Sérgio Gonçalo Fernandes Brígida

Licenciado em Ciências da Engenharia Eletrotécnica e de Computadores

Caracterização e Previsão do Tráfego Rodoviário

Dissertação para obtenção do Grau de Mestre em

Engenharia Eletrotécnica e de Computadores

Orientador: Ricardo Luís Rosa Jardim Gonçalves, Professor, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

Co-orientador: Rúben Duarte Dias da Costa, Professor, Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Setembro, 2018

Caracterização e Previsão do Tráfego Rodoviário

Este documento foi gerado utilizando o processador (pdf)LaTeX, com base no template "novathesis" [1] desenvolvido no Dep. Informática da FCT-NOVA [2].
[1] https://github.com/joaomlourenco/novathesis [2] http://www.di.fct.unl.pt

Dedicatória aos meus pais,
António e Maria Rogado,

ao meu irmão,
André Brígida,

e à minha namorada,
Ana Cláudia Lopes.

Agradecimentos

Para a realização do trabalho apresentado foi fundamental o contributo, a disponibilidade e o incentivo por parte de algumas pessoas, a quem pretendo expressar o meu apreço e gratidão.

Gostaria de começar por agradecer ao meu orientador da dissertação, o Professor Ricardo Gonçalves, por me ter dado a oportunidade de trabalhar num projeto do Group for Research in Interoperability of Systems (GRIS) e por toda a disponibilidade demonstrada ao longo da presente dissertação.

Ao meu co-orientador, o Professor e Investigador Sénior no instituto UNINOVA, Rúben Costa, gostaria de deixar um agradecimento especial por me ter assistido de forma incansável em todas as etapas deste projeto. Gostaria também de deixar um agradecimento à sua equipa, particularmente ao Paulo Alves Figueiras e ao Guilherme Guerreiro, pelas rápidas e eficazes orientações prestadas ao longo de todo este projeto.

Quero também agradecer em geral à Faculdade de Ciências e Tecnologia da Universidade Nova de Lisboa, e principalmente ao departamento de Engenharia Eletrotécnica por todo o conhecimento transmitido ao longo do meu percurso académico.

Aos meus amigos "Los Charnéqueros", gostaria de deixar um agradecimento enorme por todos os momentos passados e experiências partilhadas, que considero importantíssimos para finalizar esta etapa da minha vida. A eles peço que me desculpem pelas minhas ausências em alguns momentos.

À minha família, em especial aos meus pais e ao meu irmão, quero agradecer por serem um importante pilar na minha vida, por terem feito de mim o Homem que hoje sou e por me proporcionarem todas as condições para realizar e concluir esta etapa na minha vida. A eles agradeço do fundo coração por tudo.

Por último, um agradecimento enorme à minha mulher e namorada, Ana Lopes, pela força, pela paciência, pela amizade, pelos ensinamentos e ajuda que me prestou no decorrer destes anos, mas, especialmente por ter estado sempre a meu lado em todos os momentos.

Resumo

Ao longo dos anos, o número de viaturas a circular nas estradas tem vindo a aumentar significativamente, principalmente nos grandes centros urbanos, onde se concentra um elevado número de pessoas, empresas e atividades. Esse aumento anormal conduziu a inúmeras e nefastas consequências, como altos níveis de congestionamento de tráfego, emissões excessivas de CO_2, aumento do tempo despendido em deslocações e diminuição da qualidade de vida das pessoas.

Devido à saturação e complexidade das redes de transporte, torna-se essencial encontrar várias soluções que permitam compreender o tráfego rodoviário e prever eventuais ocorrências que possam surgir com o objetivo de minimizar os impactos negativos provocados pelo mesmo.

O objetivo desta dissertação passa por desenvolver uma metodologia de processamento de dados capaz de caracterizar e prever o fluxo rodoviário numa determinada autoestrada, através da aplicação de técnicas de *machine learning*. Para esse efeito, são analisados e usados dados históricos, recolhidos por pórticos eletrónicos instalados ao longo da autoestrada A25, em Portugal.

Esta dissertação é desenvolvida no âmbito do projeto *OPTIMUM - Research and Innovation Action* - financiado no âmbito do Horizonte 2020 - Programa - Quadro Comunitário de Investigação e Inovação da União Europeia cujo objetivo principal é explorar soluções inovadoras para colmatar o congestionamento das redes de transportes, designadamente a transferência de fluxos para vias menos congestionadas e reduzir assim os problemas de mobilidade que os cidadãos enfrentam no dia a dia.

Palavras-chave: Sistemas Inteligentes de Transporte (SIT), Modelos de Previsão de Tráfego, Prospeção de Dados, Análise de Séries Temporais, Técnicas de *Machine Learning*

Abstract

Over the years, the number of vehicles on the road has been increasing, especially in large urban centers, where a large number of people, companies and activities are concentrated. This abnormal increase has led to numerous and disastrous consequences, such as high levels of traffic congestion, excessive CO_2 emissions, increased travel time and reduced quality of life.

Due to the saturation and complexity of transport networks, it has become essential to solutions to understand road traffic on the main access roads to cities and to anticipate possible occurrences that may arise in order to minimize the negative impacts caused by it.

The objective of this dissertation is to develop a methodology of data processing able to characterize and predict the road flow in a particular highway, through the application of machine learning techniques. For this purpose, historical data collected and analyzed by electronic tolls installed along the A25 motorway in Portugal are analyzed and used.

This dissertation is developed under the OPTIMUM project - Research and Innovation Action - funded under Horizon 2020 - Community Framework Program for Research and Innovation of the European Union whose main objective is to explore innovative solutions to solve the congestion of transport networks, in particular the transfer of flows to less congested roads and thus reduce the mobility problems that citizens face on a day to day basis.

