INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Área Departamental de Engenharia de Electrónica e Telecomunicações e de Computadores

Indicador de tráfego: descoberta de padrões na cidade de Lisboa

João Pedro Tavares Vaz

(Licenciado)

Dissertação para obtenção do Grau de Mestre
em Engenharia Informática e de Computadores

Orientadores : Doutor Nuno Miguel Soares Datia
Doutora Matilde Pós-de-Mina Pato

Júri:

Presidente: Doutor Tiago Miguel Braga da Silva Dias

Vogais: Doutor João Moura Pires
Doutora Matilde Pós-de-Mina Pato

Fevereiro, 2022

INSTITUTO SUPERIOR DE ENGENHARIA DE LISBOA

Área Departamental de Engenharia de Electrónica e Telecomunicações e de Computadores

Indicador de tráfego: descoberta de padrões na cidade de Lisboa

João Pedro Tavares Vaz

(Licenciado)

Dissertação para obtenção do Grau de Mestre
em Engenharia Informática e de Computadores

Orientadores : Doutor Nuno Miguel Soares Datia
Doutora Matilde Pós-de-Mina Pato

Júri:

Presidente: Doutor Tiago Miguel Braga da Silva Dias

Vogais: Doutor João Moura Pires
Doutora Matilde Pós-de-Mina Pato

Fevereiro, 2022

À minha família que sempre me acompanhou.

Agradecimentos

Em primeiro lugar, gostaria de deixar o meu mais sincero agradecimento aos meus orientadores, Matilde Pato e Nuno Datia, por toda a motivação e apoio que sempre me ofereceram, contribuindo não só de forma essencial para a concretização deste trabalho como também para o meu crescimento pessoal.

Ao Instituto Superior de Engenharia de Lisboa (ISEL) por todas as oportunidades que me proporcionou, nomeadamente, a participação numa bolsa de investigação e a todos os professores e colegas com quem me cruzei ao longo destes últimos anos.

A todos os meus amigos que direta ou indiretamente contribuiram para me manter motivado, especialmente ao Lugy, Soraia, Diogo e Cláudia.

A toda a minha família, em especial aos meus pais, à Junior, aos meus tios, aos meus avós, à minha namorada e ao Tom, o gato!

Obrigado.

Resumo

O parque automóvel circulante em Portugal tem tido um crescimento constante, quer em número de veículos, quer na idade média dos veículos. Os congestionamentos de trânsito, com particular incidência nos centros urbanos, como a cidade de Lisboa, têm impactos negativos a diferentes escalas, quer no dia-a-dia, quer a longo prazo nos cidadãos. Provocando problemas de saúde, problemas económicos e sociais, assim como ambientais. É possível obter dados de tráfego, com baixa latência, com recurso a diferentes formas de sensorização. A partir destes e recorrendo a algoritmos de aprendizagem automática é possível estudar, compreender e prever fluxos de tráfego em zonas de interesse nos centros urbanos. Este trabalho propõem o uso de modelos preditivos para encontrar os indicadores de tráfego, e numa interface gráfica visualizá-los em contexto espácio-temporal, para diferentes momentos e pontos de interesse na cidade de Lisboa. Os resultados obtidos mostraram que através da utilização do algoritmo XGBoost, usando técnicas adequadas na geração e tratamento de características, é possível prever o tempo de atraso causado por um congestionamento com um erro a variar, aproximadamente, os 3 e os 5 minutos. Foi possível observar que a fusão de dados de tráfego e de meteorologia teve um impacto positivo na qualidade do modelo. Estes modelos podem ser integrados com a Plataforma de Gestão Integrada de Lisboa (PGIL) assim como dar origem a um *dashboard* interativo onde se observam os indicadores de tráfego, reais e previstos, num mapa da cidade, contribuindo assim para as tomadas de decisão relativas à mobilidade. São, assim, uma ferramenta que permite antecipar futuros congestionamentos, melhorar o planeamento e gestão urbana para que seja possível reduzir os congestionamentos e mitigar os seus consequentes impactos.

Palavras-chave: Sistemas Inteligentes de Monitorização, Indicadores de Fluidez de Tráfego, Visualização Interativa, Análise Preditiva, Dados Espácio-temporais

Abstract

In Portugal, the number of automobiles on the road and their average lifespan have been rising constantly. In urban centers, like Lisbon, traffic jams have a negative impact on citizens in different ways, whether on a day-to-day basis or over the long-term. They result in problems for health, economics, environmental, and social. A variety of sensorization methods can be used in order to get low-latency traffic data. Then, using automated machine learning algorithms, one can study, understand, and predict traffic flow in areas of common interest in urban centers. This project proposes the use of predictive models to create traffic indicators and visualize them on a graphic interface in a spatio-temporal context, in different points in Lisbon. It's possible to predict the delay time caused by a jam with a variant error of 3 to 5 minutes through the algorithm XGBoost, using the appropriate procedures in generating and processing characteristics. It was evident that the merging of the traffic data and meteorological data had a positive impact on the models. Models like these can be integrated with Plataforma de Gestão Integrada de Lisboa (PGIL), as they can produce an interactive *dashboard* where the real and predicted traffic indicators can be observed on a map of the city, which helps with making mobility decisions. Such *dashboards* are tools that potenciate the usage of the available data that, mixed with machine learning, provides quality insights of the city, helping the decision makers to mitigate the jams impact.

Keywords: Intelligent Monitorization Systems, Traffic Flow Indicators, Interactive Visualization, Predictive Analytics, Spatiotemporal Data

Índice

Lista de Figuras

Lista de Tabelas

Lista de Listagens

Lista de Abreviaturas e Siglas

ARM *Autoregressive Model*. 6

CML Câmara Municipal de Lisboa. 2, 5, 26

DGRNN *Deep GRU Recurrent Neural Network*. 7

EMEL Empresa Municipal de Mobilidade e Estacionamento de Lisboa. 18

GRU Gated Recurrent Unit. 11

IPMA Instituto Português do Mar e da Atmosfera. 2, 3

KNN K-*Nearest Neighbor*. 6

KPI *Key Performance Indicator*. 6, 12

LSTM *Long Short-Term Memory*. 9

MAE Erro Absoluto Médio (do inglês, *Mean Absolute Error*). 7, 39

MAPE Erro médio percentual absoluto (do inglês, *Mean Absolute Percentage Error*). 6, 7, 8

MSE Erro quadrático médio (do inglês, *Mean Squared Error*). 7

NN *Neural Networks*. 6

NOAA *National Oceanic and Atmospheric Administration*. 7

OMS Organização Mundial de Saúde. 1

PM Partículas finas. 1

RF *Random Forest*. 9

RMSE Raiz do erro quadrático médio (do inglês, *Root Mean Squared Error*). 6, 7, 8, 39

RMSLE Raiz do erro médio quadrático e logarítmico (do inglês, *Root Mean Squared Logarithmic Error*). 39

SVR *Support Vector Regression*. 6, 7, 9

1

Introdução

O parque automóvel circulante em Portugal está em constante crescimento, ultrapassando os $6,2$ milhões de veículos nos últimos anos [30]. Além disso apresenta ainda um constante envelhecimento, sendo que desde o ano de 2000 até à atualidade, a idade média dos veículos passou dos $7,2$ anos para os $12,7$ anos [6]. Sendo os veículos mais antigos também os mais poluentes, estamos perante um cenário que se tende a agravar a cada ano que passa [29].

Atualmente observa-se uma intensificação do tráfego de veículos nos grandes centros urbanos e, por consequência, um aumento dos congestionamentos. Estes por sua vez apresentam impactos bastante negativos [22, 25], destacando-se: (i) a nível económico, através do aumento do tempo despendido em deslocações e os seus possíveis custos como por exemplo, maiores consumos de combustível; (ii) a nível social, através do aumento dos níveis de *stress* e ansiedade o que por sua vez pode levar a comportamentos agressivos; e (iii) ao nível da saúde, principalmente com impactos induzidos pelo aumento da poluição sonora e do ar.

Estima-se, com base em médias globais, que cerca de 25% das Partículas finas (PM) presentes em áreas urbanas sejam provenientes do tráfego de veículos [17]. A Organização Mundial de Saúde (OMS) afirma que estas partículas são responsáveis pela morte de 7 milhões de pessoas por ano [28], pelo que a qualidade de vida da população é direta (e, indiretamente) afetada não só pela exposição a este tipo de partículas que se confirma serem nocivas para a saúde, mas também pelo tempo perdido em deslocações, tempo este que poderia ser aproveitado visando alguns retornos a nível

social.

Deste modo o congestionamento no tráfego apresenta-se como um problema em larga escala, principalmente nas áreas mais urbanas, tornando-se cada vez mais importante estudar, conhecer e tentar prever o fluxo de tráfego em determinadas vias. No entanto, a obtenção desse conhecimento não é facilmente alcançável devido à complexidade associada ao desenvolvimento de modelos que consigam prever a evolução dos congestionamentos sendo necessário observar dados existentes utilizando técnicas de aprendizagem automática, não só para identificação de pontos críticos de forma automática, mas também para a obtenção de modelos preditivos da realidade a modelar abrindo assim caminho para soluções de aplicações que visam mitigar os congestionamentos.

Estas técnicas possibilitam previsões com níveis aceitáveis de precisão, confiança, fiabilidade e sensibilidade que desempenham um papel fundamental na resolução de problemas atuais, influenciando a vida das pessoas em qualquer lugar do mundo e em múltiplos contextos revelando-se como um aspecto fundamental na construção de uma cidade inteligente. Antevisões de uma determinada situação permitem estimar várias variáveis e consequentemente formular planos para atingir as metas programadas de avanço do 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, permitindo antever dificuldades e problemas, antecipar soluções, e consequentemente reduzir custos principalmente a nível temporal refletindo-se estes também economicamente.

É perante este cenário que surge um enorme interesse em analisar, monitorizar e, sobretudo, prever permanentemente e com confiança o tráfego não só nas principais vias de entrada e saída dos grandes centros urbanos, mas também nas suas principais avenidas e locais de maior susceptibilidade a este tipo de condicionamentos. Deve atribuir-se particular atenção aos movimentos pendulares, ou seja, às deslocações entre locais de residência habitual e locais de trabalho ou estudo sendo estes responsáveis por um crescimento diário superior a 70% de população na cidade de Lisboa [21].

Ciente disto, a Câmara Municipal de Lisboa (CML), através do Laboratório de Dados Urbanos de Lisboa propôs o desafio intitulado "Criação de indicador de tráfego geral e indicadores para cada uma das principais vias de entrada na cidade" [5], ao qual pretendemos dar resposta nesta dissertação.

Assim sendo, este trabalho tem definidos como objetivos:

O1: Desenvolver um modelo preditivo capaz de prever congestionamentos no tráfego da cidade de Lisboa utilizando dados fornecidos pelo Waze e pelo Instituto Português

do Mar e da Atmosfera (IPMA), com enfoque nas principais vias de acesso à cidade;

O2: Desenvolver um indicador de fluidez de tráfego capaz de resumir para uma dada área geográfica e num determinado espaço temporal a fluidez do tráfego;

O3: Propor uma representação visual do tráfego na cidade através de um *dashboard* onde se incluem as previsões para o indicador desenvolvido num determinado espaço temporal.

1.1 Organização do documento

Este documento encontra-se estruturado através de seis capítulos. O presente capítulo apresenta a descrição do problema, a sua importância, motivação e objetivos definidos. No Capítulo 2 é analisado o estado da arte relativo ao problema com principal foco nos conjuntos de dados, algoritmos de mineração, indicadores de fluidez de tráfego e técnicas de visualização de dados. No Capítulo 3 é apresentado o sistema proposto para a resolução do problema, detalhando todos os seus componentes desde as fontes de dados à extração de conhecimento e representação gráfica. No Capítulo 4 é apresentado o desenvolvimento e análise de desempenho dos vários modelos preditivos construídos, assim como a decisão da escolha do modelo a utilizar segundo as medidas de avaliação obtidas. O desenvolvimento de indicadores de fluidez de tráfego, a representação de um *dashboard* interativo e validação dos indicadores desenvolvidos através de uma análise com utilizadores encontra-se no Capítulo 5. A tese termina com o Capítulo 6, onde são apresentadas as conclusões do trabalho desenvolvido ao longo da dissertação e o que poderá vir a ser realizado futuramente.

2

Trabalho Relacionado

Para a realização deste trabalho, foi realizada uma pesquisa e consequente estudo de diversos artigos, teses e dissertações relacionadas com o problema de modo a ser possível desenvolver uma solução mais adequada e inovadora. Foram analisados diferentes conjuntos de dados, nomeadamente de meteorologia, de trânsito[1], da plataforma de dados da CML. Outro ponto relevante foi a pesquisa de indicadores (ou métricas) de fluidez de trânsito, algoritmos que objetivam a construção de modelos de previsão e medidas de avaliação para comparação dos respetivos modelos. A complexidade e variabilidade dos dados para o estudo do tráfego tem um impacto relevante nos resultados de previsão, daí a construção de um modelo de dados robusto requerer um estudo cuidado e comparativo com trabalhos relacionados (Secção 2.1). Estes trabalhos irão ser analisados e descritos neste capítulo.

Pretende-se que a solução seja facilmente analisada e tenha uma representação visual cuidada, simples e direta permitindo ao analista obter informação, através de indicadores, do estado do tráfego em diversos pontos da cidade de modo a auxiliar eventuais tomadas de decisão. Para alcançar este objetivo foram estudadas diversas aplicações e ferramentas até à escolha daquela que se mostrou ser a mais adequada. As aplicações e ferramentas estudadas serão ilustradas neste capítulo.

[1] Aplicação Waze https://www.waze.com/

2.1 Conjuntos de dados

Foram analisados vários trabalhos, como Pirra and Diana [27] que explora o congestionamento de tráfego, neste caso, para a cidade Italiana de Turin utilizando dois conjuntos de dados recolhidos em Maio de 2017. O primeiro contém informação relacionada com o fluxo de trânsito nas principais vias da cidade com uma hora de frequência entre amostras. Contém a data e hora, o fluxo médio de tráfego em número de veículos por hora, duração de viagem em segundos e a fonte de dados como suas características para determinada via. Estes dados são gerados através de estimativas com base em modelos de fluxo de tráfego implementados pela empresa local 5T [1] e foram fornecidos pela Câmera Municipal de Turin não estando disponíveis métricas de avaliação dos modelos assim como outras informações adicionais. O segundo conjunto de dados contém informação de rotas GPS em 28 veículos que realizavam entregas na cidade apresentando assim as posições de latitude e longitude, data e hora, velocidade média do veículo e direção do trajeto como suas características. Após a combinação dos conjuntos de dados foi definido um indicador chave de desempenho, *Key Performance Indicator* (KPI), descrito posteriormente em 2.2.

Ni et al. [23] descreve a previsão do fluxo de tráfego perante eventos que atraem massas na Califórnia. A análise de previsão de tráfego, em determinadas vias, utiliza o fluxo de tráfego por hora detetado por sistemas de vigilância e dados provenientes da rede social Twitter através de *tweets*. Para a construção deste conjunto de dados foi necessário escolher (a) um conjunto de palavras adequadas ao evento em questão pelas quais se realizou a pesquisa de *tweets*, e dessa forma obter 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, e deste modo é possível obter um conjunto de dados final mais enriquecido e capaz de gerar modelos de previsão para essas mesmas 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 de indução colocados nas

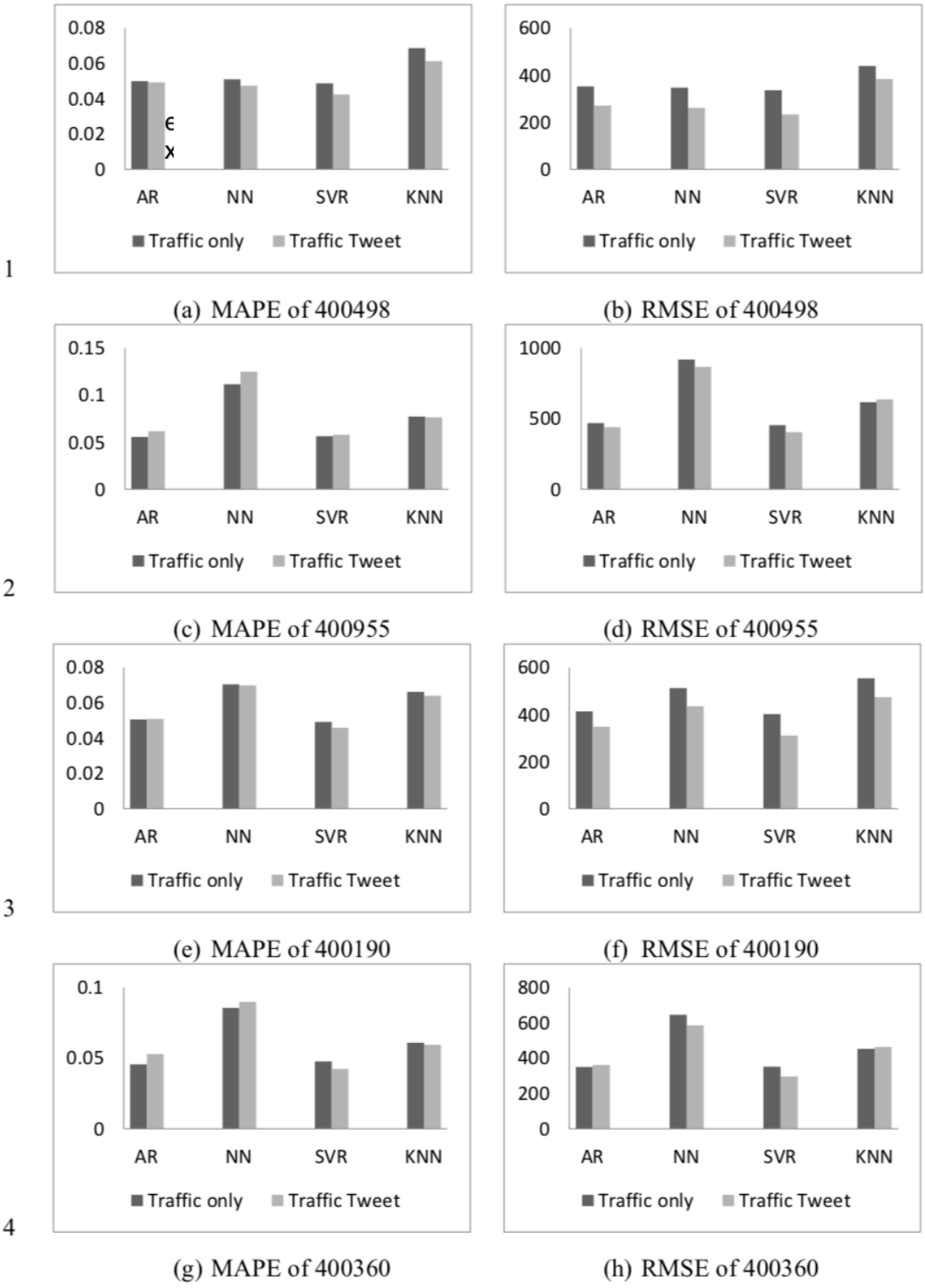
principais vias de acesso a dois espaços destinados a atrações públicas (*Oracle Arena* e *O.co Coliseum*) estão ilustrados na Figura 2.1. Estes permitem inferir que o modelo gerado através da técnica SVR apresenta os melhores resultados entre as quatro técnicas expressando maioritariamente os menores valores quer de MAPE, quer de RMSE, independentemente do conjunto de dados utilizado. Valores estes que estão próximos de $0,04$ no caso do MAPE do modelo gerado com dados do sensor 400498 e um valor próximo de 300 no caso do RMSE do modelo gerado com dados do sensor 400460. Também se conclui que os valores de MAPE não apresentam resultados tão consistentes como os valores de RMSE, em que se observa que o modelo construído através do conjunto de dados que utiliza *tweets* apresenta sempre melhor desempenho em relação ao modelo construído apenas com conjunto de dados de tráfego.

Zhang and Kabuka [39] 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 39 000 sensores espalhados pelas principais áreas metropolitanas de todo o estado da Califórnia, obtendo-se assim uma característica que representa o número de veículos que passaram pelo sensor durante um determinado intervalo de tempo e espaço. Estes dados foram combinados com dados meteorológicos provenientes do repositório da *National Oceanic and Atmospheric Administration* (NOAA), que são reportados de hora em hora, obtendo-se então características como a precipitação e temperatura máxima e mínima, para além das habituais características de tráfego. Tendo em conta que estes dados serão aplicados a uma rede neuronal, foi realizada a normalização tanto das características de tráfego como meteorológicas, de acordo com a normalização *Min-Max* apresentada na equação 2.1,

$$S = \alpha \times \frac{x - min(x)}{[max(x) - min(x)]} \tag{2.1}$$

onde S representa o valor da característica normalizada num intervalo de [0,α] sendo x o valor inicial da característica.

O conjunto de dados foi dividido em conjunto de treino e conjunto de teste contendo 90% e 10% do volume dos dados, 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, apresentando um MAE, MSE e RMSE de $7,9 \times 10^{-3}$, $3,76 \times 10^{-5}$ e $1,9 \times 10^{-3}$,



(a) MAPE of 400498 (b) RMSE of 400498

(c) MAPE of 400955 (d) RMSE of 400955

(e) MAPE of 400190 (f) RMSE of 400190

(g) MAPE of 400360 (h) RMSE of 400360

Figura 2.1: Medidas de avaliação MAPE e RMSE dos modelos gerados com dados reportados pelos sensores 400498, 400955, 400190 e 400360, respectivamente. Legenda: AR - Autoregressive Model, NN - Neural Networks, SVR - Support Vector Regression e KNN - K-Nearest Neighbor (Retirado de [23])

respectivamente. 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), entre outras podem ser vistas na Figura 2.2. Conclui-se que o modelo proposto apresenta sempre os melhores resultados, em comparação com as restantes técnicas, para as três métricas de avaliação utilizadas.

Model	Evaluation criteria	Error rate
two hidden layer simple RNN (50 hidden units for each layer)	MSE	0.0014
	MAE	0.024
	RMSE	0.037
two hidden layer DGRNN (50 hidden units for each layer)	MSE	7.45×10^{-4}
	MAE	0.019
	RMSE	0.027
two hidden layer DGRNN (500 hidden units for each layer)	**MSE**	$\mathbf{3.76 \times 10^{-5}}$
	MAE	**0.0079**
	RMSE	**0.0019**
two hidden layer LSTM (500 hidden units for each layer)	MSE	8.72×10^{-5}
	MAE	0.0090
	RMSE	0.0093
three hidden layer DGRNN (500 hidden units for each layer)	MSE	0.00044
	MAE	0.0082
	RMSE	0.02
ARIMA (0,1,1)	MSE	0.007
	MAE	0.058
	RMSE	0.083
support vector regression	MSE	0.003
	MAE	0.048
	RMSE	0.058
random forest regression	MSE	0.002
	MAE	0.029
	RMSE	0.042

Figura 2.2: Comparação de resultados após construção de modelos aplicando diversas técnicas. (Retirado de [39])

Konior et al. [19] monitoriza o impacto do tráfego rodoviário em ambiente urbano na qualidade do ar e nos níveis de ruído. Tiram partido do sistema já existente, *OnDynamic*, que monitoriza parâmetros de tráfego rodoviário através da deteção de dispositivos *bluetooth* dos condutores ao passarem por estações base, estando estas devidamente equipadas com sensores e próximas das faixas de rodagem. A passagem de um dispositivo entre duas destas estações permite a obtenção da informação da velocidade média do veículo, do tempo gasto em congestionamentos e do tempo total da viagem. A estas estações foram ainda adicionados sensores de poluição e ruído permitindo assim recolher informação da concentração de poluentes no ar e níveis de ruído, respectivamente. Deste modo, os dados são recolhidos e devidamente processados com o objetivo final de serem apresentados ao utilizador através de métodos descritos posteriormente na Secção 2.3.

2.2 Indicadores de fluidez de trânsito e algoritmos de mineração

Dados da PORDATA[2] indicam que em 2019 a Área Metropolitana de Lisboa tinha um número médio de indivíduos por Km^2 de $946,8$ e sendo a cidade de Lisboa um local propício ao tráfego, Brito [3] apresenta o fluxo de tráfego nas principais vias de entrada da cidade com base numa análise efetuada na hora de ponta da manhã ao longo de todo o ano de 2004 onde 70% do tráfego é escoado pelos corredores de Cascais, Sintra/Amadora, Amadora/Loures, A1 (entrada norte) e Ponte 25 de Abril. O autor apresenta também um indicador que calcula o tráfego médio anual de uma estrada com base na duração das contagens efetuadas resultando assim num valor que representa o número de veículos por dia por sentido. Sendo o modo de obtenção do volume de tráfego contínuo realiza-se a média da soma dos valores do tráfego médio diário mensal correspondente a cada um dos meses do ano, podendo a equação sofrer adaptações conforme o número de meses disponíveis. É também possível estimar, embora de forma menos precisa, o volume de tráfego procedente de contagens de curta duração através da utilização de fatores de ajustamento consoante a duração das contagens.

