precisão dos modelos que serão construídos posteriormente, apresenta também como beneficio a redução do volume de dados.

Com o objetivo de aumentar o desempenho dos modelos foram ainda derivadas novas variáveis a partir deste conjunto de dados, sendo estas: (i) `feriadoFimSemana`, que indica se o congestionamento ocorreu num feriado e/ou num fim de semana; (ii) `vesperaFeriadoFimSemana`, que indica se o congestionamento ocorreu na véspera de um feriado e/ou na véspera de um fim de semana; (iii) `diaDaSemana`, que indica qual o dia da semana (por extenso) em que o congestionamento ocorreu; (iv) `horaDePonta`, que indica se o congestionamento ocorreu durante as horas de ponta (7h/10h e 18h/20); e, (v) `horaDoDia`, que indica em que parte do dia ocorreu o congestionamento (manhã, tarde ou noite).

Uma solução para a incerteza relacionada com o `delay` igual a -1, foi de atribuir valores substitutos com base nos dados existentes e no domínio do problema. Dada a natureza desta variável, o valor atribuído não tem significado, então todos os valores iguais -1 foram substituídos por um valor que resulta da soma do valor máximo existente nos dados com N vezes o desvio padrão, sendo

Tabela 3. Descrição dos modelos.

Modelos	Descrição
A	conjunto de dados final sem alteração nos valores da variável dependente, i.e., com os valores de -1 presentes em `delay`
B	conjunto de dados final de onde foram removidos as instâncias que apresentam valores de -1 na variável dependente, `delay`
C	conjunto de dados final onde os valores que apresentavam -1 na variável `delay` foram substituídos por: $\max(\texttt{Delay}) + 1 \times \texttt{DesvioPadrao}$
A+, B+ e C+	extensão aos Modelos A, B e C, respectivamente, onde se acrescentaram as novas características derivadas

N o número de vezes que se pretende aumentar o valor resultante em relação ao seu valor máximo.

Foram gerados seis variantes do conjunto de dados, com as quais foram geradas seis modelos preditivos, como descrito na Tabela 3.

Todos os modelos foram construídos através da ferramenta `H2o` [6]. Dado que esta é adequada à realização da modelação, e que já se encontra em utilização na PGIL, facilitando assim uma possível integração destes mesmos modelos. Utilizou-se o algoritmo *XGBoost* [7] que é um algoritmo de aprendizagem supervisionada baseado na técnica de *boosting*. Os dados utilizados na construção destes modelos são relativos a toda a cidade de Lisboa na semana 49 do ano de 2020 e apresentam 86532 registos sendo as previsões realizadas relativas à semana seguinte cujo total de dados apresentam 97074 registos.

5 Resultados e análise

Os modelos desenvolvidos foram avaliados e comparados segundo as métricas *R Squared* (R^2), MAE, RMSE e RMSLE. Foram escolhidas estas métricas pelo facto de serem também as utilizadas nos trabalhos existentes na literatura e permitirem a análise e comparação dos diferentes modelos, assim como uma melhor percepção dos dados aquando da sua utilização individual ou em conjunto.

O R^2 representa o grau em que o valor previsto e o valor atual se movem em uníssono. Apresenta valores normalizados entre 0 e 1, sendo que 0 representa nenhuma correlação e 1 representa uma correlação total o que permite uma comparação mais directa entre modelos e verificar o impacto da transformação da variável dependente. Já no caso de MAE, que representa a média dos erros absolutos, sendo uma medida mais robusta perante *outliers* é útil para entender o tamanho do erro devido às unidades que utiliza (mesmas unidades que a variável em estudo — segundos). Quanto menor for, melhor é o desempenho do modelo. No caso de RMSE é uma medida que representa a raiz do erro quadrático médio, medindo o quão bem o modelo pode prever um valor continuo, é sensível a *outliers* ao contrário de MAE pelo que através das análise destas duas medidas em

Tabela 4. Medidas de avaliação dos modelos construídos.

Modelo	**R2**	**MAE**	**RMSE**	**RMSLE**
Modelo A	0.972	**7.779**	21.128	—
Modelo B	0.961	13.389	26.824	0.132
Modelo C	0.971	142.324	201.638	0.126
Modelo A+	**0.973**	7.841	**20.758**	—
Modelo B+	0.959	13.536	27.331	0.133
Modelo C+	0.972	142.122	201.377	**0.126**

conjunto é possível determinar a presença ou não de *outliers*. Também utiliza as mesmas unidades que o valor previsto. Quanto menor for, melhor é o desempenho do modelo. Por último, no caso de RMSLE é uma medida que representa a raiz do erro médio quadrático e logarítmico medindo a proporção entre valores reais e previstos. Quanto menor for, melhor é o desempenho do modelo. É uma medida adequada, uma vez que se considera que uma previsão mais baixa que o valor real é mais inconveniente do que uma previsão mais alta. Fazendo a ligação para o domínio do problema, chegar a uma determinada via e esta estar congestionada pensando o utilizador que não iria estar é bastante mais inconveniente do que chegar a uma via que se pensava estar com algum congestionamento e na realidade não estar. Todos os modelos usaram a parametrização por omissão do *XGBoost*.