Pirra and Diana [27] explora precisamente o desenvolvimento de um indicador de congestionamento de tráfego, que indica o tempo gasto no congestionamento de determinada via de acordo com a equação 2.2,

$$KPI = T_O - (T_{GPS} - T_{GPSss}) \tag{2.2}$$

sendo T_O o valor de duração de viagem aquando a via está totalmente livre, T_{GPS} a duração de viagem em que um veículo demorou a passar a via e T_{GPSss} a duração em que o veículo esteve totalmente parado na via. No entanto, caso se pretenda comparar os resultados obtidos para vias com diferentes comprimentos é necessário existir uma escala e uma proporcionalidade comum no valor do indicador. Para tal, o indicador também pode ser calculado em termos relativos de acordo com a equação 2.3.

$$RKPI = \frac{T_O - (T_{GPS} - T_{GPSss})}{T_O} \tag{2.3}$$

Estes indicadores podem apresentar três tipos de valores: (a) valores negativos, que correspondem ao tempo gasto no congestionamento de determinada via; (b) valor zero, que indica que o tempo de viagem foi exatamente o mesmo aquando a via está

[2]Ver: https://www.pordata.pt/Municipios/Densidade+populacional-452 PORDATA, Densidade populacional

totalmente livre; e (c) valores positivos que indicam também que a via está totalmente livre sendo que estes surgem devido ao facto dos valores referência de T_O corresponderem a uma média e existir a possibilidade dos veículos viajarem a uma velocidade superior à estimada.

Zhang and Kabuka [39] aplicam uma rede neuronal recorrente com a implementação do algoritmo de retropropagação de modo a aumentar a precisão das previsões, mas sendo que o input desta rede consiste numa sequência de vetores com as características do conjunto de dados final o algoritmo de retropropagação apresenta alguma dificuldade em treinar a sequência com a sua dependência de longo prazo devido, nomeadamente, à dissipação ou explosão do gradiente. Face a estas dificuldades foi introduzido uma Gated Recurrent Unit (GRU) que por sua vez é baseada numa long short-term memory (LSTM) com a diferença de que a GRU não apresenta células de memoria separadas proporcionando assim um maior poder computacional no treino dos dados. A GRU através do seu mecanismo que aceita a passagem de uma certa quantidade de informação, aquando da retropropagação, permite resolver os problemas relacionados com o gradiente.

A proposta dos autores consiste então na substituição dos neurónios da camada escondida por unidades GRU permitindo assim descobrir a correlação nos dados ao longo das series temporais tanto a curto como a longo prazo. O modelo proposto intitula-se de Deep GRU Recurrent Neural Network e apresenta-se na Figura 2.3.

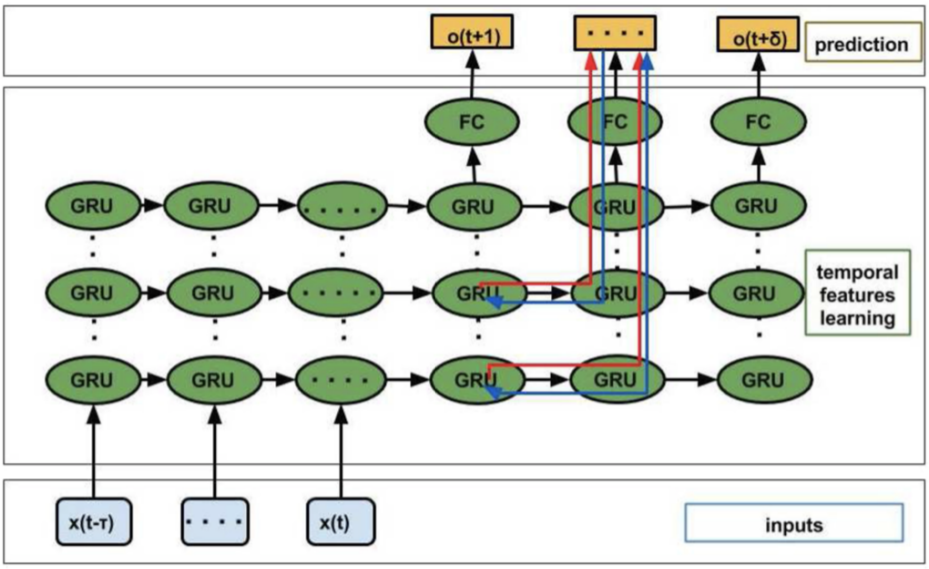


Figura 2.3: Modelo Deep GRU Recurrent Neural Network (DGRNN). (Retirado de [39])

Como mencionado em 2.1, os melhores resultados foram obtidos através da utilização

deste modelo chegando mesmo a atingir uma melhoria de 25% na precisão das previsões aquando da utilização sequencial dos dados de condições meteorológicas como entradas da rede, isto em comparação com os modelos onde não foram considerados os dados relativos a condições meteorológicas.

2.3 Apresentação de dados

No artigo [27] foi desenvolvida uma ferramenta gráfica capaz de integrar o indicador KPI desenvolvido na representação dos mapas da cidade mencionado na Secção 2.1.

Na Figura 2.4 é possível observar uma representação gráfica dos diferentes valores do indicador para diferentes períodos do dia e em diversas vias da cidade. Os valores do indicador são representados por cores correspondendo assim a cinco diferentes níveis de congestionamento sendo o roxo o nível máximo de congestionamento e o cinzento o nível mínimo de congestionamento correspondendo assim a uma via totalmente congestionada e uma via totalmente livre, respectivamente.

É também apresentada a possibilidade de visualizar o indicador tendo em conta um panorama mais geral e abrangente do centro da cidade mostrando o congestionamento médio para cada uma das vias analisadas no mês correspondente aos dados obtidos e não apenas em pequenos períodos de tempo específicos permitindo assim uma diferente análise temporal.

Konior et al. [19] implementaram um módulo para a visualização dos dados recolhidos, mencionados na Secção 2.1), sob a forma de plataforma através da utilização da biblioteca *Leaflet* [20] onde se inclui a apresentação de dados integrados num mapa assim como a apresentação de dados através de gráficos .

Ao nível dos dados integrados no mapa é possível visualizar diretamente sobre os troços de vias a contagem do número de dispositivos *bluetooth* que neles passaram nos últimos 15 minutos. É através de um esquema de cores (cinzento, verde, amarelo, laranja e vermelho) que esta informação é observada, não existindo informação de qual o método utilizado para associar os intervalos dos valores das contagens com a respectiva cor.

Ao interagir com o mapa, carregando num troço de via, é possível obter informação sobre os restantes parâmetros quer os que estão relacionados com o tráfego (velocidade média dos veículos, tempo médio gasto em congestionamentos e tempo médio total de viagens), quer os que estão relacionados com a poluição (níveis da concentração de poluentes no ar e ruído). É também possível visualizar um pequeno gráfico sobre o

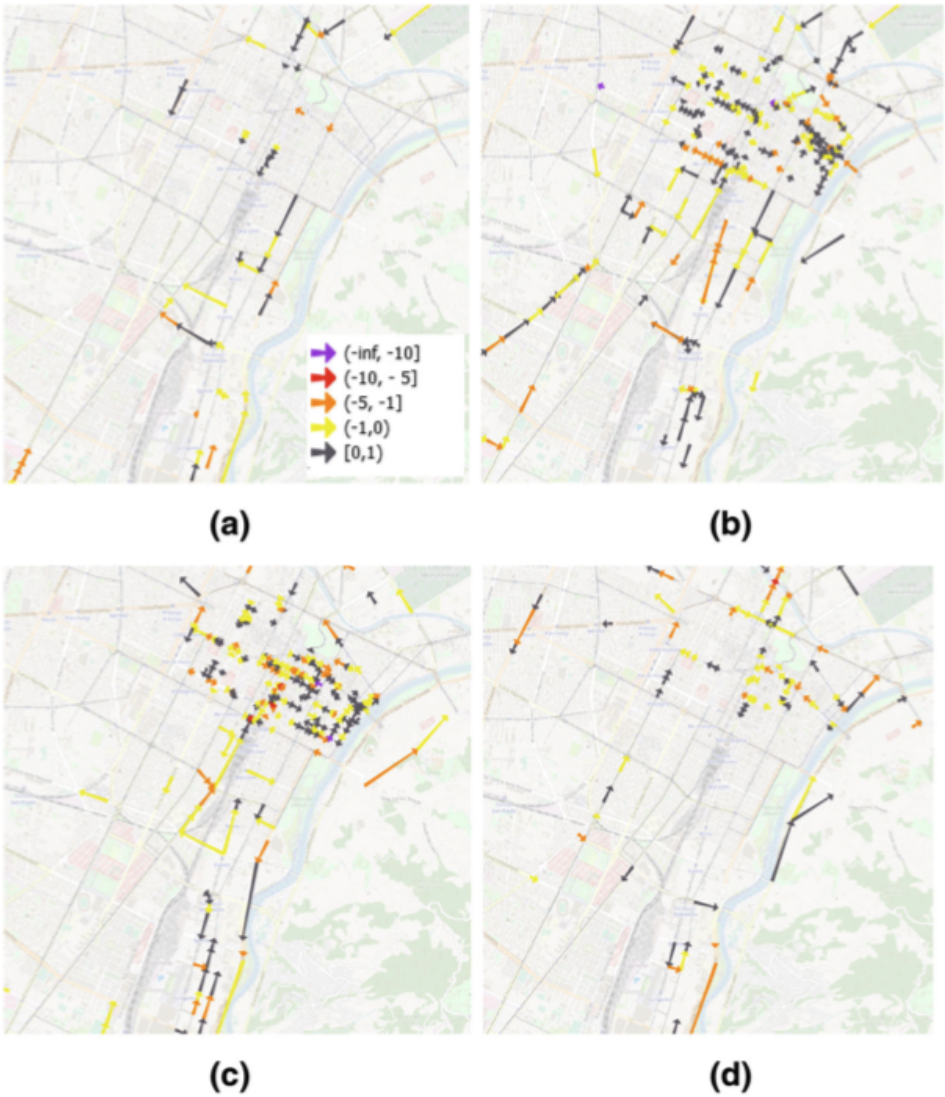


Figura 2.4: Visualização dos valores do indicador de tempo gasto em congestionamento de tráfego para diversas vias da cidade de Turin nos períodos de tempo 8:00–8:59 (a), 9:00–9:59 (b), 17:00–17:59 (c), 18:00–18:59 (d) (Retirado de [27])

mapa com os dados das últimas 24 horas do parâmetro e troço de via em questão, o que permite observar todas as mudanças num passado recente.

Ao nível dos dados apresentados em gráficos, estes são independentes do mapa, ou seja, são apresentados numa outra página da plataforma e permitem a observação de um parâmetro selecionado individualmente num determinado troço de via através de diferentes janelas temporais com um intervalo mínimo de 1 dia e um intervalo máximo de 1 mês. Desta forma já é possível uma comparação do parâmetro selecionado em determinada via a longo prazo ao contrário do gráfico integrado no mapa, mencionado no parágrafo anterior, o que permite com uma maior facilidade e segurança detetar mudanças nos dados podendo eventualmente gerar algum tipo de alerta.

Picozzi et al. [26] desenvolveram um protótipo de uma aplicação web onde integram algumas técnicas de visualização de dados de modo a permitir uma visão geral de dados relativos ao tráfego na cidade de Oulu permitindo às várias entidades interessadas a capacidade de analisar e explorar eventos relacionados com o fluxo de tráfego de modo a planear atividades de forma mais consciente na cidade.

Os dados utilizados nos testes deste protótipo foram recolhidos em Oulu através de sensores colocados em 77 cruzamentos da cidade, entre os dias 26 de Maio de 2021 e 21 de Junho de 2021. O conjunto de dados apresenta cerca de 600 000 entradas com uma resolução temporal mínima de 1 minuto entre amostras. O objetivo final da utilização destes dados passa por medir o volume de tráfego em termos de número de veículos por espaço de tempo, no entanto, não é disponibilizada qualquer tipo de informação relacionada com as características presentes no conjunto de dados obtido.

A aplicação web desenvolvida segue o mantra de Shneiderman's [32]: "*Overview first, zoom and filter, then details-on-demand*", focando-se inicialmente numa visão geral do tráfego na cidade e mediante o interesse e/ou necessidade permite observar zonas ou cruzamentos específicos com um maior nível de detalhe. Monitoriza o volume de tráfego através de contagens do número de veículos nos vários cruzamentos da cidade por períodos temporais definidos pelo utilizador que podem ir desde dia/noite até algumas semanas. De modo a obterem níveis de agregação espácio-temporais dinâmicos foram combinadas três técnicas de visualização, estando estas devidamente sincronizadas conforme a janela temporal selecionada pelo utilizador.

Na Figura 2.5 é possível observar as três técnicas de visualização utilizadas, sendo estas: (a) *Chart UI*, através de um gráfico que permite analisar e explorar as variações do volume de tráfego ao longo do tempo, (b) *Map UI*, através de um mapa onde cada cruzamento está representado com um círculo sendo que a sua cor representa a média do volume de tráfego naquele cruzamento para o período de tempo selecionado, e

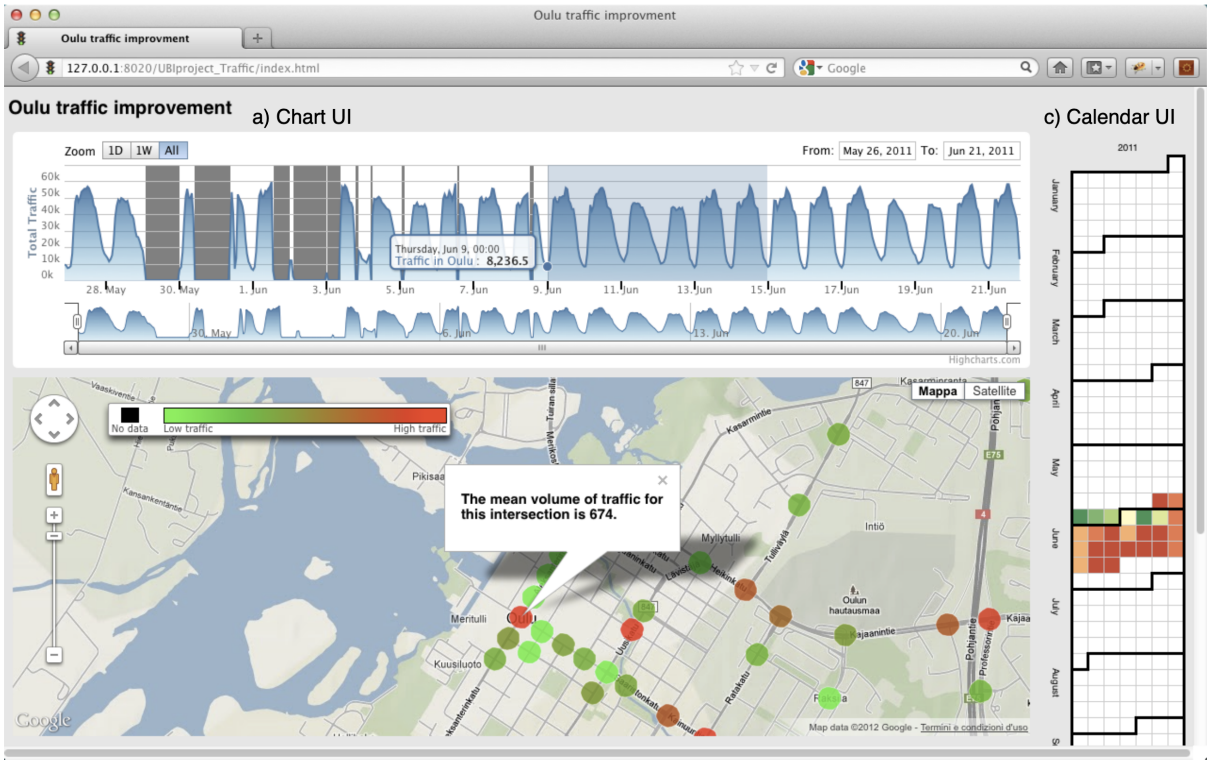


Figura 2.5: Página da aplicação web com as três técnicas de visualização utilizadas (*Chart UI*, *Map UI* e *Calendar UI*) para observação e monitorização do volume de tráfego na cidade de Oulu (Retirado de [26])

(c) *Calendar UI*, através de um calendário onde se ilustra a variação do volume de tráfego médio por dia e segundo um esquema de cores.

Esta solução parece estar bastante adequada ao problema em questão pelo que oferece uma visão geral do volume de tráfego em toda a cidade assim como a possibilidade de aplicar filtros espaciais e temporais ao dados facilitando a análise por parte do utilizador. Esta análise, por sua vez, poderá ser útil no apoio à tomada de decisões ou até mesmo na ajuda à programação de eventos na cidade, por exemplo, através da reprogramação de semáforos em alguns cruzamento em horas especificas do dia.

3

Abordagem ao problema

A proposta de 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.

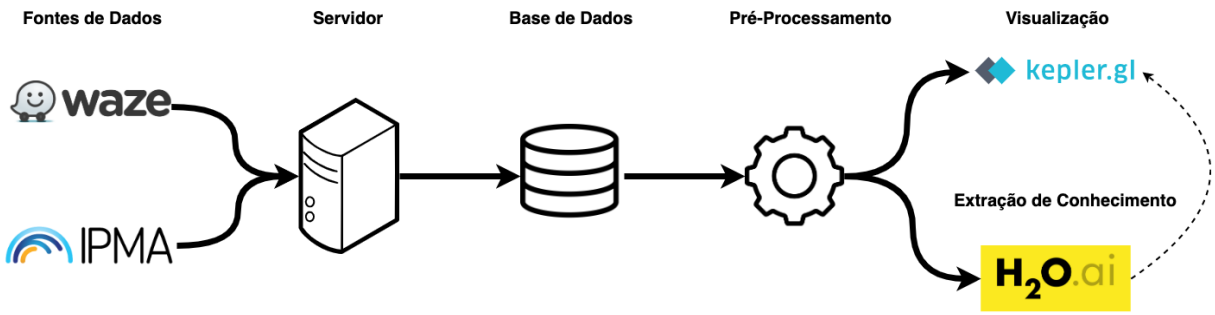


Figura 3.1: Pipeline do sistema

O pipeline do sistema é apresentado na Figura 3.1 e divide-se nos seguintes componentes: (i) Fontes de Dados, onde estão representadas as fontes dos dados usadas para criar o modelo preditivo; (ii) Servidor, que será o responsável pela realização de pedidos HTTP às fontes de dados, assim como do 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.

Nas próximas secções são descritos em detalhe cada um dos componentes.

3.1 Fontes de Dados

A principal fonte de dados utilizada é proveniente da Empresa Municipal de Mobilidade e Estacionamento de Lisboa (EMEL) que fornece dados recolhidos pelo Waze [37], relativos aos congestionamentos do tráfego na cidade de Lisboa através de dois conjuntos de dados, sendo estes denominados de *Jams* e *Irregularities*, respectivamente [9].

O primeiro, *Jams*, consiste exactamente numa lista de congestionamentos de trânsito e

Tabela 3.1: Descrição das características presentes nas propriedades do conjunto de dados *Jams*.

Nome do campo	**Descrição**
country	Código representativo do país segundo a norma ISO 3166-1 [15]
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 (esquerda, direita, frente ou nenhum)
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 via (1 = Rua, 2 = Rua principal, 3 = Auto-estrada, 4= Rampa, 5 = Caminho, 6 = Rua principal, 7 = Rua secundária, 8 = Caminho de terra batida, 9 = Passeio, 10 = Caminho pedestre, 11 = Saída, 14 = Caminho de terra batida, 15 = Travessia de barco, 16 = Escadas, 17 = Caminho privado, 18 = Caminho de ferro, 19 = Faixa exclusiva (taxi/bus), 20 = Via de acesso ou dentro de parque de estacionamento, 21 = Estrada de serviço)
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 (*bounding box*)

é gerado através do processamento interno de localizações GPS enviadas por utilizadores da aplicação, cálculos baseados em velocidades numa determinada via e também pelos relatórios gerados pelos utilizadores quando estes encontram algum congestionamento. As suas características encontram-se presentes na Tabela 3.1.

O segundo, *Irregularities*, consiste também numa lista de congestionamentos, identificados pelo sistema como irregulares em relação ao normal andamento de determinada via com base em dados históricos de velocidade. As suas características encontram-se descritas na Tabela 3.2.

Tabela 3.2: Descrição das características presentes nas propriedades do conjunto de dados *Irregularities*.

Nome do campo	**Descrição**
detection_date	Data da detecção
update_date	Data da última actualização da ocorrência
country	Código representativo do país segundo a norma ISO 3166-1 [15]
city	Nome da cidade ou estado
length	Comprimento da congestionamento irregular em metros
speed	Velocidade média do tráfego em metros por segundo
seconds	Tempo média de tráfego em segundos
trend	Tendência da irregularidade (-1 = Melhorar , 0 = Constante, 1 = Piorar)
type	Tipo de irregularidade (0 = Nenhuma, 1 = Pequena, 2 = Média, 3 = Grande, 4 = Enorme)
delay_seconds	Tempo de atraso do tráfego em comparação com a via completamente livre em segundos
regular_speed	Velocidade média geral para a via em condições normais
street	Nome da rua
end_node	Saída mais próxima do final do congestionamento irregular
start_node	Entrada mais próxima do inicio do congestionamento irregular
jam_level	Código do nível do congestionamento irregular (0 = livre de tráfego, 5 = via bloqueada)
severity	Gravidade do congestionamento irregular (0 = pouco severo, 5 = muito severo)
drivers_count	Número de *Wazers* presentes no congestionamento irregular
alerts_count	Número de alertas recebidos por *Wazers*
bbox	Conjunto de coordenadas que representam o local do congestionamento (*bounding box*)

Tabela 3.3: Descrição das características presentes nas propriedades do conjunto de dados *Closures*.

Nome do campo	**Descrição**
morada	Morada do local da ocorrência
local_referencia	Descrição de possíveis locais de referência
restricao_circulacao	Tipo de restrição de circulação (Estacionamento, Estreitamento de via, Corte total, Cortes temporários, Passeio, Mantém perfil de via)
motivo	Motivo da ocorrência
impacto	Nível de impacto no tráfego (Pouco relevante, Relevante)
periodos_condicionamentos	Datas dos períodos onde se poderá verificar acondicionamentos
creation_date	Data de criação da ocorrência
lastmod_date	Data da última actualização da ocorrência
pedido	Identificador único da ocorrência
bbox	Conjunto de coordenadas que representam o local do congestionamento (*bounding box*)

É ainda disponibilizada pela EMEL informação relativa a bloqueios, condicionamentos e restrições de vias previamente programadas na cidade de Lisboa através de um conjunto de dados denominado de *Closures*. As suas características encontram-se presentas na Tabela 3.3.

Todos estes conjuntos de dados (*Jams*, *Irregularities* e *Closures*) são apresentados no formato JSON [16], sendo que as características relacionadas com as áreas geográficas são apresentadas no formato GeoJSON [11]. Este permite codificar uma variedade de estruturas geográficas, pelo que os dados são sempre reportados com a característica tipo de *FeatureCollection* apresentando assim um conjunto de *Features* que contêm informações relacionadas com cada um dos acontecimentos reportados tais como a sua localização, que é reportada através de um conjunto de coordenadas e todas as suas propriedades.

A segunda fonte de dados utilizada é proveniente do IPMA [14] e disponibiliza informação relativa a dados meteorológicos de todo o país. Estão a ser recolhidos dois conjuntos de dados. O primeiro contém informação observada através de sensores nas diversas estações meteorológicas, espalhadas pelo país, nas últimas 24 horas. O segundo conjunto de dados contém informação relativa à previsão de dados meteorológicos até um máximo de 5 dias para a cidade de Lisboa, sendo estes obtidos automaticamente através do processamento estatístico das previsões dos dois modelos

numéricos (ECMWF [8] e AROME [2]). Estes dois conjuntos de dados estão a ser denominados *Weather* e *PrevWeather* e as suas propriedades estão apresentadas na Tabela 3.4 e na Tabela 3.5, respectivamente.

3.2 Servidor

Para garantir a recolha contínua e automática dos dados anteriormente apresentados, foi desenvolvido um sistema recorrendo ao Node-RED [24], que é uma ferramenta de desenvolvimento baseada em fluxos para programação visual que permite fazer ligações entre dispositivos de hardware, APIs e outros serviços *online*. A utilização desta ferramenta deve-se à forma simplificada com que permite criar e alterar fluxos programáticos, à sua rápida instalação e, sobretudo, à quantidade reduzida de recursos que necessita permitindo com que o servidor esteja em funcionamento constante numa máquina virtual.

O sistema encontra-se em funcionamento contínuo, realizando principalmente pedidos HTTP, sendo que todos os pedidos relacionados com o tráfego estão a ser realizados de 5 em 5 minutos. Esta periodicidade está relacionada com o facto destes dados

Tabela 3.4: Descrição das características presentes no conjunto de dados *Weather*.

Nome do campo	**Descrição**
YYYY-mm-ddThh:mi	Data e hora da observação segundo a norma ISO 8601
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

Tabela 3.5: Descrição das características presentes no conjunto de dados *PrevWeather*.

Nome do campo	**Descrição**
forecastDate	Data da previsão
idWeatherType	Id do código relativo ao tipo de tempo [a]
tMin	Temperatura mínima diária em graus centígrados
tMax	Temperatura máxima diária em graus centígrados
classWindSpeed	Id da classe da intensidade do vento [b]
predWindDir	Rumo predominante do vento (N, NE, E, SE, S, SW, W, NW)
probPrecipita	Probabilidade da precipitação
classPrecInt	ID da classe da intensidade da precipitação [c]

[a] https://api.ipma.pt/open-data/weather-type-classe.json
[b] https://api.ipma.pt/open-data/wind-speed-daily-classe.json
[c] https://api.ipma.pt/open-data/precipitation-classe.json

serem o principal objeto de estudo do trabalho, mas também pela frequência das alterações, por vezes, bastante altas. Todos os pedidos relacionados com a meteorologia estão a ser realizados a cada hora, sendo que para este tipo de dados a sua frequência de alterações não justifica um menor intervalo de tempo.

Como pode ser observado na Figura 3.2, o principal fluxo desenvolvido não só realiza os vários pedidos HTTP como também guarda os dados provenientes desses pedidos numa base de dados SQLite [33].

É também realizado um pequeno pré-processamento que consiste numa filtragem dos dados provenientes do conjunto de dados *Weather*, dado que estes contêm informação de todas as estações meteorológicas do país e apenas são pretendidas informações relativas à cidade de Lisboa. Como tal, apenas estão as ser retidas informações de 3 estações meteorológicas da cidade sendo estas a Lisboa (Geofísico) , Lisboa (G.Coutinho) e Lisboa (Tapada da Ajuda).

Para além do principal fluxo, a implementação do sistema conta ainda com dois fluxos secundários que têm como objetivo automatizar toda a gestão ao nível da base de dados e ao nível do sistema de ficheiros local nomeadamente a criação e organização de directórios onde os dados irão ser armazenados.

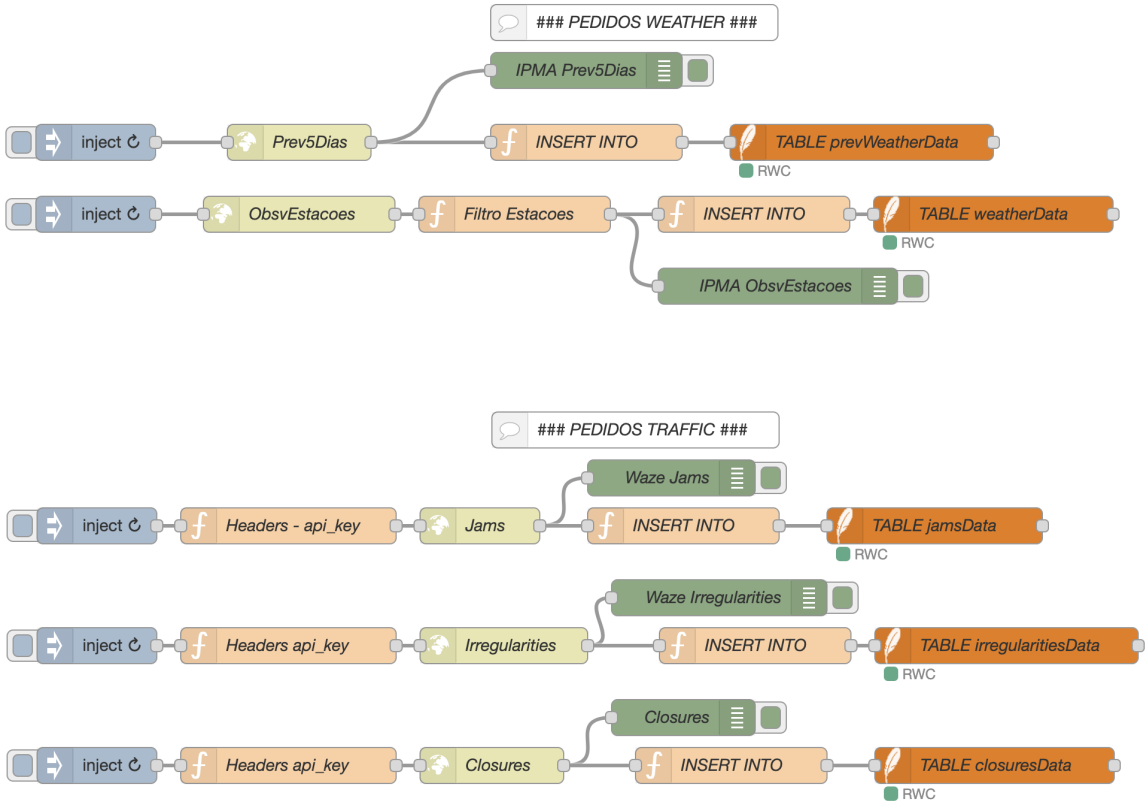


Figura 3.2: Principal fluxo programático do sistema implementado em Node-RED

3.3 Base de Dados

Os dados são armazenados em bases de dados SQLite, nomeadamente em ficheiros com a extensão *.db, visto ser um motor de base de dados que não necessita de configuração inicial nem de contas de utilizadores, possui tipos dinâmicos e de dimensões variáveis, é de fácil integração na ferramenta Node-RED e, principalmente, porque possui uma enorme facilidade na transmissão dos dados devido a ser uma base de dados orientada a documentos. Assim sendo, os dados encontram-se divididos sempre em duas bases de dados sendo uma para os dados relacionados com o tráfego que contém as tabelas *JamsData*, *IrregularitiesData* e *ClosuresData*, e uma para os dados relacionados com a meteorologia que contém as tabelas *WeatherData* e *PrevWeatherData*. Todas as tabelas são constituídas pelas seguintes colunas:

- *key* - identificador único do pedido
- *requestDate* - data e hora da realização do pedido
- *payload* - conteúdo da resposta do pedido em JSON

Foi decidido não inserir diretamente os dados obtidos em tabelas com colunas bem definidas, guardando todo o conteúdo da resposta do pedido HTTP numa única coluna (*payload*) pelo facto de, posteriormente, caso existisse alguma alteração por parte das fontes de dados, sendo que estas são controladas por entidades externas, não existisse a necessidade de criar novas tabelas ou alterar as existentes na base de dados nem alterar o fluxo programático na ferramenta Node-RED, prevenindo também a ocorrência de erros que por sua vez poderiam levar ao não armazenamento dos dados.

Tendo em conta que o volume de dados relacionado com o tráfego é elevado e tende a aumentar, foi realizada uma estimativa para prever qual iria ser o espaço ocupado pelos mesmos. Sendo que nas primeiras 9 horas de armazenamento a base de dados já alcançara um tamanho de cerca de 90 *megabytes* foi concluído que ao final de uma semana estaríamos perante uma base de dados com cerca de $1,70$ *gigabytes*. Como não se pretende trabalhar com ficheiros de elevadas dimensões, devido aos factos dos dados terem de ser: (a) transportados, por exemplo, entre o local onde estão armazenados e o local onde vão ser processados posteriormente, (b) processados, sendo que ficheiros maiores envolvem mais tempo de processamento e (c) carregados para diversas ferramentas, por exemplo, de análise e/ou visualização de dados sendo que estas nem sempre têm a capacidade de lidar adequadamente com grandes volumes de dados, decidiu-se manter uma organização semanal para ambas as bases de dados.

3.4 Pré-Processamento

Estando os dados armazenados em bases de dados SQLite, é necessário transformar o conteúdo de cada uma das tabelas para um formato onde se possa visualizar e permitir a análise por parte de outros *softwares* de uma forma mais simples. Assim sendo, o conteúdo de cada tabela está a ser convertido diretamente para ficheiros CSV através do comando indicado na Listagem 3.1, que pode ser executado na linha de comandos:

```
sqlite3 -header -separator ";" {dbInputFilePath} "Select * from jamsData;"
   > {csvOutputFilePath}
```

Listagem 3.1: Exportação de dados de uma tabela para um ficheiro csv.

Após a obtenção deste ficheiro CSV, o mesmo é carregado para o ambiente de desenvolvimento *RStudio* [31] onde é possível realizar uma análise estatística dos valores das características presentes nos conjuntos de dados permitindo assim perceber quais as características irrelevantes ou omissas para que posteriormente seja possível desenvolver e aplicar um conjunto de processamentos que objetivam a extração de um *dataset*

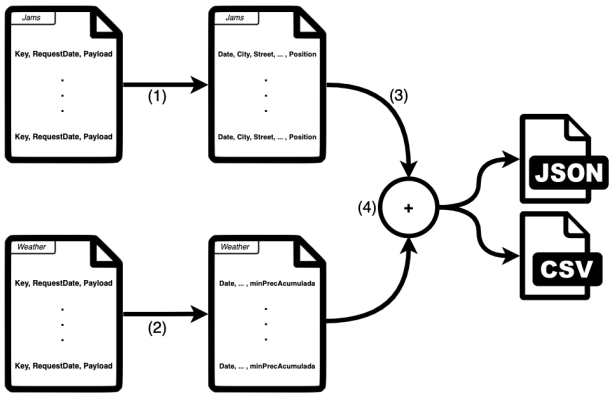


Figura 3.3: Diagrama de pré-processamentos que visa a construção do conjunto de dados que combinada os conjuntos *Jams* e *Weather*

limpo e cuidado capaz de ser facilmente integrado com as ferramentas de visualização e extração de conhecimento (Secções 3.5 e 3.6, respectivamente).

A utilização do ambiente de desenvolvimento *RStudio*, e consequentemente, da linguagem *R*, deve-se ao facto de ter todas as funcionalidades e pacotes necessários para todo o processo de desenvolvimento. Como permite analisar e manipular grandes conjuntos de dados com uma curva de aprendizagem relativamente baixa tornando-a assim numa excelente ferramenta de análise e computação estatística.

O primeiro passo do processamento, correspondente ao ponto (1) da Figura 3.3, extrai características relevantes da coluna *payload* de cada uma das linhas do ficheiro CSV, permitindo criar um novo ficheiro tabular onde cada um desses atributos ficam representados como colunas. 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 contém as coordenadas exactas (latitude e longitude) da localização da via onde ocorreu o congestionamento através de uma *Feature* do tipo *MultiLineString*. As restantes características tais como `country`, `turn_type`, `uuid` e `bbox` foram removidas devido a apresentarem valores irrelevantes, idênticos ou até mesmo inexistentes. Para o conjunto de dados *Weather*, o processamento é exactamente o mesmo e corresponde ao ponto (2) da Figura 3.3, e visto que estamos perante informações de três estações meteorológicas de Lisboa, o processamento consiste também em agregar a informação dessas três estações a 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. A nível da escolha de características,

nomeadamente para a construção de modelos, decidiu-se manter apenas as características `intensidadeVentoKM`, `temperatura` e `precAcumulada` pelo que se acredita que as restantes (`humidade`, `pressao` e `radiacao`) não têm influencia significativa no tráfego automóvel. No entanto, poder-se-á integrá-las futuramente se assim se justificar. Os restantes conjuntos de dados não estão a ser pré-processados pelo que ainda não existem intenções de os integrar num conjunto de dados final.

O processamento seguinte corresponde ao ponto (3) da Figura 3.3 e apenas é realizado no conjunto de dados *Jams*. 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 (resolução temporal mínima) 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. A resolução temporal mínima definida foi de 30 minutos, valor que vai de encontro com os interesses dos analistas da CML.

Por fim, e após aplicação de todos estes processamentos, ficamos com dois conjuntos de dados devidamente preparados com as características pretendidas que vão ser combinados através de um último processamento, correspondente ao ponto (4) da Figura 3.3, que visa criar assim o que denominamos de conjunto de dados base para o início da fase de modelação. Este processamento combina os dois conjuntos de dados através de uma junção dos mesmos, sendo a componente temporal (`datetime`) a chave utilizada para a realização da combinação. Deste modo, é obtido um conjunto de dados onde se podem observar congestionamentos num determinado local com determinadas condições meteorológicas. Este conjunto de dados é extraído em formato CSV e em formato JSON de forma a poder ser integrado em qualquer ferramenta, quer de visualização de dados quer de extração de conhecimento.

3.5 Visualização

Relativamente à visualização dos dados foi utilizada a ferramenta Kepler [18] que permite a análise de dados geoespaciais diretamente integrados num mapa. Estando a ferramenta apta a lidar com grandes quantidade de dados diretamente no *browser* considerou-se ser a mais 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.

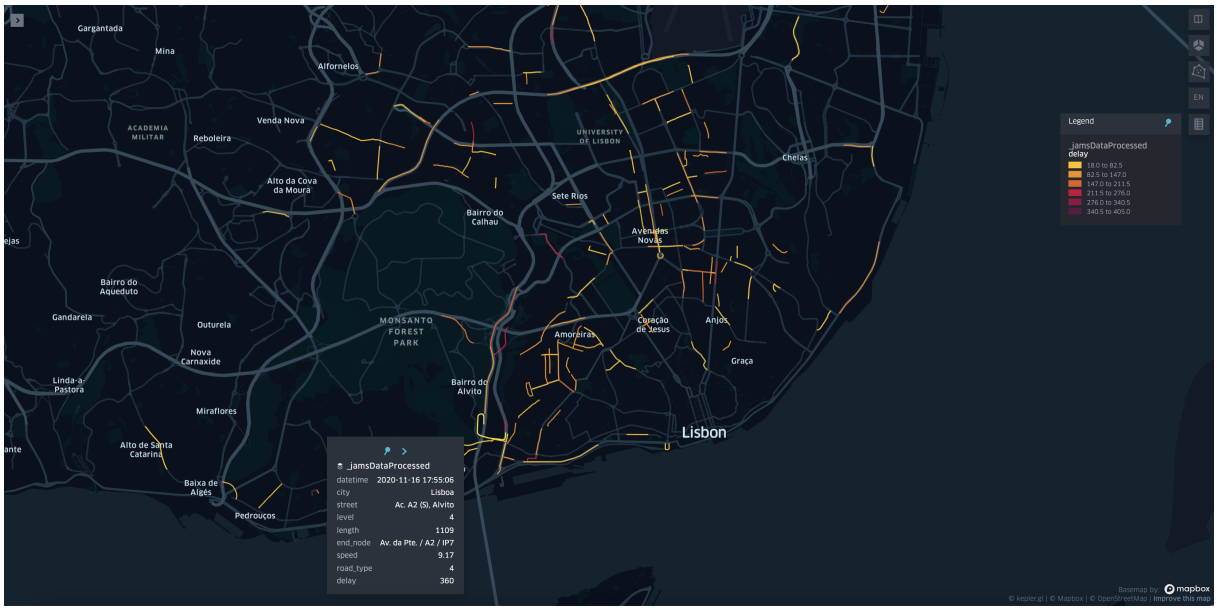


Figura 3.4: Representação gráfica geral do conjunto de dados representativo dos dias 16 a 22 do mês de Novembro de 2020 na ferramenta Kepler

Embora a ferramenta permitisse o uso de ficheiros CSV para carregamento no mapa, foi sempre utilizado o formato JSON que se mostrou ser o mais eficaz, na medida em que não apresentava quaisquer problemas no reconhecimento das características do conjunto de dados e apresentava carregamentos mais rápidos ao contrário do outro formato. Apenas é necessário que o conjunto de dados contenha uma característica no formato GeoJSON com o tipo *FeatureCollection*, sendo que neste caso cada uma das entradas do nosso conjunto de dados apenas contêm uma *Feature* com a informação das coordenadas relacionadas com determinado congestionamento de modo a que essas coordenadas possam ser reproduzidas no mapa. Os restantes campos são usados para observação e/ou manipulação gráfica aquando da interação entre o utilizador e o mapa. É possível observar um simples exemplo da sua utilização na Figura 3.4, onde estão representados os congestionamentos relativos às duas últimas semanas do mês de Novembro de 2020. De um modo geral, é possível observar com mais detalhe um congestionamento em específico apresentando todas as características relativas ao mesmo. De notar ainda que as cores dos troços dos congestionamentos representam o valor da característica `delay` (variável em análise), sendo que as cores mais claras, nomeadamente o amarelo, que representa um congestionamento mais ligeiro enquanto que as cores mais escuras representam um congestionamento mais demorado sendo este o caso da cor roxo.

Toda a zona do mapa permite uma interação total com o utilizador, especialmente os troços de vias delineados que ao serem selecionados pelo utilizador abrem uma janela

com as informações relativas ao congestionamento em questão. Existe ainda uma zona que permite ao utilizador definir todo o tipo de configurações, como por exemplo, se pretende ou não incorporar os nomes das zonas da cidade, estradas ou até mesmo rios ou oceanos no mapa, as características que pretende observar, a forma como pretende observar o valor de uma determinada característica sobre um troço de via podendo esta ser mapeada segundo um esquema de cores ou segundo a largura da linha que desenha o troço da via, entre outras. Existe também a possibilidade de aplicar filtros sobre qualquer característica presente no conjunto de dados de acordo com a gama de valores que se pretende observar.

Uma das características mais interessantes desta ferramenta é a forma como permite observar os dados no decorrer do tempo sob uma janela temporal, facilmente configurada de uma forma totalmente dinâmica, permitindo que seja possível vários tipos de observações mais detalhadas como por exemplo a análise dos dados de acordo com os movimentos pendulares ou a análise dos dados em momentos orientados a horas chave. Na Figura 3.5 são apresentados os dados relativos aos congestionamentos de tráfego na cidade de Lisboa em quatro momentos da manhã do dia 17 do mês de Novembro de 2020 e como se pode observar, o maior número de congestionamentos assim como os maiores tempos de atraso em cada congestionamento verificam-se nas horas que correspondem exactamente a movimentos pendulares, nomeadamente em a) e em b), como seria previsto. De notar que alguns períodos temporais encontram-se sobrepostos devido ao facto da ferramenta não permitir definir períodos temporais exatos, sendo esta uma limitação identificada e a ter em conta em futuras representações.

Todo este dinamismo que é permitido aquando da observação ao longo do tempo é bastante útil do ponto de vista da visualização humana, no entanto as janelas temporais da ferramenta ajustam-se exactamente à data e hora dos dados pelo que por vezes é mesmo impossível realizar uma análise deste tipo com uma dimensão exacta, isto é, o utilizador não tem total liberdade na escolha da dimensão da janela temporal. Ainda ao nível da visualização através uma janela temporal, não é possível observar qualquer tipo de desvanecimento na cor dos troços de via congestionados aquando de uma transição sendo que esta é realizada de forma abrupta o que poderá dificultar a interpretação ao nível da visualização humana.

De notar que as visualizações anteriormente apresentadas (Figuras 3.4 e 3.5) ainda contêm alguns problemas, nomeadamente a não representação gráfica dos congestionamentos onde a característica `delay` apresenta valores -1, pelo que o conjunto de dados ainda terá de passar por uma fase de modelação, e a não representação gráfica de qualquer tipo de indicador. A resolução destes problemas encontra-se descrita posteriormente nos Capítulos 4 e 5, respectivamente.

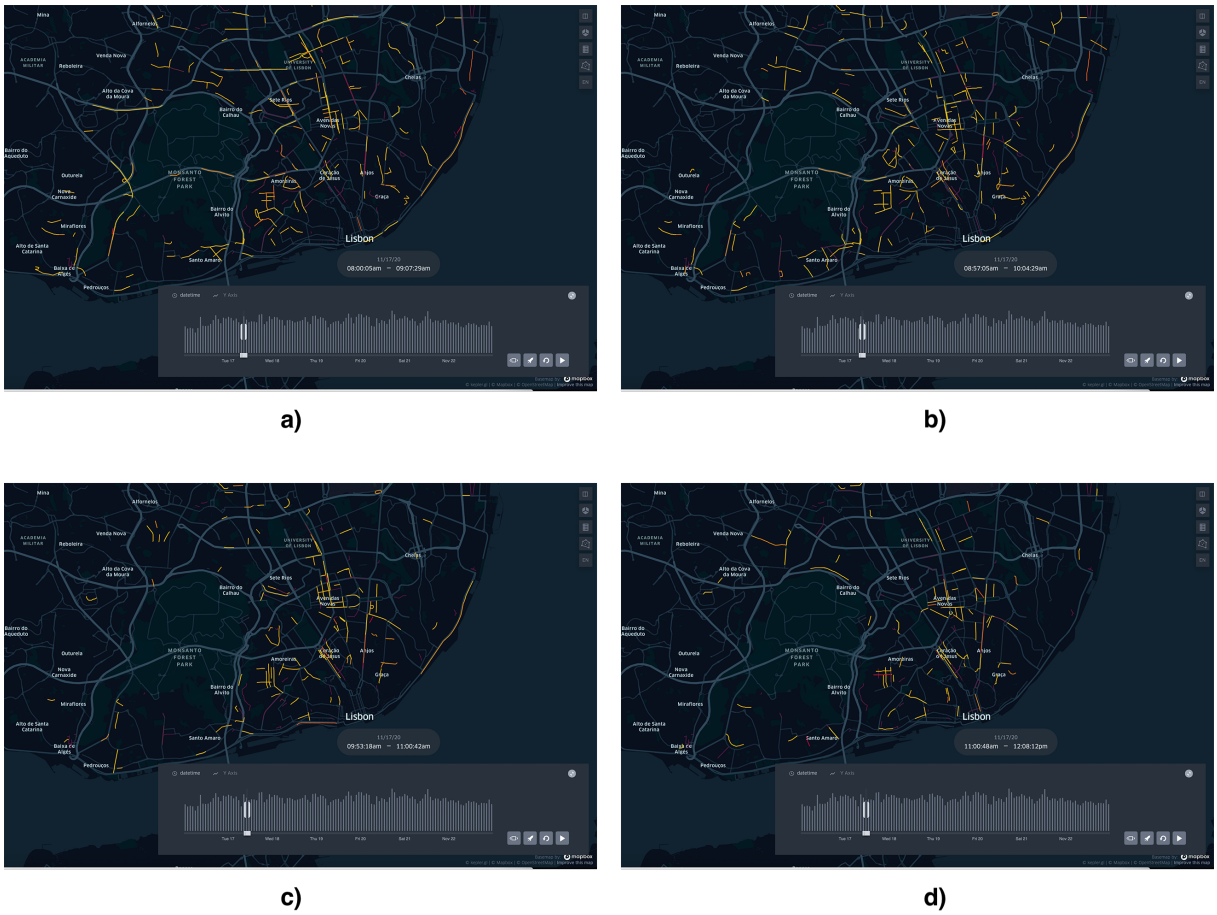


Figura 3.5: Visualização do atraso nos congestionamentos de tráfego do dia 17 do mês de Novembro de 2020 nos períodos de tempo 8:00–9:07 (a), 8:57–10:04 (b), 09:53–11:00 (c), 11:00–12:08 (d)

3.6 Extração de Conhecimento

Tendo em conta que os dados já se encontram devidamente tratados, e sendo um dos grandes objetivos do trabalho desenvolver modelos capazes de prever congestionamentos, foi utilizada a ferramenta H2O [12] que disponibiliza vários algoritmos de *machine learning* e aprendizagem automática para este fim. É uma ferramenta *open source* e pode ser utilizada através das suas bibliotecas em *R* ou *Python* assim como através da sua interface gráfica, H2O Flow [13], que possibilita a combinação da execução de código com texto, equações, tabelas e gráficos num único documento.

A característica escolhida para a previsão foi o `delay`, que corresponde ao tempo de atraso do tráfego em comparação com a via completamente livre. É reportada em segundos, pelo que através desta característica é possível ter uma noção mais real e

imediata do estado e da gravidade de um congestionamento sendo que, de modo geral, os humanos apresentam uma maior sensibilidade ao tempo do que, por exemplo, a distâncias ou velocidades. Outras características tais como o `level`, `length` ou até mesmo o `speed` também foram tidas em conta, na medida em que também conseguiriam indicar o estado de um congestionamento, no entanto com menos de precisão que o `delay`.

Inicialmente foi utilizado o algoritmo *Distributed Random Forest* [7] para a exploração da ferramenta na construção de modelos de regressão e consequente observação e manipulação dos seus resultados obtidos. Posteriormente, de forma a obter modelos mais precisos, passou-se a utilizar o algoritmo *XGBoost* [38] que é um algoritmo de aprendizagem supervisionada implementado sobre um processo de *boosting*, técnica que constrói modelos sequencialmente onde o modelo mais recente tende a corrigir lacunas dos anteriores, mais concretamente na técnica *gradient boosted decision trees* permitindo alcançar desempenhos excelentes ao nível da velocidade e performance. Na construção dos modelos é possível alterar vários parâmetros tais como colunas do conjunto de dados a ignorar, número de *folds*, critérios de paragem, entre outros pelo que o utilizador tem o poder de tomar decisões de acordo com o tipo de modelo que pretende, contudo foram utilizados os valores por omissão para cada um dos parâmetros.