Na Tabela 4 podem ser observadas os valores obtidos para as métricas de avaliação para os modelos construídos, constatando-se que ao nível da medida de avaliação R^2 os valores são todos relativamente próximos sendo os dois melhores apresentados pelos modelos A+ e C+, possivelmente devido ao fato destes apresentarem as características adicionais em relação aos modelos que apenas utilizaram o conjunto de dados original. Relativamente aos piores valores desta métrica, estes foram obtidos pelos modelos B e B+ constatando-se que ao remover efectivamente as entradas com o valor de `delay` com -1 estamos a prejudicar o desempenho do modelo para além de estarmos a retirar-lhe a capacidade de previsão em vias totalmente congestionadas. Em relação à medida de avaliação MAE, observa-se que o melhor valor foi obtido através do modelo A, com um erro médio absoluto de menos de 8 segundos. No entanto esta métrica pode ser enganadora aquando observada neste modelo em questão pois estão a ser previstos valores positivos muito próximos de 0 para entradas que deveriam apresentar valores -1 (via totalmente congestionada), fazendo com que esta medida seja relativamente baixa embora a previsão não esteja de todo correta. Já os valores, desta medida, obtidos pelos modelos C e C+ aparentam um pior desempenho na medida em que apresentam valores bastante mais elevados mas tendo em conta a transformação realizada na variável `delay` são valores perfeitamente aceitáveis e, sobretudo, mais reais (erro médio absoluto de aproximadamente de 2 minutos e 22 segundos) tendo em conta o domínio do problema. Em relação à medida

Figura 2. Representação gráfica dos resultados na ferramenta Kepler

de avaliação RMSE, observa-se que o melhor valor foi obtido através do modelo A+ apresentando um erro de aproximadamente 21 seg. A análise desta métrica segue exactamente a mesma linha de pensamento da análise do MAE pelo que os modelos C e C+ também aqui apresentam valores mais significativos (erro de aproximadamente 3 min e 21 seg) indicando uma maior proximidade a valores reais assim como a possível presença de *outliers*. Por último, em relação à medida de avaliação RMSLE pode-se observar que o melhor valor obtido é proveniente do modelo C+. No entanto, todos os restantes modelos apresentam valores bastante próximos deste, à excepção dos modelos A e A+ onde não foi possível o cálculo desta medida devido à presença de valores negativos.

Com base na análise e comparação de todas estas medidas entre os modelos desenvolvidos, foi seleccionado o **Modelo C+** para a realização de previsões por ter sido considerado um dos melhores modelos mas também pelo facto dos seus resultados se aproximarem mais da realidade tendo em conta o domínio do problema.

5.1 Integração em *Dashboard*

Após a realização de previsões com base no modelo seleccionado, estas foram integradas num mapa interactivo da cidade de forma a que possam ser mais facilmente analisadas. A integração foi realizada através da ferramenta Kepler [16], que permite a análise de dados georeferenciados integrados num mapa. Estando a ferramenta apta a lidar com grandes quantidade de dados, considerou-se ser adequada a utilizar. Permite várias operações ao nível de filtragem e visualização dos dados, nomeadamente omitir as características pretendidas, definir determinada característica sob uma escala de cores, reproduzir os dados perante um determinado intervalo de tempo desde a data inicial presente nos dados até à data final, entre outros. Esta integração pode ser observada na Figura 2, onde as cores dos troços dos congestionamentos representam o valor da característica `delay`, sendo que as cores mais claras, nomeadamente o amarelo, representam

um congestionamento mais ligeiro enquanto que as cores mais escuras representam um congestionamento mais demorado sendo este o caso da cor roxo. É possível ainda observar com mais detalhe um congestionamento em específico, ao carregar no seu troço, apresentando este todas as características relativas ao mesmo. Esta integração tem como objecto mostrar a integração do modelo do `H2o` num *dashboard*, bem como identificar os problemas de visualização analítica neste domínio de problema.

6 Conclusão

Neste trabalho apresenta-se um sistema completo, desde a recolha dos dados à visualização das previsões de congestionamento de trânsito na cidade de Lisboa. Em particular, investigou-se o impacto que a transformação da variável associada ao tempo de congestionamento, `delay`, tem no modelo final. Este sistema pode ser aplicado a outros centros urbanos, necessitando apenas de uma recolha de dados meteorológicos e informação de congestionamentos de tráfego do local.