Os modelos gerados podem ser extraídos da ferramenta H2O através de ficheiros zip e integrados numa aplicação Java, permitindo assim a realização de previsões em novos conjuntos de dados. De modo a testar esta funcionalidade, foi desenvolvida uma pequena aplicação Java que serviu como forma de ganhar conhecimento para o que se poderá vir a desenvolver futuramente utilizando todo este ambiente. No entanto, os modelos ao serem extraídos do H2O e integrados no Java deixa de ser possível a análise de métricas de avaliação dos mesmos pelo que se existir essa necessidade as métricas têm de ser construídas pelo programador. Assim sendo, foi sempre utilizada a interface gráfica (H2O Flow) de modo a facilitar a construção dos modelos assim como a observação e análise das métricas de avaliação.

De modo a continuar a testar a ferramenta, assim como a dar início a uma observação mais detalhada da importância das características na construção de modelos, e, visto que a escolha de características fará parte de um processo contínuo ao longo deste trabalho, foi construído um modelo com dados somente das vias de entrada e saída da Ponte 25 de Abril relativos à semana 49 do ano de 2020 de onde é possível retirar algumas conclusões acerca da importância das características. Na Figura 3.6 é possível observar a importância das características presentes no conjunto de dados na construção de um modelo através do algoritmo de *XGBoost* perante uma escala normalizada, concluindo assim que para este caso em particular as cinco características que mais

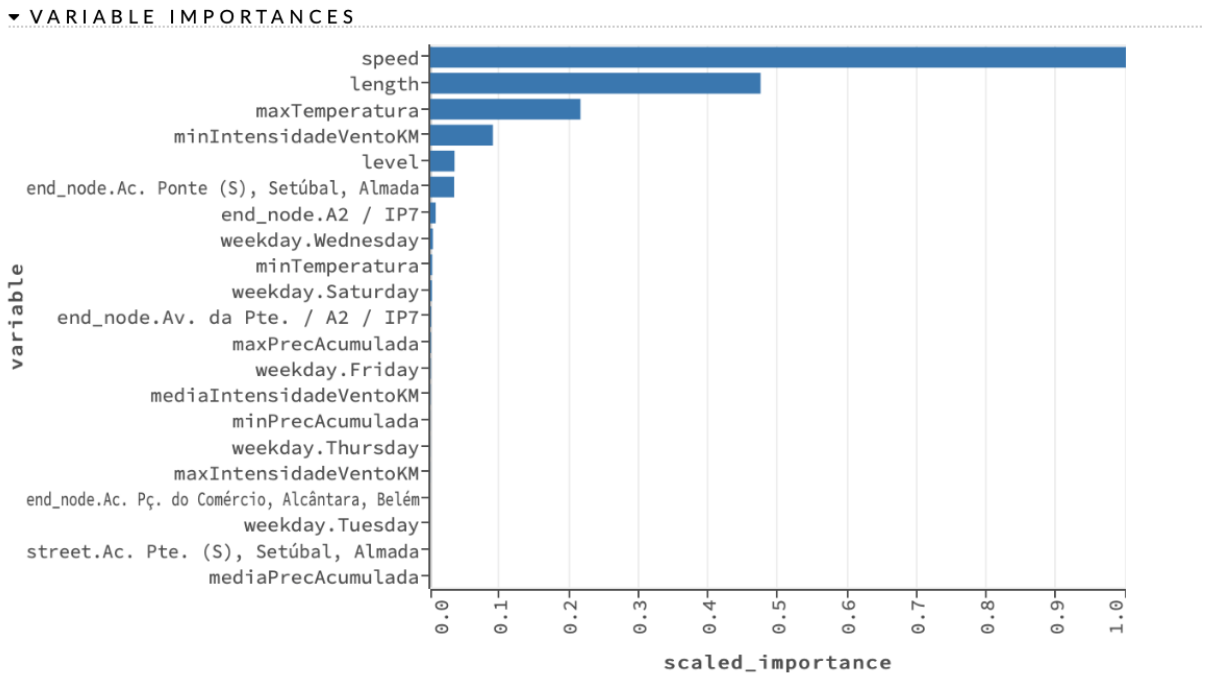


Figura 3.6: Importância das características presentes no conjunto de dados filtrado para entradas e saídas da Ponte 25 de Abril (Semana 49 do ano de 2020)

contribuíram para a previsão foram a velocidade do tráfego, o comprimento do congestionamento, a temperatura máxima observada, a intensidade mínima do vento e o nível de congestionamento.

Seria esperado que características, por exemplo, relacionadas com a precipitação tivessem uma maior importância na construção do modelo de previsão e tal não se verificou possivelmente devido ao conjunto de dados ser demasiado curto na medida em que este apenas abrange uma semana estando também filtrado para conter apenas as vias que constituem entradas e saídas da Ponte 25 de Abril. Concluí-se então que dependendo da diversidade das características presentes no conjunto de dados, estas possam ter importâncias bem distintas o que nos leva a pensar em acrescentar outras características ao conjunto de dados.

As métricas escolhidas para a avaliação dos modelos gerados foram as seguintes:

- *R Squared* (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.
- *Mean Squared Error* (MAE) - Representa a média dos erros absolutos sendo que as suas unidades são iguais às do valor previsto. Quanto menor for, melhor é o desempenho do modelo.

- *Root Mean Squared Error* (RMSE) - Representa a raiz do erro quadrático médio medindo o quão bem o modelo pode prever um valor continuo, também utiliza as mesmas unidades que o valor previsto. Quanto menor for, melhor é o desempenho do modelo.
- *Root Mean Squared Logarithmic Error* (RMSLE) - 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.

Foram escolhidas estas métricas por permitirem a análise e comparação dos modelos em através de vários aspectos. No caso de R^2 apresenta um valor normalizado o que permite uma comparação mais directa, já no caso de MAE é uma medida mais robusta perante *outliers* e é bastante útil para entender o tamanho do erro devido às unidades que utiliza (mesmas unidades que a variável em estudo - segundos), no caso de RMSE é uma medida sensível a *outliers* ao contrário de MAE pelo que através das análise destas duas medidas em conjunto é possível determinar a presença ou não de *outliers*, e por fim, no caso de RMSLE é uma medida bastante poderosa no sentido em que uma previsão mais baixa é mais desagradável do que uma previsão mais alta, ou seja, 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, esta medida é utilizada com este sentido.

Com base em todas estas análises e testes preliminares é definido então o algoritmo a utilizar na construção de modelos de previsão (*XGBoost*), as métricas para avaliação dos modelos (R^2, MAE, RMSE e RMSLE) e o conjunto de dados base à modelação através da combinação dos conjuntos de dados relativos aos congestionamentos de tráfego (*Jams*) e à meteorologia (*Weather*) assim como algumas características neles a manter e a resolução temporal mínima a utilizar.

4

Engenharia de Modelos e Dados

Encontrando-se já construído o conjunto de dados base sob o qual se quer desenvolver, aplicar e analisar uma série de modelos de previsão, conjunto este que combina as características dos conjuntos *Jams* (Tabela 3.1) e *Weather* (Tabela 3.4) e cujas características estão presentes na Tabela 4.1, pretende-se que estes modelos sejam capazes de prever a característica `delay`, ou seja, o valor do tempo de atraso dos congestionamentos.

Tabela 4.1: Descrição das características presentes nas propriedades do conjunto de dados base à modelação que combina as características dos conjuntos *Jams* e *Weather*.

Nome do campo	**Descrição**
datetime	Data e hora do congestionamento
city	Nome da cidade ou estado
street	Nome da rua
level	Nível de congestionamento de tráfego (0 = via completamente livre, 5 = via completamente congestionada)
length	Comprimento do congestionamento em metros
end_node	Saída mais próxima do final do congestionamento
speed	Velocidade média do tráfego em metros por segundo

continua na próxima página

Nome do campo	Descrição
road_type	Tipo de via (1 = Rua, 2 = Rua principal, 3 = Auto-estrada, 4= Rampa, 5 = Caminho, 6 = Rua principal, 7 = Rua secundária, 8 = Caminho de terra batida, 9 = Passeio, 10 = Caminho pedestre, 11 = Saída, 14 = Caminho de terra batida, 15 = Travessia de barco, 16 = Escadas, 17 = Caminho privado, 18 = Caminho de ferro, 19 = Faixa exclusiva (taxi/bus), 20 = Via de acesso ou dentro de parque de estacionamento, 21 = Estrada de serviço)
delay	Tempo de atraso do tráfego em comparação com a via completamente livre em segundos (-1 = via completamente congestionada)
position	Conjunto de coordenadas que representam o local do congestionamento (Objecto *GeoJSON*)
minIntensidadeVentoKM	Valor mínimo de intensidade do vento registada a 10 metro de altura em kilometros por hora das três estações meteorológicas de Lisboa
maxIntensidadeVentoKM	Valor máximo de intensidade do vento registada a 10 metro de altura em kilometros por hora das três estações meteorológicas de Lisboa
mediaIntensidadeVentoKM	Valor médio de intensidade do vento registada a 10 metro de altura em kilometros por hora das três estações meteorológicas de Lisboa
minTemperatura	Valor mínimo da média da temperatura do ar registada a 1.5 metros de altura numa hora em graus centígrados das três estações meteorológicas de Lisboa
maxTemperatura	Valor máximo da média da temperatura do ar registada a 1.5 metros de altura numa hora em graus centígrados das três estações meteorológicas de Lisboa
mediaTemperatura	Valor médio da média da temperatura do ar registada a 1.5 metros de altura numa hora em graus centígrados das três estações meteorológicas de Lisboa
minPrecAcumulada	Valor mínimo do valor acumulado da precipitação registada a 1.5 metros de altura numa hora em milímetros das três estações meteorológicas de Lisboa
maxPrecAcumulada	Valor máximo do valor acumulado da precipitação registada a 1.5 metros de altura numa hora em milímetros das três estações meteorológicas de Lisboa
mediaPrecAcumulada	Valor médio do valor acumulado da precipitação registada a 1.5 metros de altura numa hora em milímetros das três estações meteorológicas de Lisboa

Para alcançar este objetivo foi então realizada uma análise estatística do conjunto de

dados como um foco especial na característica `delay`, o que permitiu identificar problemas relacionados com a mesma e que poderiam por em causa o desenvolvimento e desempenho dos modelos de previsão. Ao longo deste capítulo, estes problemas são identificados assim como são apresentadas soluções para os mesmos.

Contudo, o foco principal deste capítulo é o desenvolvimento e a análise de desempenho de modelos de previsão construídos segundo várias variantes do conjunto de dados com o objetivo de obter um ou mais modelos capazes de realizar previsões com um desempenho adequado ao domínio do problema.

4.1 Análise estatística do conjunto de dados

O conjunto de dados utilizado para esta análise é relativo a toda a cidade e contém informação recolhida entre o dia 11 de Novembro e 20 de Dezembro do ano de 2020, representando assim as semanas 46 a 51 desse mesmo ano, totalizando 1 780 507 registos. De notar que nesta análise são apenas incluídas seis semanas, pelo que estas podem não ser representativas das restantes semanas do ano.

Através de análises preliminares já se conhecia de antemão que existia um número elevado de registos que apresentavam a característica `delay` com um valor de -1, que representa uma via completamente congestionada, pelo que foi realizada uma contagem desses mesmos registos obtendo-se então um total de 1 508 350 registos onde esta característica apresentava o valor de -1 em comparação com 272 157 registos onde já apresentavam valores capazes de quantificar efectivamente o tempo de atraso dos congestionamentos. Foi também realizada uma análise estatística, que pode ser observada na Tabela 4.2, onde se observam os valores mínimo, 1º quartil, mediana, média, 3º quartil e máximo da característica `delay` no conjunto de dados original, que inclui todos os registos, e no conjunto de dados filtrado onde foram retirados todos os registos que apresentavam o valor de -1 na característica `delay`, permitindo assim observar com maior detalhe o impacto que a grande quantidade de valores -1 tem no conjunto de dados. Esse impacto pode ser visto, por exemplo, nos valores apresentados no 3º quartil onde para o conjunto original foi obtido um valor de -1, mostrando que pelo menos 75% da amostra ordenada apresentava valores negativos.

Estamos assim perante um cenário onde cerca de 85% dos registos deste conjunto de dados apresenta um valor negativo para quantificar um atraso, e dado que não existem atrasos negativos estamos perante alguns problemas: (1) a incerteza gerada quanto ao tempo real de atraso do congestionamento pela presença do valor negativo, (2) a utilização de modelos de previsão mediante valores com esta característica, pelo que os

Tabela 4.2: Análise estatística da característica `delay`.

Conjunto de Dados	Mínimo	1º Quartil	Mediana	Média	3º Quartil	Máximo
Original	-1	-1	-1	20.68	-1	5244
Filtrado	2	82	105	140.8	150	5244

modelos assumirão valores positivos próximos de 0 como previsões adequadas a um valor de -1 (via totalmente congestionada), o que não é de todo correto tendo em conta que valores positivos próximos de 0 representam atrasos bastante ligeiros, o que culminará numa eventual distorção das métricas sob as quais serão avaliados os modelos, e (3) a dificuldade na forma sob a qual irão ser representados graficamente estes valores na ferramenta de visualização, tendo em conta que têm de ser comparados com os restantes atrasos.

Na Figura 4.1 é possível observar um gráfico com a contagem total do número de congestionamentos por hora do dia das seis semanas em que os dados foram recolhidos onde se observa que os valores máximos foram obtidos em horas de ponta, sobretudo no período da tarde, onde se chegou a registar mais de 90 000 congestionamentos numa única hora. A obtenção destes valores máximos para estas horas específicas (horas de ponta) pode-se também associar de certo modo aos movimentos pendulares sendo que são horas de bastante movimento na cidade. Nota-se um crescimento acentuado no número de congestionamentos, sobretudo, entre as 7 e as 8 horas da manhã e entre as 16 e as 17 horas da tarde, o que poderá ser um bom indicador de que seria interessante adicionar ao conjunto de dados uma característica que indique se o congestionamento ocorreu ou não em hora de ponta, podendo esta ser definida, por exemplo, entre as 7 e as 10 horas da manhã e as 16 e as 19 horas da tarde tal como se pode observar na marcação a tracejado presente no gráfico.

Foi também realizada uma outra análise que realiza a contagem total do número de congestionamentos por dia da semana nas seis semanas em que os dados foram recolhidos onde se observou que os dias da semana que registaram um maior número de congestionamentos foram sexta-feira, quinta-feira e sábado com um total de 313 466, 294 109 e 262 622 congestionamentos, respectivamente. Por outro lado, o dia da semana onde foi registado menos congestionamentos foi segunda-feira com um total de 213 297 congestionamentos. Perante estes resultados surge a possibilidade de se integrar no conjunto de dados uma característica com o dia da semana em que o congestionamento ocorreu.

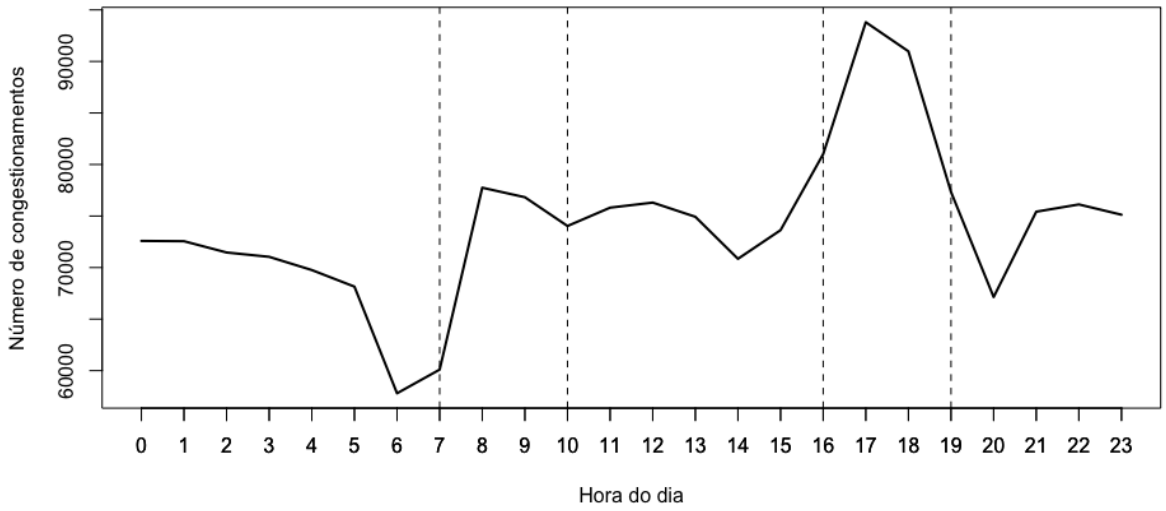


Figura 4.1: Contagem total do número de congestionamentos por hora do dia (Semanas 46 a 51 do ano de 2020)

4.2 Modelação de previsão de atraso

Encontrando-se já definido o algoritmo que irá ser utilizado para a construção dos modelos de previsão (*XGBoost*), mais concretamente, modelos de regressão devido à característica sob a qual se quer realizar previsões (`delay`) ser do tipo numérico. Foi ainda pensada a utilização de modelos de classificação, no entanto, seria necessário estabelecer um conjunto intervalos de valores e associá-los a classes, o que retiraria às previsões algum nível de precisão em comparação com um modelo de regressão.

Todos os modelos foram construídos através da ferramenta H2O. Dado que esta é adequada à realização da modelação, e que já se encontra em utilização na CML, facilitando assim uma possível integração destes mesmos modelos. Todos os modelos utilizaram a parametrização por omissão da ferramenta para o algoritmo *XGBoost*.

Os dados utilizados na construção dos modelos posteriormente apresentados são relativos a toda a cidade de Lisboa na semana 49 do ano de 2020 e apresentam um total de 86532 registos (conjunto de treino) sendo as previsões realizadas relativas à semana seguinte cujo total de dados apresentam 97074 registos (conjunto de teste).

Com o objetivo de aumentar o desempenho dos modelos foram ainda derivadas novas características a partir do conjunto de dados base (Tabela 4.1), 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/20h); 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 a de atribuir valores substitutos com base nos dados existentes e no domínio do problema. Dada a natureza desta variável e tendo em conta essa incerteza, todos os valores iguais a -1 foram substituídos por um valor que resulta da soma do valor máximo existente nos dados com N vezes o desvio padrão, sendo N o número de vezes que se pretende aumentar o valor resultante em relação ao seu valor máximo. Esta solução permite também resolver os problemas (mencionados anteriormente na Secção 4.1) da utilização de valores negativos em modelos de previsão assim como a questão da dificuldade de representação gráfica destes valores em ferramentas de visualização.

Foram então geradas seis variantes do conjunto de dados base (Tabela 4.1), com as quais foram construídos seis modelos preditivos, como descrito na Tabela 4.3.

Na Tabela 4.4 podem ser observados 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.

Tabela 4.3: Descrição dos modelos.

Modelo	**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

Tabela 4.4: 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**

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 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 o mais adequado mas também pelo facto dos seus resultados se aproximarem mais da realidade tendo em conta o domínio do problema.

Sendo que o conjunto de dados sob o qual foi construído este modelo tirou partido da utilização da formula $\max(\texttt{Delay}) + N \times \texttt{DesvioPadrao}$ para a substituição dos valores

da característica `delay`, neste caso com *N = 1*, foram também construídos outros modelos com diferentes valores de *N*. Nomeadamente, com *N = 2, 3 e 4* onde se observou que as métricas de avaliação desses modelos eram idênticas à do modelo previamente obtido (Modelo C+) pelo que se decidiu manter um valor de *N = 1* visto que este não alterava significativamente a distribuição dos valores de `delay`.

4.2.1 Grau de correlação com a variável dependente

Tendo em conta que para a realização de previsões, isto num contexto do mundo real, não teremos disponíveis as características `speed`, `length` e `level` sendo estas a velocidade média do tráfego, o comprimento do congestionamento e o nível de congestionamento, respectivamente, e de acordo com alguns resultados preliminares, nomeadamente, a importância das características observada anteriormente na Figura 3.6 foi medido o grau de correlação entre estas características e a variável dependente `delay` através de coeficientes de correlação para entender o quão fundamentais são para um bom desempenho por parte dos modelos.

A correlação de variáveis baseia-se em métodos estatísticos com o objetivo de medir as relações entre variáveis e o que elas representam, nomeadamente, entender como se comporta uma em função da outra. [10]

Neste trabalho vão ser utilizados os coeficientes de correlação de *Pearson* (ou correlação linear) e de *Spearman*. [4]

Pearson representa o grau de relação entre duas variáveis quantitativas, sendo esse grau exprimido em valores entre -1 e 1 onde os valores negativos representam uma correlação negativa, isto é, quando o valor de uma variável diminui, o valor da outra aumenta ou vice-versa e os valores positivos representam uma correlação positiva, ou seja, quando o valor de uma variável aumenta, o valor da outra variável também aumenta. Quanto mais os valores se aproximarem dos extremos (-1 e 1) mais forte é a correlação entre as variáveis, sendo que o valor 0 indica que não existe nenhuma correlação entre as variáveis.

Spearman representa uma medida de correlação não paramétrica onde não exige necessariamente que a relação entre as variáveis seja linear, ou seja, as variáveis tendem a mover-se na mesma direção relativa mas não necessariamente a uma taxa constante, pelo que esta medida avalia a relação monotónica entre as variáveis. Esta medida exprime-se através dos mesmos valores que o coeficiente de correlação de *Pearson*.

Na tabela 4.5 é possível observar os valores obtidos dos coeficientes de correlação entre as características `speed`, `length` e `level` em relação à variável dependente `delay`.

Tabela 4.5: Coeficientes de correlação entre as características `speed`, `length` e `level` em relação à variável dependente `delay`.

Coeficiente de correlação	**Speed**	**Length**	**Level**
Pearson	-0.748	-0.472	0.881
Spearsman	-0.917	-0.704	0.949

Tabela 4.6: Medidas de avaliação dos modelos de previsão onde não foram incluídas as características `speed`, `level` e `length`.

Modelo	**R2**	**MAE**	**RMSE**	**RMSLE**
Sem a característica `speed`	0.969	157.041	210.297	—
Sem a característica `length`	0.970	157.199	209.472	0.255
Sem a característica `level`	**0.972**	**142.885**	**201.525**	**0.134**
Sem as características `level` e `speed`	0.940	186.521	293.641	—
Sem as características `level` e `length`	0.970	157.986	209.646	0.266
Sem as características `length` e `speed`	0.968	161.906	214.691	0.292
Sem as características `speed`, `level` e `length`	0.796	331.737	542.711	—

Em relação ao coeficiente de correlação de *Pearson*, este indica-nos claramente que as características `speed` e `level` têm um grau de correlação alto com a característica `delay` apresentando valores próximos dos extremos da escala, ao contrário da característica `length` que não apresenta uma correlação tão elevada. Em relação ao coeficiente de correlação de *Spearman*, este demonstra um grau de correlação elevadíssimo para as características `speed` e `level` assim como um grau de correlação elevado para a característica `length` que nos confirma que também esta característica tem um peso importante no desempenho dos modelos de previsão.

As diferenças nos valores obtidos entre os dois tipos de coeficientes de correlação deve-se à questão das relações de linearidade ou de não linearidade, sendo esta diferença melhor observada no caso da característica `length` onde a correlação é muito maior segundo o coeficiente de correlação de *Spearman* do que segundo o coeficiente de correlação de *Pearson*.

Com base nestas conclusões foi então decidido construir mais alguns modelos de previsão segundo o conjunto de dados base (Tabela 4.1), sendo que agora não serão incluídas estas três características (`speed`, `length` e `level`), para que se possa analisar as métricas desses mesmos modelos e perceber o quão prejudicado ficará o seu desempenho em relação aos demais.

Na Tabela 4.6 é possível observar as medidas de avaliação dos modelos de previsão

construídos sem as características `speed`, `level` e `length`. Os resultados das métricas de avaliação obtidas para os modelos construídos com um conjunto de dados que não incluíram uma ou duas destas características indicam que estes modelos não demonstram melhor desempenho em relação ao Modelo C+ (Tabela 4.4) tal como era esperado, contudo observam-se níveis de desempenho muito próximos.

No entanto, o modelo em que nos pretendemos focar será o modelo construído sem as três características pelo facto, anteriormente mencionado, de que numa situação mais aproximada da realidade as mesmas não estarão disponíveis em tempo útil à realização de previsões. Ao nível das suas medidas de avaliação é o modelo onde se observou até à data os piores valores das mesmas, apresentando um R^2, MAE, e RMSE de $0,796$, $331,737$ e $542,711$, respectivamente. Ou seja, a nível de MAE e RMSE estamos perante erros na ordem de cerca de 6 e 9 minutos, respectivamente. Em relação ao RMSLE, não foi possível a obtenção do cálculo do seu valor.

Estes resultados, em comparação com o Modelo C+ (Tabela 4.4) que apresenta um R^2, MAE, e RMSE de $0,972$, $142,122$, $201,37$, respectivamente, demonstram uma grande queda de desempenho na medida em que para o R^2 diminuiu cerca de 15%, o MAE aumentou em cerca de 133% e o RMSE aumentou em cerca de 170%. Posto isto, confirma-se claramente que estas características (`speed`, `level` e `length`) são realmente essenciais para o aumento do desempenho das previsões.

4.2.2 Integração de dados históricos

Existindo a confirmação de que as características `speed`, `level` e `length` possuem um papel importante no desempenho das previsões, foi decidido incorporar no conjunto de dados, que visa construir os modelos (Tabela 4.1), informação histórica referentes às três características em questão.

Para tal, e estando os dados sob os quais estamos a trabalhar organizados segundo uma resolução mínima temporal de 30 minutos onde cada entrada do conjunto de dados representa um congestionamento, foram mantidas informações históricas das características em questão observadas nos últimos três congestionamentos (30m, 1h e 1h30), ou seja, mantidos os dados históricos da última hora e meia. No caso dos congestionamentos onde não existe esta informação, as características históricas ficaram com um valor vazio. Não foi considerado manter dados históricos mais antigos uma vez que se partiu do pressuposto de que o passado recente serviria para caracterizar um determinado congestionamento.

Tabela 4.7: Descrição das características que contêm a informação histórica acrescentadas ao conjunto de dados base (Tabela 4.1).

Nome do campo	Descrição
speed{1,2,3}_median	Mediana da característica `speed` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
speed{1,2,3}_mean	Média da característica `speed` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
speed{1,2,3}_max	Máximo da característica `speed` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
level{1,2,3}_median	Mediana da característica `level` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
level{1,2,3}_mean	Média da característica `level` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
level{1,2,3}_max	Máximo da característica `level` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
length{1,2,3}_median	Mediana da característica `length` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
length{1,2,3}_mean	Média da característica `length` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
length{1,2,3}_max	Máximo da característica `length` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
delay{1,2,3}_median	Mediana da característica `delay` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
delay{1,2,3}_mean	Média da característica `delay` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
delay{1,2,3}_max	Máximo da característica `delay` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.

Ao ser desenvolvida esta abordagem pensou-se que também poderia ser bastante vantajoso para os modelos manter informação histórica da própria característica a prever (`delay`). Assim sendo, foram acrescentados os dados históricos que se podem observar na Tabela 4.7. De notar que cada uma das entradas da tabela corresponde a uma característica observada na última meia hora, uma hora e hora e meia pelo que origina sempre três novas colunas no conjunto de dados.

Pelo facto de no mesmo conjunto de 30 minutos poderem existir múltiplos congestionamentos para as mesmas ruas, incluiu-se a mediana, a média e o valor máximo dos dados históricos obtidos de modo a dar mais opções de escolha de características ao modelo de previsão. De notar, no entanto, que não se incluiu o valor mínimo pelo que este representaria uma perspectiva otimista do problema e não é de todo o pretendido para o cenário em questão.

Tabela 4.8: Medidas de avaliação do modelo de previsão C+ e modelos onde não foram incluídas as características `speed`, `level` e `length` com e sem dados históricos.

Modelo	**R2**	**MAE**	**RMSE**	**RMSLE**
Modelo C+	**0.972**	**142.122**	**201.377**	**0.126**
Modelo C+ (sem as características `speed`, `level` e `length`) **sem** dados históricos	0.796	331.737	542.711	—
Modelo C+ (sem as características `speed`, `level` e `length`) **com** dados históricos	0.936	189.428	304.899	—

Na Tabela 4.8 é possível observar as medidas de avaliação dos dois modelos de previsão, anteriormente construídos, Modelo C+ e Modelo C+ onde não foram incluídas as características `speed`, `level` e `length` (sem incluir dados históricos) e também o novo modelo, idêntico ao anterior mas com a inclusão dos dados históricos de modo a ser possível comparar directamente os três.

Observando estes resultados, e como era de esperar o pior dos três modelos é o que não inclui as três características acima mencionadas nem os dados históricos. No entanto, o novo modelo construído que embora também não inclua estas três características, ao incluir os dados históricos garante uma melhoria bastante significativa nas suas medidas de avaliação. Estas medidas chegam mesmo a aproximar-se das medidas obtidas com o Modelo C+, modelo construído com todas as características disponíveis, onde fora obtido um valor de R^2 de $0,972$ em comparação com o agora obtido no modelo com dados históricos de $0,936$, o que representa uma diminuição de aproximadamente 4%. Em relação aos valores de MAE e RMSE, para o Modelo C+ foram obtidos os valores $142,122$ e $201,377$, em comparação com os valores agora obtidos no modelo com dados históricos de $189,428$ e $304,899$, o que representa um aumento de aproximadamente 33% e 51%, respectivamente. A maior diferença observa-se na medida RMSE em cerca de 100 segundos, no entanto, o valor obtido de $304,899$ segundos representa aproximadamente 5 minutos e encontra-se ainda num intervalo de tempo aceitável para o erro em questão. Assim sendo, o modelo que inclui os dados históricos anteriormente apresentados demonstra ser o mais indicado para a realização de previsões, sobretudo, em tempo real onde não é de todo possível o acesso às características `speed`, `level` e `length`.

4.2.3 Modelo de zonas

Tendo este trabalho como objetivo a previsão de congestionamentos em toda a cidade de Lisboa, tem também um especial interesse em determinadas zonas, nomeadamente,

Tabela 4.9: Conjunto de ruas selecionadas para a zona da Ponte 25 de Abril.

Street	**End_node**
Pte. 25 Abril / Viaduto Norte	Pte 25 de Abril
Ac. Pte. (S), Setúbal, Almada	A2 / IP7
Ac. Pte. (S), Setúbal, Almada	Ac. Ponte (S), Setúbal, Almada
Ac. Pte. 25 de Abril, A2 (S)	Lisboa (Campo de Ourique)
IP7 / Eixo N-S	Lisboa (Alvito)
Ac. Av. da Ponte, A2 (S), IC20, Almada, Caparica	Lisboa (Br. da Liberdade)
Pte. 25 de Abril	HGO Almada
Ac. A2 (S), Alvito	Av. da Pte. / A2 / IP7
Pte. 25 Abril / Viaduto Norte	A2 / IP7
Ac. Pç. do Comércio, Alcântara, Belém	Ac. Pç. do Comércio, Alcântara, Belém
Ac. 1, A5, Cascais, IC19, Sintra, Monsanto	Lisboa (Campo de Ourique)

Tabela 4.10: Conjunto de ruas selecionadas para a zona da Calçada de Carriche.

Street	**End_node**
Cç. de Carriche	Lisboa (Lumiar)
Cç. de Carriche	Cç. de Carriche
IC22	Sr. Roubado
A8	Sr. Roubado
Estr. do Desvio	Cç. de Carriche

pontos de entrada e de saída da cidade onde, por norma, se verifica um maior fluxo de tráfego. Sendo todos os modelos anteriores desenvolvidos com o objetivo de prever não zonas específicas mas sim toda a cidade, foi então estudada a hipótese de ter um modelo para cada zona específica.

Foram escolhidas duas zonas de interesse, Ponte 25 de Abril e Calçada de Carriche, de onde foram selecionadas um conjunto de ruas (`street` e `end_node`), de acordo com os dados fornecidos pelo Waze, que podem ser observadas nas Tabelas 4.9 e 4.10, respectivamente.

Tendo por base o conjunto de dados onde se inclui os dados históricos anteriormente descritos, foram construídos modelos com dados relativos às ruas selecionadas para cada uma destas duas zonas e comparados com um modelo que utiliza dados de toda a cidade.

Tabela 4.11: Medidas de avaliação dos modelos de previsão para a zona da Ponte 25 de Abril.

Modelo	**R2**	**MAE**	**RMSE**	**RMSLE**
Ponte 25 de Abril (1 Semana)	0.588	153.065	223.217	0.616
Ponte 25 de Abril (2 Semanas)	**0.825**	**107.236**	**145.561**	0.449
Ponte 25 de Abril (4 Semanas)	0.716	113.391	185.366	**0.438**
Modelo com dados de toda a cidade (1 Semana)	0.799	117.453	156.014	**0.438**

Ao utilizar o conjunto de dados relativo à semana 49 do ano de 2020, e após filtrar o mesmo para os conjuntos de ruas selecionadas notou-se que se obteve um número bastante reduzido de registos, mais concretamente, 89 registos para a zona da Ponte 25 de Abril e 33 registos para a zona da Calçada de Carriche. De modo a aumentar o número de registos, decidiu-se também construir modelos com dados relativos a 2 semanas (Semana 48 e 49) e 4 semanas (Semana 46 a 49).

Para a realização das previsões foi sempre utilizado um conjunto de dados filtrado para os conjuntos de ruas selecionadas com dados relativos à semana 50 do ano de 2020 para cada uma das duas zonas de interesse.

Nas Tabelas 4.11 e 4.12 é possível observar as medidas de avaliação dos modelos de previsão construídos para a zona da Ponte 25 de Abril e Calçada de Carriche, respectivamente.

Em relação aos modelos construídos para a zona da Ponte 25 de Abril, observa-se que os piores resultados são obtidos no modelo construído somente com dados de 1 semana muito possivelmente devido ao reduzido número de registos. Em contrapartida, os melhores resultados observam-se no modelo construído com dados de 2 semanas e não 4 semanas, pelo que o aumento do número de registos para a construção do modelo não significa diretamente um aumento na qualidade do mesmo. O segundo melhor modelo, foi o construído com os dados de toda a cidade apresentando medidas de avaliação bastante próximas dos melhores resultados obtidos.

Em relação aos modelos construídos para a zona da Calçada de Carriche, os resultados obtidos demonstram, de modo geral, um fraco desempenho por parte dos modelos. Por exemplo, na medida de avaliação R^2 chega-se a obter valores negativos, indicando estes que os modelos em questão não estão de todo adequados aos dados

Tabela 4.12: Medidas de avaliação dos modelos de previsão para a zona da Calçada de Carriche.

Modelo	**R2**	**MAE**	**RMSE**	**RMSLE**
Calçada de Carriche (1 Semana)	-0.113	95.729	137.653	0.611
Calçada de Carriche (2 Semanas)	-0.0832	110.831	135.816	0.635
Calçada de Carriche (4 Semanas)	**0.238**	**84.201**	**113.935**	**0.514**
Modelo com dados de toda a cidade (1 Semana)	0.180	90.614	118.154	0.535

sob os quais se quer realizar as previsões. Os melhores resultados obtidos são provenientes do modelo construído com dados de 4 semanas, no entanto, os resultados do modelo construído com os dados de toda a cidade também se aproximam bastante a estes, à semelhança do que foi observado para os modelos da zona da Ponte 25 de Abril.

Conclui-se assim que para os dados que foram utilizados não existe vantagem em considerar zonas específicas para a criação de modelos, pelo que o tempo e a complexidade computacional que envolve a criação destes modelos pode não justificar o aumento pouco significativo da qualidade e desempenho dos mesmos. No entanto, não exclui a possibilidade de que com outros dados e/ou com a incorporação de outras características essa vantagem possa existir.

4.2.4 Discussão de resultados

Tendo em conta toda a análise realizada sobre o desempenho dos diferentes modelos de previsão construídos, sendo que estes tiveram por base diversos conjuntos de dados onde foi testada a integração de novas características assim como a remoção de outras às quais não seria possível o acesso em tempo útil para a realização de previsões, obtêm-se um conjunto de dados com um total de 62 características que pode ser observado no Anexo A.

A partir deste conjunto de dados é possível construir o modelo de previsão que se mostrou ser o mais adequado à realização de previsões de tempos de atraso em congestionamentos (*Modelo C+* ***com*** *dados históricos* - Tabela 4.8) no qual foram obtidos erros médios de previsão entre os 3 e os 5 minutos segundo as métricas de avaliação MAE e RMSE, respectivamente.

5

Indicadores de fluidez de tráfego

Nesta fase do trabalho, já tendo os conjuntos de dados relacionados com os congestionamentos de tráfego devidamente preparados, isto é, tanto os congestionamentos realmente observados assim como os congestionamentos previstos, existe a necessidade de ter uma representação visual destes mesmos dados. Esta representação visual consiste, essencialmente, num *dashboard* onde se integram os dados num mapa interativo da cidade de Lisboa onde estes podem ser observados ao longo do tempo e de forma sumariada através da utilização de indicadores de fluidez de tráfego.

Como mencionado anteriormente na Secção 3.5, a ferramenta de visualização Kepler foi utilizada para explorar a representação visual dos congestionamentos. Contudo, após uma exploração mais detalhada e aprofundada da mesma, foram encontradas algumas limitações nas suas funcionalidades. A ferramenta não permite aplicar o mesmo filtro temporal em mais do que um conjunto de dados em simultâneo, ou seja, iria exigir que os conjuntos de dados relacionados com os indicadores a desenvolver neste capítulo tivessem de estar no mesmo conjunto de dados que os congestionamentos para que toda a informação pudesse ser observada ao longo do tempo e de forma dinâmica. Outra limitação da ferramenta é não permitir ao utilizar definir intervalos de valores para representar uma característica do conjunto de dados segundo um esquema de cores, disponibilizando apenas as opções *quantize*, que define esses intervalos de forma totalmente uniforme e *quantile*, que define os intervalos segundo os quantis calculados a partir da distribuição dos valores da característica.

Podendo estas funcionalidades pôr em causa a qualidade da representação visual dos

dados decidiu-se explorar outras possibilidades. Foi então testada e, consequentemente, utilizada a ferramenta de visualização Unfolded Studio [34] que é uma extensão da ferramenta Kepler, fornecida através de um modelo *Software as a Service* (SaaS), com uma série de novas funcionalidades, tais como, a possibilidade de sincronizar filtros para conjuntos de dados distintos, aplicar filtros temporais dinâmicos com intervalos de tempo bem definidos, definir intervalos de valores específicos para representar características, entre outras.

Assim sendo, este capítulo tem como principal objetivo representar visualmente os congestionamentos na cidade assim como desenvolver indicadores capazes de resumir de forma simplificada o estado do tráfego em toda a cidade assim como em determinadas zonas de interesse.

5.1 Representação de congestionamentos

Embora já se encontre bem definido o conjunto de dados representativo dos congestionamentos, como se pretende mostrar os congestionamentos segundo o seu tempo de atraso (`delay`) foram acrescentadas duas novas características ao conjunto de dados: (a) `delayMinutes`, que corresponde ao tempo de atraso do tráfego em comparação com a via completamente livre em minutos e segundos; e (b) `nivelCongestionamento`, que corresponde ao nível de congestionamento.

Em relação à característica `delayMinutes`, esta é apenas uma conversão da característica `delay` que se expressa em segundos, para valores expressos em minutos e segundos. Esta nova característica tem como objetivo simplificar a leitura do tempo de atraso do congestionamento por parte do analista, uma vez que a observação de tempos somente em segundos poderá ser pouco natural e, por isso, não ser tão rapidamente assimilada.

Relativamente à característica `nivelCongestionamento`, esta foi incorporada pelo facto de existir a necessidade de ter um escala sobre a qual o congestionamento irá ser categorizado segundo o seu tempo de atraso e consequentemente representado visualmente. A escala definida foi baseada na literatura existente [40] e adaptada à realidade da cidade, podendo apresentar os seguintes valores: (1) Não congestionado, para atrasos entre os [0..5[minutos; (2) Pouco congestionado, para atrasos entre os [5..8[minutos ; (3) Congestionado, para atrasos entre os [8..12[minutos; (4) Muito congestionado, para atrasos entre os [12..20[minutos ; e (5) Parado, para atrasos iguais ou superiores a 20 minutos.

Partindo para a representação gráfica dos congestionamentos e tendo como ponto de partida a representação apresentada anteriormente na Secção 3.5, foi estudado um novo esquema de cores pelo facto de que as cores mais escuras anteriormente utilizadas, sobretudo, o roxo, nem sempre permitirem uma fácil identificação dos congestionamentos mais críticos devido ao seu contraste com o fundo escuro do mapa da cidade. Por outro lado, decidiu-se manter o critério de que as cores mais claras representam congestionamentos menos demorados e as cores mais escuras representam congestionamentos mais demorados, obtendo-se assim uma escala de cores que vai desde o branco (Não congestionado) até ao vermelho (Parado), associando desta forma as cores seleccionadas com os valores da característica `nivelCongestionamento`.

Esta forma de representação aparenta ser adequada na medida em que permanecerá estável para qualquer que seja a distribuição associada aos valores dos tempos de atraso.

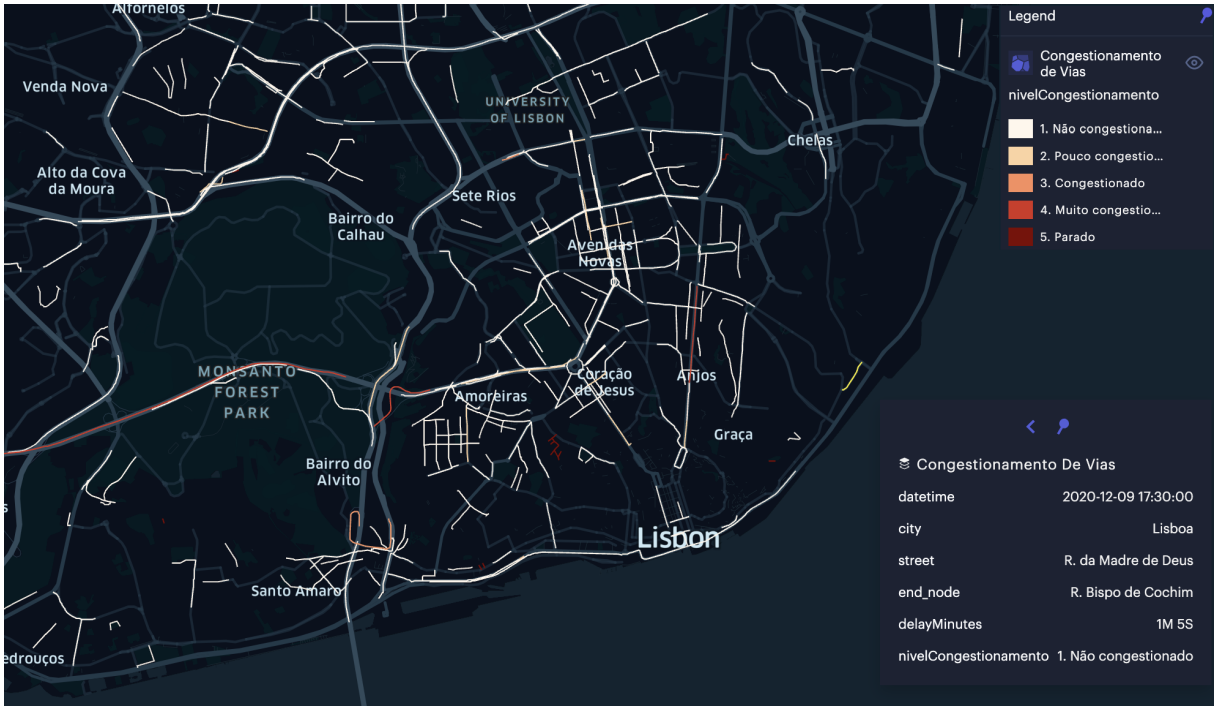


Figura 5.1: Congestionamentos na cidade de Lisboa

Na Figura 5.1 encontram-se representados graficamente todos os congestionamentos observados na cidade de Lisboa no dia 9 de Dezembro de 2020 entre as 17h30 e as 17h59, possibilitando uma identificação imediata de congestionamentos mais críticos, como por exemplo, na autoestrada A5 (troço de via que passa por Monsanto) ou nos acessos à autoestrada do sul (em direção à Ponte 25 de Abril). Observa-se ainda o caso onde o utilizador pretende obter mais informações acerca de um determinado congestionamentos, neste caso, o congestionamento em questão encontra-se no lado

direito da imagem representado com a cor amarela, devido ao clique do utilizador, e através de uma nova janela (canto inferior direito) é possível observar as características presentes no conjunto de dados para o congestionamento em questão.

Neste caso especifico é possível observar, por exemplo, a característica `delayMinutes` que indica que o tempo de atraso deste congestionamento é de 1 minuto e 5 segundos. São também apresentadas as características `datetime`, `city`, `street`, `end_node` e `nivelCongestionamento`. As restantes características presentes no conjunto de dados não estão aqui representadas devido à configuração que foi utilizada, podendo estas ser mostradas ou ocultadas segundo a escolha do utilizador.

5.2 Representação de Indicadores

5.2.1 Formulação do problema

Tornando-se bastante difícil precisar o estado da fluidez de tráfego na cidade tanto de um modo geral como para zonas especificas somente com base na observação dos congestionamentos, principalmente, em horas com um número elevado de congestionamentos. Surge a necessidade da existência de indicadores de fluidez de tráfego que indiquem de forma rápida e clara o estado da fluidez de tráfego na cidade. Estando a resolução temporal mínima dos dados definida em 30 minutos, os indicadores terão de ser capazes de resumir a fluidez do tráfego baseando-se no número de congestionamentos assim como nos tempos de atraso de cada congestionamento observados a cada meia hora.

Os indicadores traduzem-se assim na fórmula (5.1) desenvolvida que multiplica o valor médio dos tempos de atraso dos congestionamentos (`delay`) por um valor normalizado entre 0 e 1, segundo a técnica *Min-Max*, da contagem do número de congestionamentos em cada intervalo de tempo (30 minutos).

Deste modo, e segundo uma breve análise ao comportamento desta fórmula, garante-se que em cenários: (1) onde o número de congestionamentos seja muito baixo ou próximo de zero, independentemente do valor médio dos tempos de atraso dos congestionamentos o valor da fórmula ficará sempre próximo do mínimo. Neste caso em concreto, não se pretendia de todo que o valor da fórmula fosse elevado na medida em que um número baixo de congestionamentos não tem a capacidade de afetar significativamente a fluidez do tráfego na cidade. Este cenário observa-se sobretudo em

períodos noturnos onde normalmente existe um baixo número de congestionamentos na cidade mas com tempos de atrasos elevados; e (2) onde o número de congestionamentos seja elevado ou próximo do seu valor máximo, o valor da fórmula será ajustado segundo o valor médio dos tempos de atraso dos congestionamentos. Este cenário observa-se principalmente nas horas de ponta onde existe um maior número de congestionamentos sendo que o valor médio dos tempos de atraso desses congestionamentos poderá variar bastante.

Por fim, decidiu-se também normalizar os valores da fórmula (5.1) entre 0 e 1, segundo a técnica *Min-Max*, uma vez que se pretende que o indicador seja classificado segundo diferentes níveis de fluidez de tráfego com base na fórmula desenvolvida. Deste modo, torna-se possível categorizar e consequentemente representar visualmente o indicador. A definição da escala dos níveis de fluidez de tráfego sob os quais os indicadores irão ser classificados foi baseada em ferramentas de monitorização de tráfego existentes [36], e resume-se através dos seguintes valores: (1) Livre, para valores entre os $[0..0,2[$; (2) Fluido, para valores entre os $[0,2..0,4[$; (3) Pouco fluido, para valores entre os $[0,4..0,6[$; (4) Denso, para valores entre os $[0,6..0,8[$; e (5) Engarrafamento, para valores entre os $[0,8..1]$.

Para a criação dos indicadores de fluidez de tráfego geral e de zonas, foram desenvolvidos dois novos conjuntos de dados intitulados de *IndicadorGeral* e *IndicadorZonas*, respectivamente, cujas características são as mesmas e podem ser observadas na Tabela 5.1. Embora ambos os conjuntos de dados tenham as mesmas características, decidiu-se não os agregar não só por uma questão de organização mas também porque se poderá querer ocultar individualmente cada um deles na ferramenta de visualização.

5.2.2 Indicador geral

Este indicador apresenta uma resolução espacial que envolve toda a cidade, ou seja, tem em conta todos os congestionamentos observados.

Relativamente à representação gráfica do indicador, foi estudado um esquema de cores que fosse capaz de traduzir o estado da fluidez do tráfego de uma forma o mais natural possível à percepção humana. Pelo que se decidiu utilizar um esquema baseado nas cores de um semáforo uma vez que estas fazem parte do quotidiano da população, obtendo-se assim uma escala de cores que vai desde o verde (Livre) até ao vermelho (Engarrafamento), associando desta forma as cores seleccionadas com os valores da característica `nivelIndicador`.

Tabela 5.1: Descrição das características presentes nas propriedades dos conjuntos de dados *IndicadorGeral* e *IndicadorZonas*.

Nome do campo	**Descrição**
datetime	Data e hora do indicador
zona	Nome da zona a que o indicador se refere
lat	Latitude
lon	Longitude
delayMean	Média dos tempos de atraso dos congestionamentos em segundos
delayMeanMinutes	Média dos tempos de atraso dos congestionamentos em minutos e segundos
numCongestionamentos	Contagem do número de congestionamentos
formula	Valor da aplicação da fórmula $\texttt{delayMean} \times \textit{ValorNormalizado}(\texttt{numCongestionamentos})$ (5.1)
formulaNormalizada	Valor normalizado da característica `formula`
nivelIndicador	Nível do indicador (Livre, Fluido, Pouco fluido, Denso ou Engarrafamento)

Ao nível da localização do indicador no mapa, e sendo este definido por um ponto (latitude e longitude) segundo as características `lat` e `lon`, pretende-se que o indicador não fique em sobreposição com qualquer outro elemento do mapa que possa conter informações. Assim sendo, e tendo em conta o enquadramento geográfico da cidade foi decidido colocar o indicador já fora da cidade, na zona do rio Tejo.