Com base nos resultados obtidos, observa-se que é possível prever o tempo de atraso causado por congestionamentos, com erros a variar entre os 2 e 3 minutos (estimativa pessimista), aquando da utilização de todas as características disponíveis das duas fontes de dados usadas. O tratamento que envolveu a variável `delay`, ou seja, a transformação dos seus valores de -1 mostrou-se adequada, na medida em que se consegue obter uma estimativa do valor da variável, permitindo uma modelação mais aproximada da realidade. Os modelos que envolveram esta transformação (C e C+) apresentaram ambos métricas de avaliação muito idênticas, com R^2, MAE, RMSE e RMSLE a apresentarem valores aproximados de 0.97, 142, 201 e 0.12 respectivamente. O modelo C+ consegue, no entanto, ter um desempenho ligeiramente superior ao modelo C indicando, o contributo positivo da derivação das novas variáveis na previsão dos congestionamentos. Estas previsões foram integradas no mapa da cidade permitindo assim a realização de uma análise preliminar do estado do tráfego na cidade de Lisboa, validando também a possibilidade de incorporar este mapa num dos *dashboards* da PGIL. Desta forma, consegue-se observar e monitorizar de uma maneira mais simples e directa o tráfego na cidade facilitando a tomada de decisões por parte dos utilizadores que objectivam a construção e realização de um plano de mobilidade mais sustentável.

Como trabalho futuro, pretende-se desenvolver modelos específicos para os seus principais eixos rodoviários da cidade. Vão ser exploradas mais variáveis derivadas, como as médias móveis, para a estimação do congestionamento a curto-prazo (e.g. 1 h). Em termos de visualização analítica, estão em estudo diferentes formas de comunicar o congestionamento ao analista, tornando mais rápida e fácil a identificação de potenciais focos de congestionamento.

Referências

1. CML: Lxdatalab. online: `https://lisboainteligente.cm-lisboa.pt/lxdatalab`

2. CML: Lxdatalab — criação de um indicador de tráfego geral. online: https://lisboainteligente.cm-lisboa.pt/lxdatalab/desafios/criacao-indicador-de-trafego-geral-e-indicadores-para-cada-uma-das-principais-vias-de-entrada-na-cidade
3. Din, I.U., Guizani, M., Rodrigues, J.J., Hassan, S., Korotaev, V.V.: Machine learning in the internet of things: Designed techniques for smart cities. Future Generation Computer Systems **100**, 826–843 (2019). https://doi.org/https://doi.org/10.1016/j.future.2019.04.017, https://www.sciencedirect.com/science/article/pii/S0167739X19304030
4. Dinis, D., Ângelo Palos Teixeira, Barbosa-Póvoa, A.: Foresim-bi: A predictive analytics decision support tool for capacity planning. Decision Support Systems **131**, 113266 (2020). https://doi.org/https://doi.org/10.1016/j.dss.2020.113266, https://www.sciencedirect.com/science/article/pii/S016792362030021X
5. Guillen, P., Komac, U.: Cities Are More Important Than Ever, pp. 7–9. Springer Singapore, Singapore (2020). https://doi.org/10.1007/978-981-15-5741-5_3, \url{https://doi.org/10.1007/978-981-15-5741-5_3}
6. H2O: H2o.ai — open-source machine learning platform for the enterprise. online: https://www.h2o.ai/
7. H2o.ai: XGBoost — optimized distributed gradient boosting. online: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/xgboost.html
8. Homem, P.: Parque automóvel de portugal supera os 6,2 milhões de veículos. online: https://www.dn.pt/dinheiro/nunca-houve-tantos-carros-e-tao-envelhecidos-a-circular-em-portugal-11248374.html (9 2019)
9. ISO: Norma iso 3166-1. online: https://www.iso.org/iso-3166-country-codes.html
10. Metz, D.: Tackling urban traffic congestion: The experience of london, stockholm and singapore. Case Studies on Transport Policy **6**(4), 494–498 (2018)
11. Nadrian, H., Mahmoodi, H., Taghdisi, M.H., Aghemiri, M., Babazadeh, T., Ansari, B., Fathipour, A.: Public health impacts of urban traffic jam in sanandaj, iran: A case study with mixed-method design. Journal of Transport & Health **19**, 100923 (2020), \url{https://www.sciencedirect.com/science/article/pii/S2214140520301274}
12. Ni, M., He, Q., Gao, J.: Using social media to predict traffic flow under special event conditions. In: The 93rd annual meeting of transportation research board (2014)
13. Paiva, K.M., Cardoso, M.R.A., Zannin, P.H.T.: Exposure to road traffic noise: Annoyance, perception and associated factors among brazil's adult population. Science of The Total Environment **650**, 978–986 (2019), \url{https://www.sciencedirect.com/science/article/pii/S0048969718334594}
14. Serrador, A., Tremoceiro, J., Cota, N., Cruz, N., Datia, N.: iLX - A Success Case in Public Tender Methodology. In: ProjMAN 2018 - International Conference on Project MANagement (2018)
15. Taborda, R., Datia, N., Pato, M., Pires, J.M.: Exploring air quality using a multiple spatial resolution dashboard—a case study in lisbon. In: 2020 24th International Conference Information Visualisation (IV). pp. 140–145. IEEE (2020)
16. Uber: Kepler — data agnostic, webgl empowered, high-performance web application for geospatial analytic visualization. online: https://kepler.gl/
17. Zhang, D., Kabuka, M.R.: Combining weather condition data to predict traffic flow: a gru-based deep learning approach. IET Intelligent Transport Systems **12**(7), 578–585 (2018)