Tendo em conta a área envolvente deste indicador (toda a cidade), foi decidido colocar o valor "Fluidez de Tráfego" na característica `zona` permitindo assim rotular o indicador uma vez que este é um caso especifico.

Na Figura 5.2 encontra-se representado graficamente o indicador geral de fluidez de tráfego calculado para a cidade de Lisboa no dia 9 de Dezembro de 2020 entre as 17h30 e as 17h59, possibilitando ao analista ter uma noção quase imediata da fluidez do tráfego na cidade de um modo geral para o espaço temporal em questão. Observa-se também o caso onde o utilizador pretende obter mais informações acerca do indicador em questão, e através de um clique no indicador, obtém-se uma janela (canto inferior direito) onde é possível verificar mais algumas das restantes características presentes no conjunto de dados para o indicador em questão.

Neste caso é observado, por exemplo, a contagem do número de congestionamentos

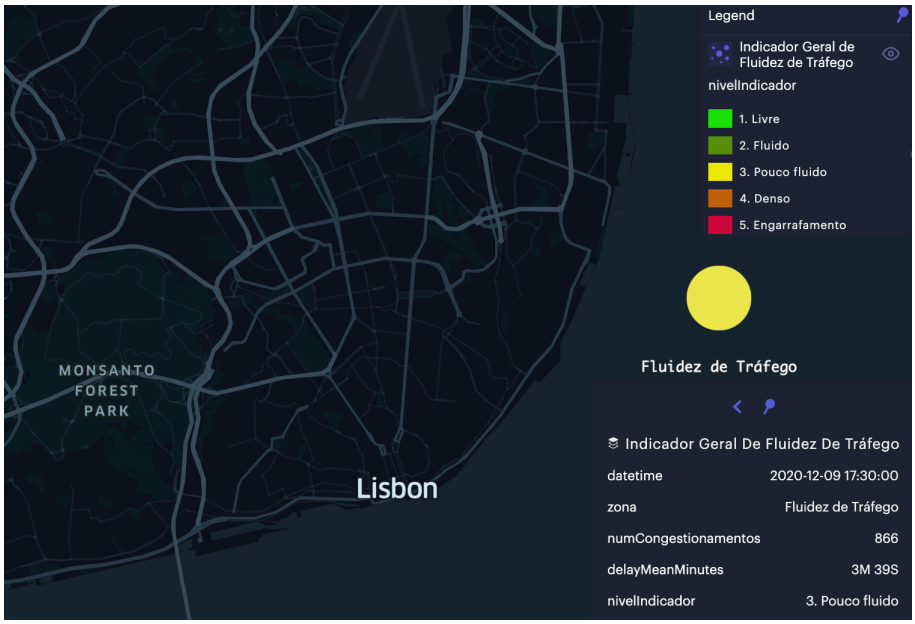


Figura 5.2: Indicador geral de fluidez de tráfego na cidade de Lisboa.

(`numCongestionamentos`) naquela meia hora que apresenta um valor de 866 congestionamentos ou o tempo médio de atraso dos congestionamentos (`delayMeanMinutes`) que apresenta um valor de 3 minutos e 39 segundos. De notar que restantes características presentes no conjunto de dados que não se encontram nesta janela deve-se à configuração escolhida na ferramenta de visualização, existindo sempre a possibilidade de as mostrar.

5.2.3 Indicadores de zonas

Este tipo de indicador apresenta uma resolução espacial que envolve apenas uma determinada área da cidade sendo esta previamente definida através de um conjunto de ruas, ou seja, tem em conta somente os congestionamentos observados no conjunto de ruas predefinido.

Foram então criadas três zonas que correspondem a entradas e saídas da cidade sendo estas a Ponte 25 de Abril, IC19 e Calçada de Carriche. Foi ainda criada uma outra zona que corresponde a uma área mais interior da cidade, Avenidas Novas, pelo que é uma zona bastante movimentada da cidade e demonstra a possibilidade de produzir indicadores de fluidez de tráfego para outras áreas que não apenas as entradas e saídas da cidade. O conjunto de ruas (`street` e `end_node`) utilizado para a criação das zonas encontra-se na Tabela 5.2.

Com base nos congestionamentos de toda a cidade, são filtrados estes conjuntos de

ruas e aplica-se a fórmula (5.1) (Secção 5.2.1) e consequentemente desenvolve-se o indicador para cada uma das zonas individualmente criando assim o conjunto de dados que o representa. Na eventualidade de não existirem congestionamentos numa zona, o valor de (5.1) é colocado a zero e, consequentemente, o indicador apresentará um nível de fluidez de tráfego com o valor 'Livre'.

Tabela 5.2: Conjunto de ruas (`street` e `end_node`) utilizado para a criação de zonas.

Zona	`Street`	`End_node`
Ponte 25 de Abril	Pte. 25 Abril / Viaduto Norte	Pte 25 de Abril
	Ac. Pte. (S), Setúbal, Almada	A2 / IP7
	Ac. Pte. (S), Setúbal, Almada	Ac. Ponte (S), Setúbal, Almada
	Ac. Pte. 25 de Abril, A2 (S)	Lisboa (Campo de Ourique)
	IP7 / Eixo N-S	Lisboa (Alvito)
	Ac. Av. da Ponte, A2 (S), IC20, Almada, Caparica	Lisboa (Br. da Liberdade)
	Pte. 25 de Abril	HGO Almada
	Ac. A2 (S), Alvito	Av. da Pte. / A2 / IP7
IC19	IC19	EMFA, Buraca
	IC19	EMFA (Alfragide)
	IC19	Buraca
	IC19	Lisboa (Bairro da Boavista)
	IC19	Calhariz de Benfica (Lisboa)
	Ac. IC19 Lisboa	Buraca
	Ac. 1, IC19 IC17 CRIL Algés, A5 Cascais, Campismo	Calhariz de Benfica (Lisboa)
	Ac. IC19 Lisboa	EMFA, Buraca
Calçada de Carriche	Cç. de Carriche	Lisboa (Lumiar)
	Cç. de Carriche	Cç. de Carriche
	IC22	Sr. Roubado
	A8	Sr. Roubado
	Estr. do Desvio	Cç. de Carriche
	Av. Padre Cruz	Campo Grande, Lisboa
	Av. Padre Cruz	Lisboa (Lumiar)
	Av. Padre Cruz	Telheiras
	Av. Padre Cruz	Lumiar (Lisboa)
Avenidas Novas	Pç. do Dq. de Saldanha	Av. da República
	Pç. do Dq. de Saldanha	Av. Fontes Pereira de Melo
	Av. da República	Av. Miguel Bombarda
	Av. da República	Pç. do Dq. de Saldanha
	Av. da República (lateral)	Av. Visc. de Valmor
	Av. da República (lateral)	Av. da República (lateral)
	Av. da República	Av. da República (túnel)
	Av. da República (túnel)	Av. da República
	Av. da República (lateral)	Av. da República (central)
	Av. Elias Garcia	Av. da República
	Av. da República	Av. da República
	Av. João Crisóstomo	Av. da República
	Av. Dq. d Ávila	Av. da República
	Av. Elias Garcia	Av. da República (túnel)
	Av. Elias Garcia	R. Arco do Cego
	Av. Barbosa du Bocage	Av. da República (lateral)
	Av. Defensores de Chaves	Av. João XXI
	Av. Visc. de Valmor	Av. da República (lateral)
	Av. João Crisóstomo	R. D.ª Filipa de Vilhena
	Av. 5 de Outubro	Av. das Forças Armadas
	Av. 5 de Outubro	Av. de Berna
	Av. 5 de Outubro	Av. Visc. de Valmor
	Av. 5 de Outubro	Av. Miguel Bombarda
	Av. 5 de Outubro	R. Pinheiro Chagas
	Av. de Berna	Av. João XXI
	Av. de Berna	Pç. de Espanha
	Av. Mq. de Tomar	Av. de Berna

A representação gráfica deste tipo de indicadores é em tudo idêntica à representação gráfica do indicador geral, alterando evidentemente a localização do indicador no mapa (latitude e longitude) com base na área que este representa assim como o nome da zona a que o indicador se refere.

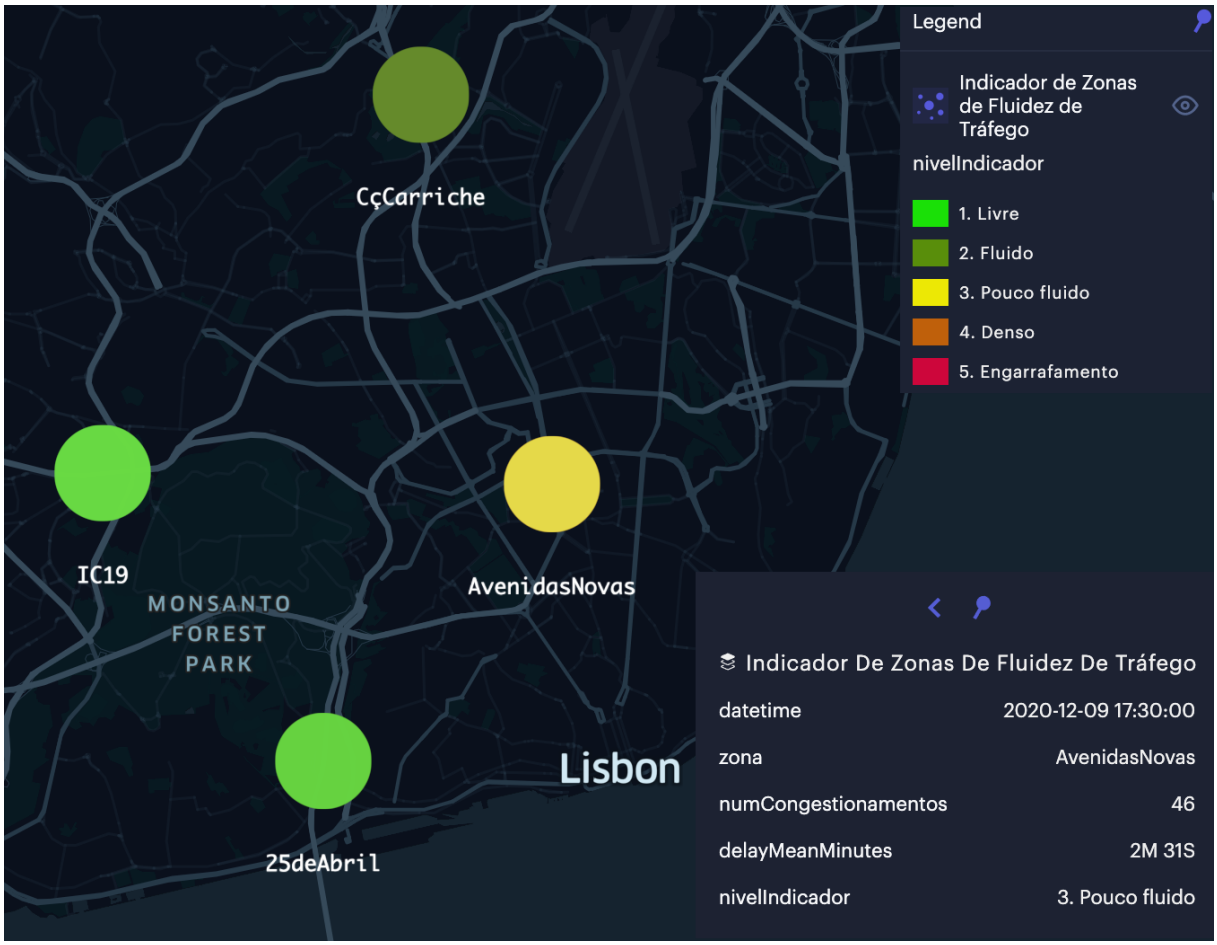


Figura 5.3: Indicador de zonas de fluidez de tráfego na cidade de Lisboa

Na Figura 5.3 encontram-se representados graficamente os indicadores de zonas de fluidez de tráfego calculado para as zonas da Ponte 25 de Abril, IC19, Calçada de Carriche e Avenidas Novas no dia 9 de Dezembro de 2020 entre as 17h30 e as 17h59, possibilitando ao analista ter uma noção quase imediata da fluidez do tráfego em cada zona especifica para o espaço temporal em questão. Tal como para o indicador geral, também nestes indicadores é possível o utilizador obter mais informações acerca dos mesmo, através de um clique onde se obtém uma janela onde é possível verificar mais algumas das restantes características presentes no conjunto de dados como se pode observar no canto inferior direito da figura.

Neste caso é observado para a zona das Avenidas Novas, por exemplo, a contagem do número de congestionamentos (`numCongestionamentos`) naquela meia hora que

apresenta um valor de 46 congestionamentos ou o tempo médio de atraso dos congestionamentos (`delayMeanMinutes`) que apresenta um valor de 2 minutos e 31 segundos. De notar que restantes características presentes no conjunto de dados que não se encontram nesta janela deve-se à configuração escolhida na ferramenta de visualização, existindo sempre a possibilidade de as mostrar.

5.2.4 Indicador de métrica de previsões

Tendo em conta que os dados a representar graficamente podem não só ser dados observados mas também dados previstos segundo um modelo foi ponderada a decisão de dispor de um indicador capaz de representar as métricas de avaliação do modelo em relação às previsões realizadas de modo a transmitir alguma confiança sobre as mesmas ao analista.

Foi então decidido incorporar também esse indicador, neste caso, sobre a forma de contorno (*Outline*) do indicador geral de fluidez de tráfego. Para tal, foram adicionadas as métricas de avaliação R^2, MAE, RMSE e RMSLE como características ao conjunto de dados representativo do indicador geral de fluidez de tráfego e escolhida a métrica R^2 para ser representada graficamente. Esta escolha não se deve a nenhum motivo em concreto pelo que este indicador poderá ser representado por qualquer uma das restantes métricas.

Todas as métricas de avaliação obtidas foram consequência da realização de previsões de tempos de atraso de congestionamentos em intervalos de 30 minutos, separadamente e segundo o Modelo C+ com dados históricos (Secção 4.2) treinado com dados de uma semana.

Os valores da métrica são associados a uma escala de cores que vai desde o azul (R^2 mais baixo) até ao branco (R^2 mais alto) para que não seja provocado um conflito visual com as cores utilizadas no indicador geral de fluidez.

Na Figura 5.4 encontram-se representado graficamente o indicador de métrica de previsão resultante da previsão de congestionamentos realizada para o dia 9 de Dezembro de 2020 entre as 17h30 e as 17h59, possibilitando ao analista ter uma rápida noção da qualidade das previsões efectuadas para o espaço temporal em questão. Neste caso, através do indicador é observada a cor branca que representa um valor de R^2 superior a $0,8$, e analisando essa métrica com maior detalhe através da janela despoletada através de um clique no indicador podemos confirmar que o valor real de R^2 é de $0,93$. Nessa mesma janela, é ainda possível observar os valores das restantes métricas.

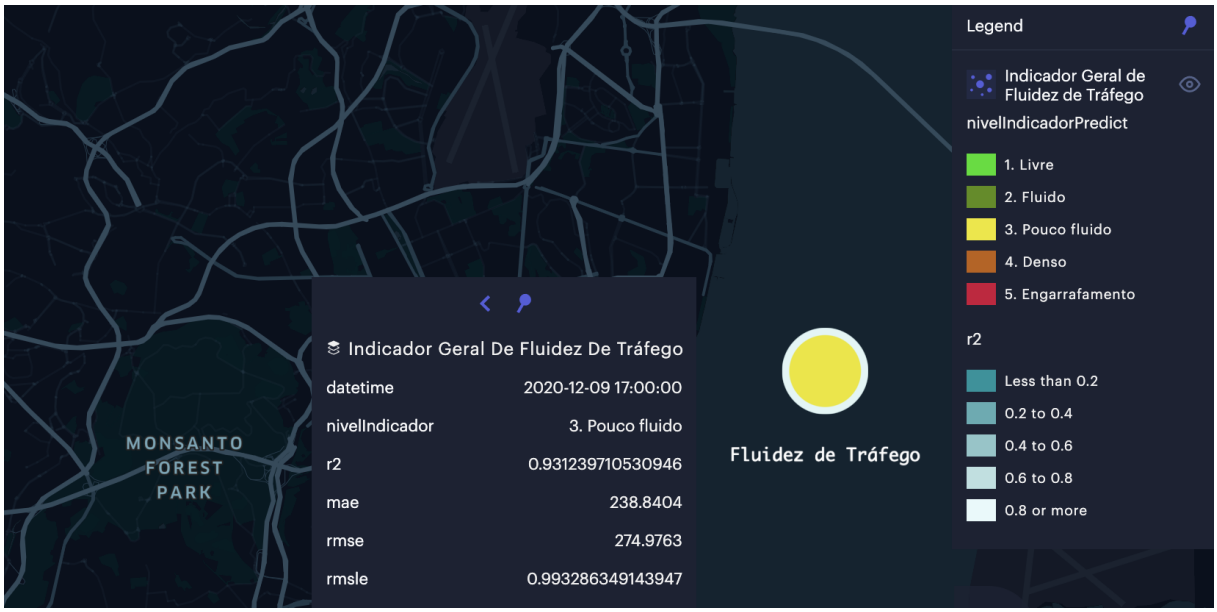


Figura 5.4: Indicador de métrica de previsão

5.3 Apresentação do *dashboard*

Sendo o principal objetivo deste capitulo a obtenção de um *dashboard* interativo, foram integrados na ferramenta de visualização os conjuntos de dados representativos dos congestionamentos e dos indicadores desenvolvidos resultando assim numa possível proposta de *dashboard*.

Na Figura 5.5 encontram-se representados graficamente todos os congestionamentos e os indicadores resultantes da previsão de tempos de atraso de congestionamentos para o dia 11 de Dezembro de 2020 entre as 17h30 e as 17h59 segundo o Modelo C+ com dados históricos (Secção 4.2) treinado com dados da semana anterior. Através de todo este conjunto de indicadores disponíveis, o analista consegue ter uma percepção imediata da previsão da fluidez do tráfego tanto de um modo geral na cidade como para as quatro zonas especificas sob as quais foram desenvolvidos indicadores assim como constatar a segurança que estas previsões transmitem através da métrica R^2. Consegue ainda observar a previsão do nível de congestionamento em ruas especificas e tem a possibilidade de omitir tanto os congestionamentos como os indicadores de modo a só observar o que realmente pretende.

Neste caso em concreto, é possível observar-se diretamente que o indicador geral de fluidez reporta o nível de fluidez "Engarrafamento", ou seja, previu-se que a fluidez de tráfego na cidade estivesse no pior nível de fluidez com um valor de R^2 superior a $0,8$ transmitindo alguma confiança nas previsões. Em relação aos indicadores de

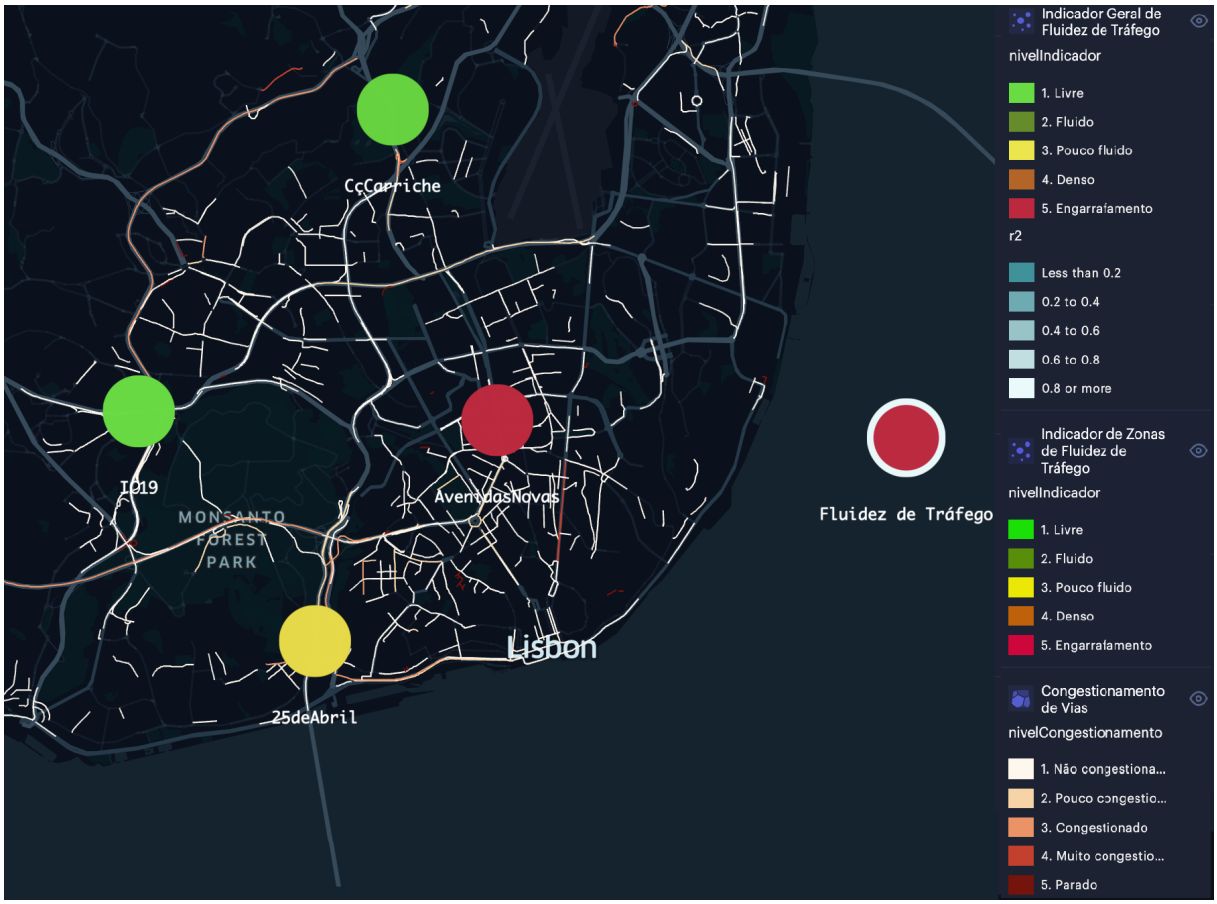


Figura 5.5: *Dashboard* de monitorização de fluidez de tráfego

zonas observam-se os níveis de fluidez de "Pouco fluido", "Livre", "Livre" e "Engarrafamento" para as zonas da Ponte 25 de Abril, IC19, Calçada de Carriche e Avenidas Novas, respectivamente. Em relação aos congestionamentos, observa-se que a maioria está com o nível de congestionamento de "Pouco congestionado" no entanto existem alguns congestionamentos espalhados pela cidade onde se observa o nível de congestionamento de "Parado".

A solução proposta de *dashboard* encontra-se disponível *online* através do *url* `https://tinyurl.com/yjayb59k` onde se observam os dados dos congestionamentos e dos indicadores baseados em previsões de tempos de atraso de congestionamentos realizadas em dados relativos à semana 50 do ano de 2020 segundo o Modelo C+ com dados históricos (Secção 4.2) treinado com dados da semana anterior. Através deste *dashboard* é possível filtrar os dados que se pretende observar no momento, observar os detalhes de cada congestionamento ou indicador em particular, observar as alteração na fluidez do tráfego de forma dinâmica ao longo da semana segundo intervalos de 30 minutos, entre outras.

5.4 Validação da representação dos indicadores

Uma vez desenvolvidos e apresentados os indicadores de fluidez de tráfego surge o interesse de os validar, isto é, perceber se a forma de representar a fluidez de tráfego é adequada e, sobretudo, perceber se a percepção da fluidez de tráfego é correctamente resumida pelos indicadores. Para tal, foi efectuado um teste com utilizadores, com base num questionário desenvolvido através da plataforma *Google Forms* de modo a permitir a validação dos indicadores desenvolvidos através de análises com utilizadores.

O questionário desenvolvido (Anexo B) apresenta um total de 18 perguntas, com uma participação de 29 inquiridos dos quais $51,7\%$ tinham idades compreendidas entre os 15 e os 24 anos, 31% entre os 25 e os 34 anos, $6,9\%$ entre os 35 e os 44 anos e $10,3\%$ entre os 45 e os 54 anos onde $51,7\%$ dos mesmos alega já ter tido contacto com ferramentas de visualização de dados assim como $79,3\%$ afirmam conduzir regularmente, no entanto, apenas $37,9\%$ afirma conduzir regularmente na cidade de Lisboa.

O questionário apresenta três perguntas iniciais de preparação que pretendem que o participante identifique o nível de fluidez de tráfego apresentado pelos indicadores (geral e de zonas) e o nível de congestionamento de uma via, separadamente e apenas como modo de apresentação dos indicadores e ponto de partida para as restantes perguntas.

As perguntas seguintes questionam o utilizador sobre: (1) qual seria o nível adequado para o indicador geral de fluidez de tráfego com base nos indicadores de zonas observados; (2) qual seria o nível adequado para o indicador geral de fluidez de tráfego com base nos níveis de congestionamentos das vias observados; (3) qual o nível de consistência (1 - nada consistente a 5 - bastante consistente) entre os níveis de fluidez de tráfego observado no indicador geral e nos indicadores de zonas; e (4) qual o nível de consistência (1 - nada consistente a 5 - bastante consistente) entre o nível de fluidez de tráfego observado no indicador geral e os níveis de congestionamentos observados nas vias. Para cada uma destas situações existem sempre duas perguntas associadas porém com cenários distintos. Possibilitando assim questionar o utilizador perante um cenário mais simples de resposta e perante um cenário mais complexo, o que permite uma análise mais legitima das respostas para cada tipo de pergunta.

Por fim, pede-se ainda ao utilizador que classifique a forma como os indicadores resumem (1- resume pessimamente a 5 - resume perfeitamente) a fluidez de tráfego ao longo do dia com base num vídeo que apresenta a fluidez de tráfego na cidade de Lisboa no dia 12 de Novembro de 2020 entre as 9 horas da manha e as 9 horas da noite.

5.4.1 Análise de poder estatístico do teste

Foi realizada uma análise de poder estatístico que permitirá estimar o tamanho ideal da amostra para detetar com confiança um dado efeito. A análise foi realizada para o questionário desenvolvido com poderes de teste de 70%, 80% e 90%, tamanho do efeito a variar entre os $0,2$ e os $0,8$, ou seja, pequenos a grandes tamanhos de efeito e um α (nível de significância) de $0,05$.

Na Figura 5.6 encontram-se os resultados desta análise onde podemos observar que para o tamanho da amostra obtida, isto é, 29 inquiridos é possível detetar efeitos de tamanho médio a grande com um poder de teste maior ou igual a 80%, encontrando-se este valor estabelecido universalmente como o valor adequado para o poder de teste. Dependendo do tamanho do efeito ainda é possível alcançar um poder de teste superior tal como se pode observar pela área sombreada a verde, por exemplo, com um tamanho de efeito de aproximadamente de $0,63$ é possível alcançar 90% de poder de teste.

Deste modo, comprova-se que o tamanho da amostra obtida encontra-se adequado estatisticamente para o nível de significância utilizado oferecendo assim uma maior confiança nos resultados obtidos.

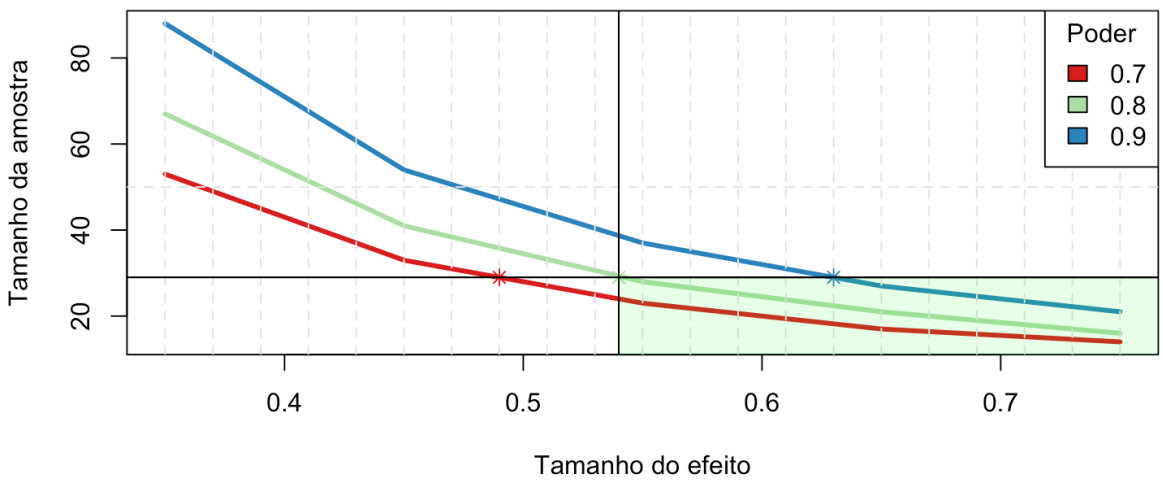


Figura 5.6: Análise de poder estatístico do teste com utilizadores. Com o número de participantes, e verificando-se um efeito médio/alto, consegue-se um poder estatístico adequado, a variar entre $0,8$ e $0,9$, que minimiza os erros de tipo 2.

5.4.2 Análise de respostas

A análise das respostas ao questionário permitirá avaliar se os indicadores apresentados são capazes de transparecer claramente e de forma adequada informação relativa ao estado da fluidez de tráfego com base na taxa de sucesso das respostas.

As três perguntas iniciais de preparação perguntam o nível geral de fluidez de tráfego na cidade, o nível de fluidez de tráfego na zona da Ponte 25 de Abril e o nível de congestionamento do tráfego na Avenida Almirante Reis e apresentam uma percentagem de resposta correta de 89,7%, 96,6% e 82,8%, respectivamente. Estes resultados revelam que a grande maioria dos inquiridos foi capaz de identificar correctamente o nível dos indicadores, permitindo concluir que os níveis dos mesmos e as cores utilizadas para os representar aparentam estar adequadas à analise de um utilizador comum.

Em relação às perguntas 4 e 5, que questionam qual seria o nível adequado para o indicador geral de fluidez de tráfego com base nos indicadores de zonas observados, das quais as respostas corretas são "Livre" e "Pouco fluido", respectivamente, os resultados obtidos podem ser observados na Figura 5.7. Estes mostram que para a pergunta 4 a taxa de respostas corretas é superior a 80% sendo que as restantes respostas indicam o nível "Fluido", que é o nível do indicador imediatamente seguinte ao correto. Para a pergunta 5, a taxa de respostas corretas é superior a 60% sendo que as restantes respostas indicam os níveis "Fluido" e "Denso", que são níveis adjacentes ao correto.

Mediante o exposto, pode-se afirmar que os níveis observados nos indicadores de fluidez de tráfego de zonas permitem, de certo modo, ter uma percepção geral adequada da fluidez de tráfego na cidade.

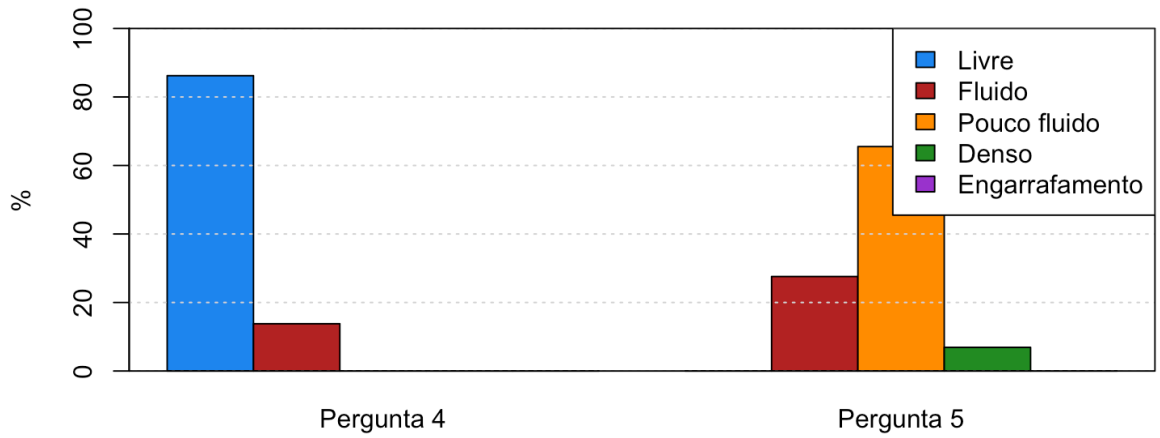


Figura 5.7: Indicador geral de fluidez de tráfego com base nos indicadores de zonas observados

Em relação às perguntas 6 e 7, que questionam qual seria o nível adequado para o indicador geral de fluidez de tráfego com base nos níveis de congestionamentos das vias observados, das quais as respostas corretas são "Engarrafamento" e "Fluido", respectivamente, os resultados obtidos podem ser observados na Figura 5.8. Estes mostram que para a pergunta 6 a taxa de respostas corretas é nula e para a pergunta 7 essa taxa ronda os 45%.

Perante estes resultados, conclui-se que somente com base nos níveis de congestionamentos de vias torna-se difícil ou até mesmo impossível ter uma percepção geral correta da fluidez de tráfego na cidade pelo que se observou sempre uma percentagem mais elevada de respostas incorretas do que corretas.

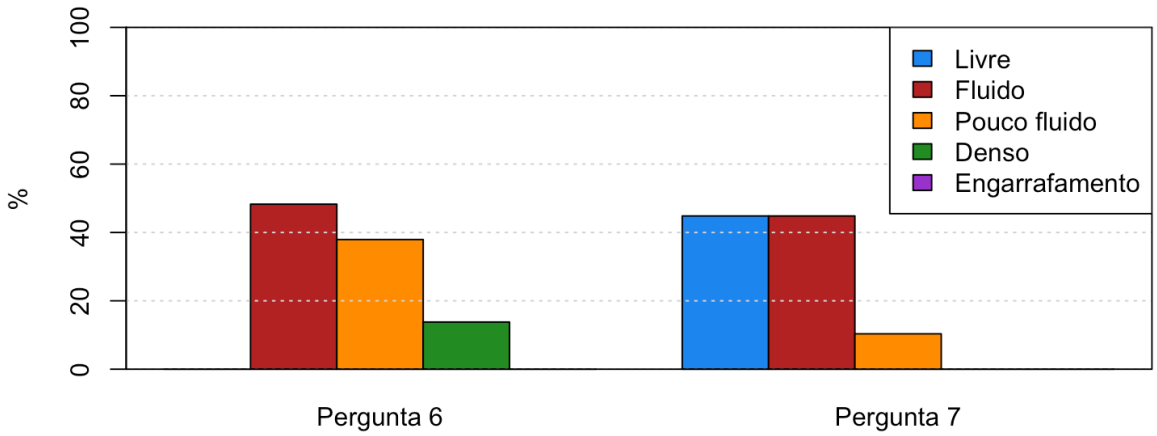


Figura 5.8: Indicador geral de fluidez de tráfego com base nos níveis de congestionamentos das vias observadas

Em relação às perguntas 8 e 9, que questionam qual o nível de consistência entre os níveis de fluidez de tráfego observado no indicador geral e nos indicadores de zonas, os resultados obtidos podem ser observados na Figura 5.9. Estes mostram que para a pergunta 8, mais de 50% dos inquiridos avaliou o nível de consistência com um valor igual ou superior 4, dos quais 24% indicou mesmo um nível de consistência máximo. Para a pergunta 9, cerca de 40% dos inquiridos avaliou o nível de consistência com um valor igual ou superior a 4, dos quais 10% indicou mesmo um nível de consistência máximo. Em comparação com os resultados obtidos na pergunta 8, embora estejam próximos, seria de esperar uma diminuição na avaliação do nível de consistência uma vez que esta pergunta abordava um cenário mais complexo à análise humana. De notar que em ambas as perguntas não existiu nenhuma resposta relativa ao nível de consistência mínimo.

Desta maneira os níveis de fluidez de tráfego dos indicadores geral e de zonas aparentam estar consistentes entre si de um modo geral. Foram calculados os valores de

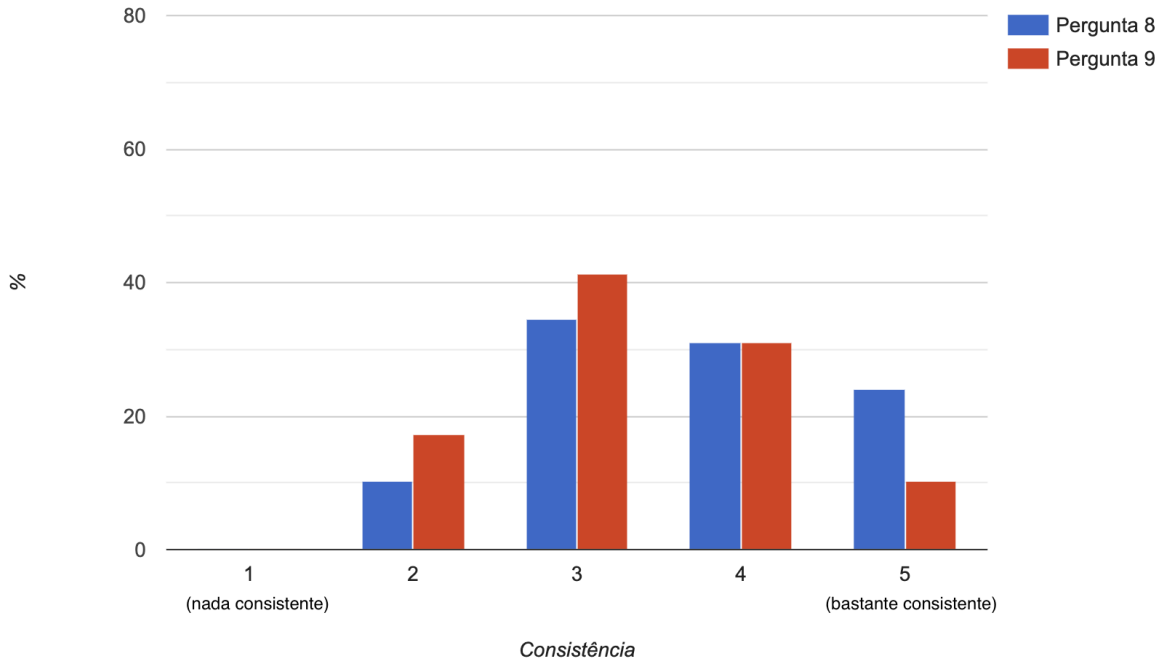


Figura 5.9: Níveis de fluidez de tráfego observado no indicador geral e nos indicadores de zonas

p-value com um nível de significância de 0.05 e assumindo como resposta correta um nível de consistência maior ou igual que 4 , obtendo-se para ambas as perguntas um valor inferior a 0.03.

Em relação às perguntas 10 e 11, que questionam qual o nível de consistência entre o nível de fluidez de tráfego observado no indicador geral e os níveis de congestionamentos observados nas vias, os resultados obtidos podem ser observados na Figura 5.10. Estes mostram que cerca de 45% e de 58% avaliou o nível de consistência com um valor igual ou superior 4, respectivamente. No entanto, na pergunta 10, cerca de 41% dos inquiridos avaliou o nível de consistência com um valor igual ou inferior a 2. De notar ainda que em ambas as perguntas cerca de 7% dos inquiridos avaliou o nível de consistência como nada consistente.

Deste modo, é possível complementar a conclusão da análise das respostas às pergunta 6 e 7 na medida em que é difícil ter uma percepção geral correta da fluidez do tráfego na cidade somente com base nos níveis de congestionamento de vias, o que consequentemente faz com que estes dois indicadores (congestionamento de vias e fluidez de tráfego geral) não se encontrem com um alto nível de consistência, pelo menos à vista humana, pelo que fará todo o sentido incluir o indicador de zonas de fluidez de tráfego como um indicador intermédio entre o indicador de congestionamento de

vias e o indicador geral de fluidez de tráfego. Foram calculados os valores de *p-value* com um nível de significância de 0.05 e assumindo como resposta correta um nível de consistência maior ou igual que 4, obtendo-se para a pergunta 10 e 11 o valor de 0.5 e 0.003, respectivamente.

Em relação à pergunta 12, que questiona quão corretamente os indicadores resumem a fluidez de tráfego ao longo do dia, os resultados obtidos podem ser observados na Figura 5.11. Estes mostram que cerca de 79% dos inquiridos avaliou o resumo da fluidez de tráfego através dos indicadores com um valor igual ou superior 4, não existindo nenhuma resposta relativa ao nível de consistência mínimo e, somente, cerca de 3% dos inquiridos avaliou o nível de consistência com o valor 2. Esta avaliação foi bastante positiva muito possivelmente pelo facto de que nesta pergunta o inquirido teve acesso a todos os indicadores em simultâneo facilitando assim a observação da fluidez de tráfego na cidade. Foi calculado o valor de *p-value* com um nível de significância de 0.05 e assumindo como resposta correta um nível de avaliação maior ou igual que 4, obtendo-se um valor inferior a 0.0005.

Tendo em conta a análise de todas as respostas, em suma, pode-se concluir que é possível identificar corretamente o nível do indicador geral de fluidez de tráfego a partir dos níveis dos indicadores de zonas de fluidez de tráfego sendo que estes demonstram

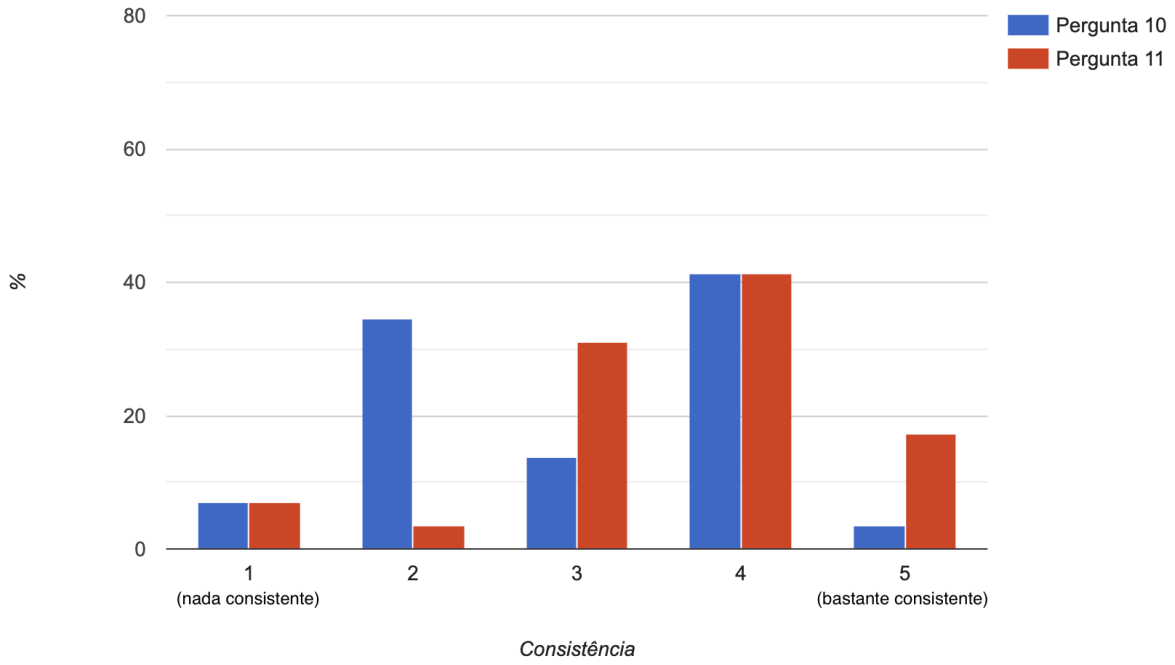


Figura 5.10: Nível de fluidez de tráfego observado no indicador geral e os níveis de congestionamentos observados nas vias

estar com bons níveis de consistência entre si. No entanto, o mesmo não se verifica na identificação do nível do indicador geral de fluidez de tráfego a partir dos níveis de congestionamento observados nas vias, onde se verificou uma maior dificuldade por parte dos inquiridos. Aquando da observação de todos os indicadores em conjunto, os resultados foram bastante satisfatórios e uma vez que, numa situação real de utilização do *dashboard*, o utilizador tem acesso a todos os indicadores pode-se afirmar que a fluidez de tráfego é corretamente resumida pelos mesmos assim como aparenta estar a ser representada de forma adequada uma vez que a grande maioria dos inquiridos foi capaz de identificar corretamente os indicadores e os níveis que estes exibiam.

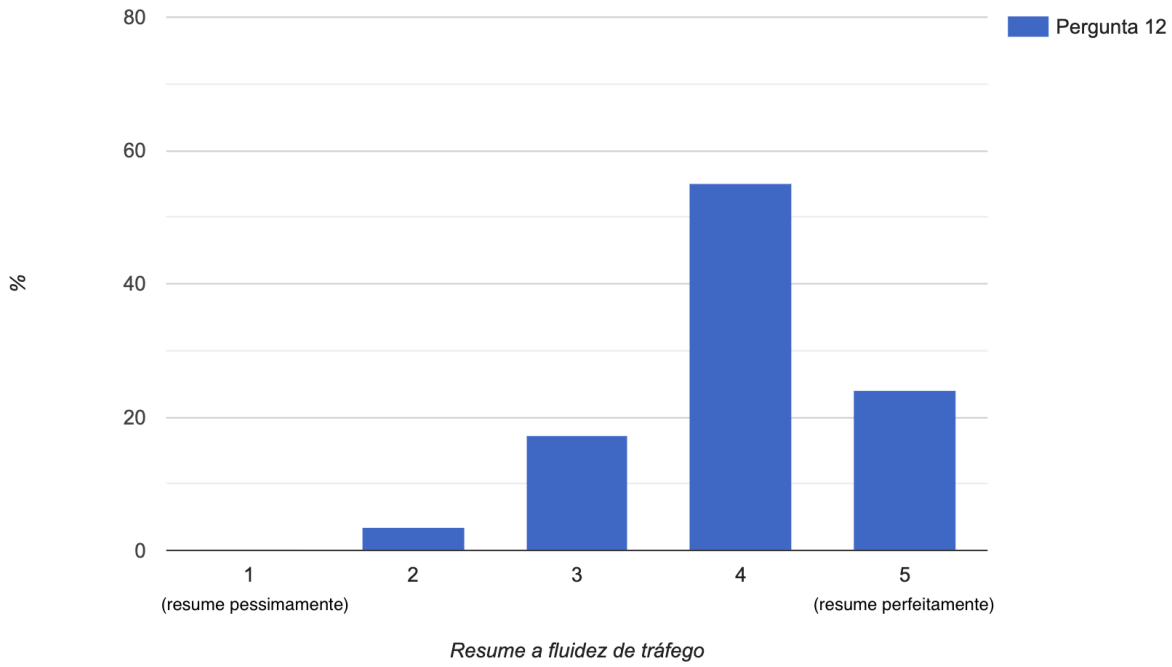


Figura 5.11: Indicador geral de fluidez de tráfego com base nos níveis de congestionamentos das vias observados

6

Conclusões

Com a intensificação do tráfego de veículos nos grandes centros urbanos, com um aumento dos congestionamentos, e com os problemas que estes causam, é cada vez mais necessário disponibilizar metodologias e ferramentas de apoio à decisão aos decisores, quer políticos, quer operacionais.

Suportado pelas transformações tecnológicas que a Câmara Municipal tem vindo a fazer na sua infra-estrutura tecnológica, visando torná-la uma "cidade inteligente", o Laboratório de Dados Urbanos de Lisboa propôs o desafio intitulado "Criação de indicador de tráfego geral e indicadores para cada uma das principais vias de entrada na cidade", o qual serviu como base para o desenvolvimento desta dissertação.

Com base no desafio, propôs-se nesta dissertação desenvolver (i) um modelo preditivo capaz de prever congestionamentos no tráfego da cidade de Lisboa, (ii) focando-se na criação de um indicador de fluidez do tráfego em zonas chave da cidade, e (iii) propondo uma representação visual desse indicador num mapa da cidade através num um *dashboard* interactivo.

6.1 Conclusões

A realização desta dissertação cumpriu todos os objetivos propostos. Foi desenvolvido um modelo de previsão com base em dados relativos a congestionamentos de tráfego e em dados meteorológicos fornecidos pelo Waze e IPMA, respectivamente. Este modelo

mostrou-se capaz de estimar o tempo de atrasado de congestionamentos com níveis de desempenho e erro adequados ao contexto do problema. Tanto quanto se sabe, é um dos poucos trabalhos que estima o tempo de atraso, com uma abordagem única ao tratamento da variável dependente — `delay`.

Foram desenvolvidos indicadores de fluidez de tráfego capazes de resumir o estado da fluidez de tráfego numa dada área geográfica e num determinado espaço temporal. Um dos indicadores desenvolvidos destina-se a resumir a fluidez do tráfego em zonas especificas da cidade, concretizado para as zonas da Ponte 25 de Abril, IC19, Calçada de Carriche e Avenidas Novas. Outro indicador é geral para a cidade e mostra, de forma concisa, a fluidez de tráfego. Todos os indicadores representam um período temporal de 30 minutos. Os testes efectuados durante o desenvolvimento mostram que os indicadores conseguem resumir, numa escala *likert*, de forma adequada o nível de fluidez do tráfego.

Foi proposta uma representação visual do tráfego na cidade de Lisboa através de um *dashboard* interativo sobre um mapa, com base nos resultados esperados do desafio proposto pela CML e nas necessidades mais relevantes dos seus analistas. Este *dashboard* inclui os indicadores desenvolvidos, tem a capacidade de apresentar dados observados assim como dados previstos segundo o modelo desenvolvido e permite ao utilizador interagir com toda a informação disponibilizada. Foi realizado um teste com utilizadores para validação dos indicadores desenvolvidos, sob a forma de questionário, o que permitiu aferir que os indicadores se encontram representados de forma simples e cuidada e possuem a capacidade de resumir corretamente o fluxo de tráfego. Os resultados do teste mostram que a representação dos indicadores em mapa permite aos utilizadores terem a percepção correta do estado da fluidez do tráfego, na maioria da situações.

6.2 Publicações

Como resultado do trabalho desenvolvido, foi escrito e apresentado um artigo no 12º simpósio de informática — INForum 2021, intitulado "Mobilidade urbana sustentável: plataforma inteligente de monitorização" [35].

6.3 Trabalho futuro

Relativamente ao que poderá ser desenvolvido como trabalho futuro, sugere-se a automatização do sistema proposto como solução na medida em que este, para além da

recolha contínua de dados, realização de pré-processamentos associados e previsão de congestionamentos, possa também construir novos modelos de previsão caso exista essa necessidade. Por exemplo, por desvios na qualidade do modelo. Por fim, atualizar dinamicamente o *dashboard* com a nova informação.

Embora os resultados obtidos, relativamente aos modelos de previsão desenvolvidos sejam bastante positivos, sugere-se que sejam incluídas mais fontes de dados para integrar o conjunto de dados utilizado na construção dos mesmos. Por exemplo, integrar informação sobre o tipo de ruas. Além disso, reforça-se que a fonte de dados WAZE reporta informação baseada somente nos seus utilizadores, que pode estar incompleta para algumas zonas da cidade pelo que é sugerido integrar mais dados de tráfego que as próprias cidades possam eventualmente ter acesso.

Ao nível dos indicadores de fluidez de tráfego é sugerido a criação de mais zonas específicas, nomeadamente, mais entradas e saídas da cidade para o desenvolvimento de indicadores o que permitiria aumentar o nível de detalhe do ponto de vista da análise à fluidez do tráfego. Quanto à visualização de dados, seria bastante interessante realizar uma validação do indicador de métrica de previsão sendo que este não foi avaliado no questionário realizado.

Por último espera-se que o *dashboard* desenvolvido possa vir a ser integrado e utilizado na CML e posteriormente contribua para a deteção e mitigação de congestionamentos na cidade assim como para o desenvolvimento de uma cidade mais sustentável.

Referências

[1] *5t*. URL: http://www.5t.torino.it/en/.

[2] *Arome*. URL: https://www.ipma.pt/pt/enciclopedia/otempo/previsao.numerica/index.html?page=arome.xml.

[3] João Miguel Branco de Brito, "Caracterização da flutuação do tráfego na cidade de lisboa", Tese de Doutoramento, Faculdade de Ciências e Tecnologia, 2012. URL: https://run.unl.pt/bitstream/10362/8436/1/Brito_2012.pdf.

[4] Joost CF de Winter, Samuel D Gosling & Jeff Potter, "Comparing the pearson and spearman correlation coefficients across distributions and sample sizes: A tutorial using simulations and empirical data.", *Psychological methods*, vol. 21, n.º 3, pág. 273, 2016.

[5] Lisboa Inteligente. URL: https://lisboainteligente.cm-lisboa.pt/lxdatalab/desafios/criacao-indicador-de-trafego-geral-e-indicadores-para-cada-uma-das-principais-vias-de-entrada-na-cidade.

[6] Ilidia Pinto, *Nunca houve tantos carros e tão envelhecidos a circular em portugal*, https://www.dn.pt/dinheiro/nunca-houve-tantos-carros-e-tao-envelhecidos-a-circular-em-portugal-11248374.html, Diário de Noticias, 2019.

[7] *Distributed random forest*. URL: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/drf.html.

[8] *Ecmwf*. URL: https://www.ecmwf.int/en/forecasts.

[9] Emel, *Emel open data*. URL: https://emel.city-platform.com/opendata/.

[10] Jerome Friedman, Trevor Hastie, Robert Tibshirani et al., *The elements of statistical learning*, 10. Springer series in statistics New York, 2001, vol. 1.

[11] *Geojson*. URL: https://geojson.org/.

[12] *H2o*. URL: https://www.h2o.ai/.

[13] *H2o flow*. URL: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/flow.html.

[14] IPMA, *Ipma api*. URL: http://api.ipma.pt/#services.

[15] ISO 3166-1, *Norma iso 3166-1*. URL: http://en.wikipedia.org/wiki/ISO_3166-1.

[16] *Json*. URL: https://www.json.org/json-en.html.

[17] Federico Karagulian, Claudio A. Belis, Carlos Francisco C. Dora, Annette M. Prüss-Ustün, Sophie Bonjour, Heather Adair-Rohani & Markus Amann, "Contributions to cities' ambient particulate matter (PM): A systematic review of local source contributions at global level", *Atmospheric Environment*, vol. 120, páginas 475 –483, 2015, ISSN: 1352-2310. URL: http://www.sciencedirect.com/science/article/pii/S1352231015303320.

[18] *Kepler*. URL: https://kepler.gl/.

[19] Aleksander Konior, Krzysztof Brzozowski, Andrzej Maczyński & Artur Ryguła, "A concept of extension of the ondynamic system with module for monitoring road traffic impact on the urban environment", *Archives of Transport System Telematics*, vol. 9, 2016. URL: http://yadda.icm.edu.pl/baztech/element/bwmeta1.element.baztech-0a49c85c-ed0f-4df9-a1a9-d7d71b19f85b.

[20] *Leaflet*. URL: https://leafletjs.com/.

[21] Câmera Municipal de Lisboa, "Economia de lisboa em números", 2019. URL: https://issuu.com/camara_municipal_lisboa/docs/economia_de_lisboa_em_numeros__2019.

[22] Haidar Nadrian, Hassan Mahmoodi, Mohammad Hossein Taghdisi, Mehran Aghemiri, Towhid Babazadeh, Bahjat Ansari & Asaad Fathipour, "Public health impacts of urban traffic jam in sanandaj, iran: A case study with mixed-method design", *Journal of Transport & Health*, vol. 19, pág. 100 923, 2020, ISSN: 2214-1405. URL: https://www.sciencedirect.com/science/article/pii/S2214140520301274.

[23] Ming Ni, Qing He & Jing Gao, "Using social media to predict traffic flow under special event conditions", em *The 93rd annual meeting of transportation research board*, 2014.

[24] *Node-red*. URL: https://nodered.org/.

[25] Karina Mary Paiva, Maria Regina Alves Cardoso & Paulo Henrique Trombetta Zannin, "Exposure to road traffic noise: Annoyance, perception and associated factors among brazil's adult population", *Science of The Total Environment*, vol. 650, páginas 978–986, 2019, ISSN: 0048-9697. URL: https://www.sciencedirect.com/science/article/pii/S0048969718334594.

[26] Matteo Picozzi, Nervo Verdezoto, Matti Pouke, Jarkko Vatjus-Anttila & Aaron John Quigley, "Traffic visualization-applying information visualization techniques to enhance traffic planning", em *GRAPP 2013 IVAPP 2013-Proceedings of the International Conference on Computer Graphics Theory and Applications and International Conference on Information Visualization Theory and Applications*, SciTePress, 2013. URL: https://research-repository.st-andrews.ac.uk/handle/10023/8828.

[27] Miriam Pirra & Marco Diana, "Integrating mobility data sources to define and quantify a vehicle-level congestion indicator: An application for the city of turin", *European transport research review*, vol. 11, n.º 1, pág. 41, 2019. URL: https://link.springer.com/article/10.1186/s12544-019-0378-0.

[28] World Health Organization. (2016). "9 out of 10 people worldwide breathe polluted air", URL: https://www.who.int/news-room/air-pollution.

[29] M. FRANCIS PORTELA, *Veículos com mais de 10 anos fazem 90% da poluição automóvel*, 2018. URL: https://www.motor24.pt/motores/ecologia/veiculos-10-anos-fazem-90-da-poluicao-automovel/399311/.

[30] Paulo Homem, *Parque automóvel de portugal supera os 6,2 milhões de veículos*, https://www.dn.pt/dinheiro/nunca-houve-tantos-carros-e-tao-envelhecidos-a-circular-em-portugal-11248374.html, Pós-venda, set. de 2019.

[31] *Rstudio*. URL: https://www.sqlite.org/index.html.

[32] B. Shneiderman, "The eyes have it: A task by data type taxonomy for information visualizations", em *Proceedings 1996 IEEE Symposium on Visual Languages*, 1996, páginas 336–343. DOI: 10.1109/VL.1996.545307.

[33] *Sqlite*. URL: https://www.sqlite.org/index.html.

[34] *Unfolded studio*. URL: https://studio.unfolded.ai/home.

[35] João Vaz, Nuno Datia & Matilde Pós-de Mina Pato, "Mobilidade urbana sustentável: Plataforma inteligente de monitorizaçao", *Inforum-Simpósio de Informática*, páginas 1–12, 2021.

[36] *Viamichelin*. URL: https://www.viamichelin.pt/web/Trafego.

[37] *Traffic data*, jan. de 2021. URL: https://wazeopedia.waze.com/wiki/USA/Traffic_data.

[38] *Xgboost*. URL: https://docs.h2o.ai/h2o/latest-stable/h2o-docs/data-science/xgboost.html.

[39] Da Zhang & Mansur R Kabuka, "Combining weather condition data to predict traffic flow: A gru-based deep learning approach", *IET Intelligent Transport Systems*, vol. 12, n.º 7, páginas 578–585, 2018.

[40] Pengjun Zhao & Haoyu Hu, "Geographical patterns of traffic congestion in growing megacities: Big data analytics from beijing", *Cities*, vol. 92, páginas 164–174, 2019, ISSN: 0264-2751. DOI: https://doi.org/10.1016/j.cities.2019.03.022. URL: https://www.sciencedirect.com/science/article/pii/S0264275119301891.

A

Conjunto de dados utilizado para a modelação

Tabela A.1: Descrição das características presentes nas propriedades do conjunto de dados utilizado na construção do Modelo C+ com dados históricos.

Nome do campo	**Descrição**
datetime	Data e hora do congestionamento
city	Nome da cidade ou estado
street	Nome da rua
end_node	Saída mais próxima do final do congestionamento
road_type	Tipo de via (1 = Rua, 2 = Rua principal, 3 = Auto-estrada, 4= Rampa, 5 = Caminho, 6 = Rua principal, 7 = Rua secundária, 8 = Caminho de terra batida, 9 = Passeio, 10 = Caminho pedestre, 11 = Saída, 14 = Caminho de terra batida, 15 = Travessia de barco, 16 = Escadas, 17 = Caminho privado, 18 = Caminho de ferro, 19 = Faixa exclusiva (taxi/bus), 20 = Via de acesso ou dentro de parque de estacionamento, 21 = Estrada de serviço)
delay	Tempo de atraso do tráfego em comparação com a via completamente livre em segundos (-1 = via completamente congestionada)
position	Conjunto de coordenadas que representam o local do congestionamento (Objecto *GeoJSON*)

continua na próxima página

Nome do campo	Descrição
minIntensidadeVentoKM	Valor mínimo de intensidade do vento registada a 10 metro de altura em kilometros por hora das três estações meteorológicas de Lisboa
maxIntensidadeVentoKM	Valor máximo de intensidade do vento registada a 10 metro de altura em kilometros por hora das três estações meteorológicas de Lisboa
mediaIntensidadeVentoKM	Valor médio de intensidade do vento registada a 10 metro de altura em kilometros por hora das três estações meteorológicas de Lisboa
minTemperatura	Valor mínimo da média da temperatura do ar registada a 1.5 metros de altura numa hora em graus centígrados das três estações meteorológicas de Lisboa
maxTemperatura	Valor máximo da média da temperatura do ar registada a 1.5 metros de altura numa hora em graus centígrados das três estações meteorológicas de Lisboa
mediaTemperatura	Valor médio da média da temperatura do ar registada a 1.5 metros de altura numa hora em graus centígrados das três estações meteorológicas de Lisboa
minPrecAcumulada	Valor mínimo do valor acumulado da precipitação registada a 1.5 metros de altura numa hora em milímetros das três estações meteorológicas de Lisboa
maxPrecAcumulada	Valor máximo do valor acumulado da precipitação registada a 1.5 metros de altura numa hora em milímetros das três estações meteorológicas de Lisboa
mediaPrecAcumulada	Valor médio do valor acumulado da precipitação registada a 1.5 metros de altura numa hora em milímetros das três estações meteorológicas de Lisboa
feriadoFimSemana	Indica se o congestionamento ocorreu num feriado e/ou num fim de semana
vesperaFeriadoFimSemana	Indica se o congestionamento ocorreu na véspera de um feriado e/ou na véspera de um fim de semana
diaDaSemana	Indica qual o dia da semana (por extenso) em que o congestionamento ocorreu
horaDePonta	Indica se o congestionamento ocorreu durante as horas de ponta (7h/10h e 18h/20h)
horaDoDia	Indica em que parte do dia ocorreu o congestionamento (manhã, tarde ou noite)
horas	Horas extraídas da característica `datetime`
minutos	Minutos extraídas da característica `datetime`
ano	Ano extraído da característica `datetime`

continua na próxima página

Nome do campo	**Descrição**
mes	Mês extraído da característica `datetime`
dia	Dia extraído da característica `datetime`
speed{1,2,3}_median	Mediana da característica `speed` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
speed{1,2,3}_mean	Média da característica `speed` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
speed{1,2,3}_max	Máximo da característica `speed` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
level{1,2,3}_median	Mediana da característica `level` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
level{1,2,3}_mean	Média da característica `level` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
level{1,2,3}_max	Máximo da característica `level` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
length{1,2,3}_median	Mediana da característica `length` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
length{1,2,3}_mean	Média da característica `length` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
length{1,2,3}_max	Máximo da característica `length` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
delay{1,2,3}_median	Mediana da característica `delay` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
delay{1,2,3}_mean	Média da característica `delay` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.
delay{1,2,3}_max	Máximo da característica `delay` nos últimos: (1) 30 minutos, (2) 1 hora e (3) 1 hora e 30 minutos.

B

Questionário: Indicador de Tráfego

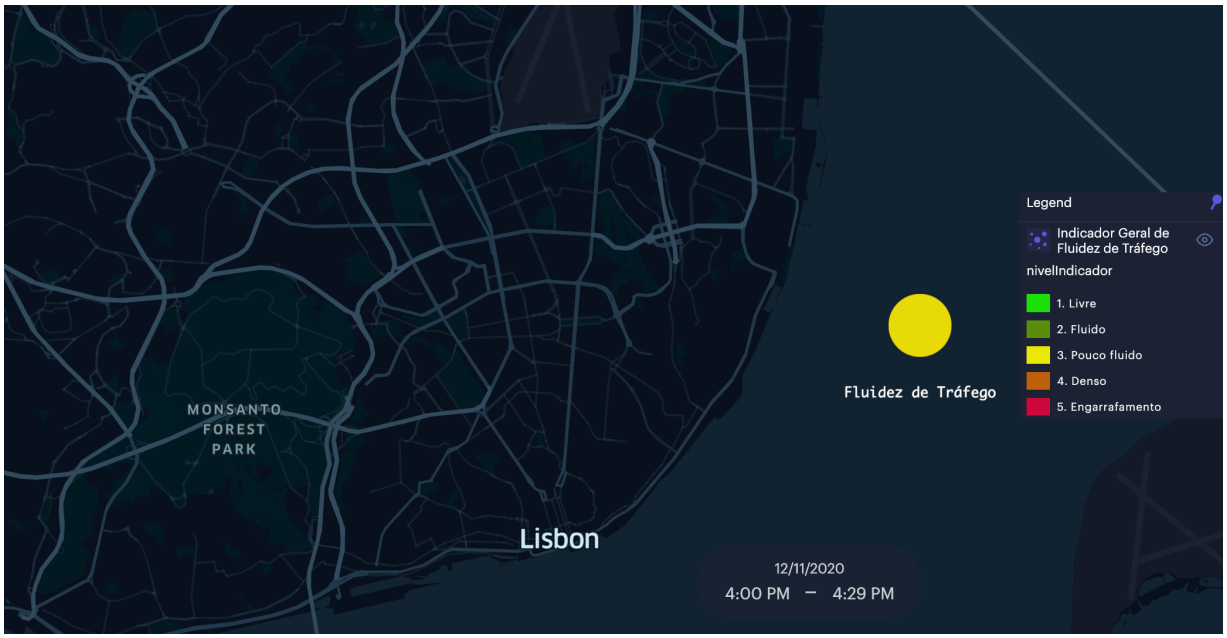
Questionário - Indicador de Tráfego

Este questionário insere-se no âmbito do desenvolvimento de uma Tese de Mestrado em Engenharia Informática e de Computadores intitulada de “Indicador de tráfego: descoberta de padrões na cidade de Lisboa” - ISEL.
Este questionário tem uma duração de aproximadamente 8 minutos.
Todas as suas respostas serão tratadas de forma totalmente anónima.

Obrigado pela sua participação.
João Vaz

*Obrigatório

Indicador Geral de Fluidez de Tráfego para a cidade de Lisboa dia 12/11/2020 entre as 16h e as 16h29.



1. Com base na figura anterior, qual o nível geral de fluidez de tráfego na cidade? *

Marcar apenas uma oval.

- Livre
- Fluido
- Pouco fluido
- Denso
- Engarrafamento

Indicador de Zonas de Fluidez de Tráfego para a cidade de Lisboa dia 12/11/2020 entre as 16h e as 16h29.

2. Com base na figura anterior, qual o nível de fluidez de tráfego na zona da Ponte 25 de Abril? *

Marcar apenas uma oval.

- ◯ Livre
- ◯ Fluido
- ◯ Pouco fluido
- ◯ Denso
- ◯ Engarrafamento

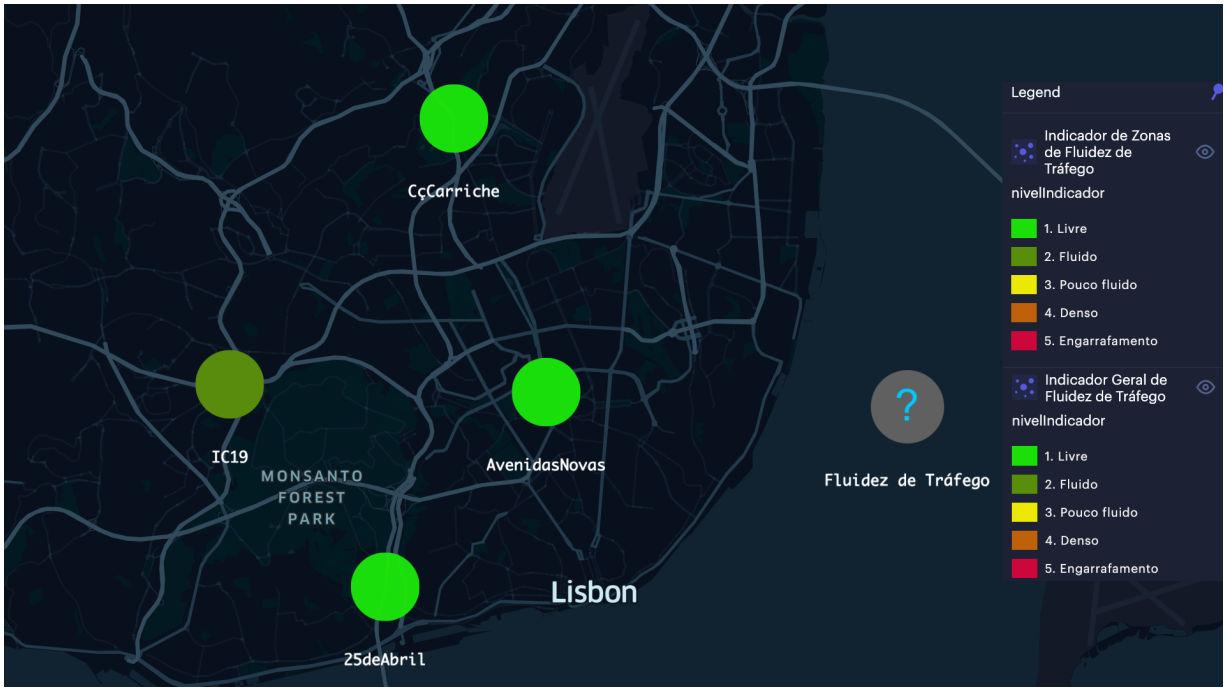
Nível de Congestionamento de Vias na cidade de Lisboa dia 12/11/2020 entre as 16h e as 16h29.

3. Com base na figura anterior, qual o nível de congestionamento do tráfego na Avenida Almirante Reis (indicada no mapa com uma seta azul)? *

Marcar apenas uma oval.

- Não congestionado
- Pouco congestionado
- Congestionado
- Muito congestionado
- Parado

Indicador de Zonas vs Global para a Fluidez de Tráfego



4. Com base nos indicadores de zonas observados na figura anterior, qual acha que seria um nível adequado para o indicador geral de fluidez de tráfego na cidade? *

Marcar apenas uma oval.

- ◯ Livre
- ◯ Fluido
- ◯ Pouco fluido
- ◯ Denso
- ◯ Engarrafamento

Indicador de Zonas vs Global para a Fluidez de Tráfego

5. Com base nos indicadores de zonas observados na figura anterior, qual acha que seria um nível adequado para o indicador geral de fluidez de tráfego na cidade? *

Marcar apenas uma oval.

- ◯ Livre
- ◯ Fluido
- ◯ Pouco fluido
- ◯ Denso
- ◯ Engarrafamento

Nível de Congestionamento de Vias vs Indicador Global de Fluidez de Tráfego

6. Com base no nível de congestionamento das vias observadas na figura anterior, qual acha que seria um nível adequado para o indicador geral de fluidez de tráfego na cidade? *

Marcar apenas uma oval.

- ◯ Livre
- ◯ Fluido
- ◯ Pouco fluido
- ◯ Denso
- ◯ Engarrafamento

Nível de Congestionamento de Vias vs Indicador Global de Fluidez de Tráfego

7. Com base no nível de congestionamento das vias observadas na figura anterior, qual acha que seria um nível adequado para o indicador geral de fluidez de tráfego na cidade? *

Marcar apenas uma oval.

- ◯ Livre
- ◯ Fluido
- ◯ Pouco fluido
- ◯ Denso
- ◯ Engarrafamento

Consistência entre Indicador Geral e de Zonas de Fluidez de Tráfego

8. Com base na figura anterior, classifique a consistência do nível de fluidez apresentado no Indicador Geral (Pouco fluido) tendo somente em conta os níveis de fluidez apresentados nos Indicadores de Zonas numa escala de 1 (nada consistente) a 5 (bastante consistente). *

Marcar apenas uma oval.

1	2	3	4	5
◯	◯	◯	◯	◯

Consistência entre Indicador Geral e de Zonas de Fluidez de Tráfego

9. Com base na figura anterior, classifique a consistência do nível de fluidez apresentado no Indicador Geral (Fluido) tendo somente em conta os níveis de fluidez apresentados nos Indicadores de Zonas numa escala de 1 (nada consistente) a 5 (bastante consistente). *

Marcar apenas uma oval.

1	2	3	4	5
◯	◯	◯	◯	◯

Consistência entre Indicador Geral de Fluidez de Tráfego e Nível de Congestionamento de Vias

10. Com base na figura anterior, classifique a consistência do nível de fluidez apresentado no Indicador Geral (Engarrafamento) tendo somente em conta os níveis de congestionamento das vias apresentadas numa escala de 1 (nada consistente) a 5 (bastante consistente). *

Marcar apenas uma oval.

1	2	3	4	5
◯	◯	◯	◯	◯

Consistência entre Indicador Geral de Fluidez de Tráfego e Nível de Congestionamento de Vias

11. Com base na figura anterior, classifique a consistência do nível de fluidez apresentado no Indicador Geral (Denso) tendo somente em conta os níveis de congestionamento das vias apresentadas numa escala de 1 (nada consistente) a 5 (bastante consistente). *

Marcar apenas uma oval.

1	2	3	4	5
◯	◯	◯	◯	◯

Fluidez de Tráfego na cidade de Lisboa (12/11/2020)

http://youtube.com/watch?v=udxAxzIiNQo

12. Com base no video anterior, classifique a forma como os indicadores resumem a fluidez do tráfego ao longo do dia numa escala de 1 (resume pessimamente) a 5 (resume perfeitamente). *

Marcar apenas uma oval.

1	2	3	4	5
◯	◯	◯	◯	◯

13. Caso pretenda deixar algum comentário ou sugestão sobre os indicadores de fluidez de tráfego observados pode fazê-lo aqui.

14. Já teve alguma experiência a lidar com ferramentas de visualização de dados? *

Marcar apenas uma oval.

- Sim
- Não

15. Conduz regularmente? *

Marcar apenas uma oval.

- Sim
- Não

16. Conduz regularmente na cidade de Lisboa? *

Caso a resposta anterior seja "Não", esta pergunta será considerada como "Não".

Marcar apenas uma oval.

- Sim
- Não

17. Quão bem conhece as principais ruas/zonas da cidade de Lisboa onde foram apresentados os indicadores nas imagens anteriores numa escala de 1 (Não conheço) a 5 (Conheço perfeitamente)? *

Marcar apenas uma oval.

1	2	3	4	5
◯	◯	◯	◯	◯

18. Qual a sua idade? *

Marcar apenas uma oval.

- ◯ 15-24
- ◯ 25-34
- ◯ 35-44
- ◯ 45-54
- ◯ 55+