Keywords: Intelligent Transport Systems (SIT), Traffic Prediction Models, Data Mining, Time Series Analysis, Machine Learning Techniques

Índice

Lista de Figuras

Lista de Tabelas

Siglas

AD	Árvores de Decisão.
ATIS	Advanced Traveller Information System.
ATMS	Advanced Traffic Management System.
CRISP-DM	Cross Industry Standard Process for Data Mining.
DM	Data Mining.
GRIS	Group for Research in Interoperability of Systems.
IP	Infraestruturas de Portugal.
KDD	Knowledge Discovery from Databases.
MAE	Mean Absolute Error.
MAPE	Mean Absolute Percentage Error.
ML	Machine Learning.
RB	Redes Bayesianas.
RF	Random Forest.
RNAs	Redes Neuronais Artificiais.
SIT	Sistemas Inteligentes de Transporte.
TIC	Tecnologias de Informação e Comunicação.
UE	União Europeia.

Capítulo 1

Introdução

Este capítulo tem como intuito abordar o contexto e a motivação para o desenvolvimento da presente dissertação. São identificados os objetivos e contribuições que se pretendem atingir e é apresentada a estrutura geral deste documento.

1.1 Contextualização

As cidades da União Europeia (UE) são hoje palco de elevada concentração de pessoas, empresas e atividades. Estas são o lar de mais de 70% da população da UE e geram mais de 80% do Produto Interno Bruto (PIB) europeu [1]. Apesar das cidades serem símbolo de riqueza, inovação e tecnologia, é nelas que se concentra uma enorme quantidade de problemas, como o desemprego, a poluição e a criminalidade [2].

Um dos problemas atuais que existe nas cidade prende-se com as condições de tráfego rodoviário. Atualmente, as condições de tráfego rodoviário são um dos problemas de mobilidade urbana que afeta diariamente a qualidade de vida dos cidadãos europeus [1]. Ao longo dos anos, o número de viaturas a circular nas estradas tem vindo a aumentar significativamente conduzindo a inúmeras e nefastas consequências:

- Aumento do congestionamento do tráfego rodoviário;
- Aumento do número de acidentes;
- Aumento do tempo despendido em deslocações;
- Aumento dos níveis de poluição;
- Diminuição da qualidade de vida das pessoas;
- Entre outras [1].

Segundo a Comissão Europeia[1], o custo total provocado pelo congestionamento de tráfego rodoviário é de 80 mil milhões de euros anuais e cerca de 23% das emissões de CO_2 libertadas para a atmosfera são provenientes dos transportes nas zonas urbanas [1].

Apesar do aumento substancial do número de viaturas a circular nas estradas, existem outros fatores que podem influenciar as condições de tráfego rodoviário, nomeadamente as condições meteorológicas, ocorrência de eventos especiais (como por exemplo, concertos ou jogos de futebol) ou a própria topologia da rede.

Tendo em conta a crise económica que assombra o mundo contemporâneo, a opção de investir em novas infraestruturas tornou-se pouco viável [1]. Daí que, nos dias de hoje, seja imperativo encontrar meios alternativos, financeiramente atraentes, que possibilitem solucionar os vários problemas da sociedade atual, nomeadamente, o problema da mobilidade urbana.

É neste contexto que se inserem os Sistemas Inteligentes de Transporte (SIT), um conjunto de mecanismos de auxílio à mobilidade urbana, que procuram modernizar o sector dos transportes recorrendo ao uso das Tecnologias de Informação e Comunicação (TIC) [3]. Estes sistemas visam proporcionar serviços inovadores relacionados com os diferentes modos de transporte, a fim de tornar o uso das redes de transporte mais eficiente, limpo, seguro e inteligente [3].

Alguns exemplos desses sistemas são: os painéis de mensagem variável, sistemas de partilha de carros, semáforos inteligentes, veículos autónomos, entre muitos outros. A Figura 1.1 ilustra alguns exemplos desses sistemas atualmente implementados.

Figura 1.1: Exemplos de Sistemas Inteligentes de Transporte. Fonte:[4]

[1] Órgão executivo da União Europeia

1.2 Motivação e Objetivo

As condições de tráfego rodoviário representam atualmente um dos problemas de mobilidade urbana. Nesse sentido, é importante encontrar uma solução que permita compreender o tráfego rodoviário nas principais vias de acesso às cidades e prever eventuais ocorrências que possam surgir, com o objetivo de minimizar os impactos negativos provocados pelo mesmo.

A integração de modelos de previsão a curto prazo das condições de tráfego aliado aos Sistemas Inteligentes de Transporte surge então como uma mais valia para o sector da mobilidade e transportes, uma vez que permite a criação de Sistemas Avançados de Gestão de Tráfego (Advanced Traffic Management System (ATMS)) e de Sistemas Avançados de Informações ao Viajante (Advanced Traveller Information System (ATIS)) altamente eficientes e em tempo-real [5].

Ao serem fornecidas atempadamente informações sobre o estado rede de transportes aos usuários do sistema de transporte, estas permitem reduzir a incerteza no processo de tomada de decisão e assim antecipar possíveis problemas em vez de lidar com os problemas após os mesmos já terem ocorrido. Os ATMS e ATIS são atualmente considerados de extrema importância no contexto das infraestruturas rodoviárias porque permitem:

- aos gestores de infraestruturas, uma gestão mais ativa e planeada das redes de transportes.

- aos condutores, rotas mais otimizadas a nível de tempos de viagem, bem como a escolha de percursos alternativos;

Com o desenvolvimento desta dissertação pretende-se através da análise de dados, desenvolver uma metodologia de processamento de dados capaz de caracterizar e prever o fluxo rodoviário numa determinada autoestrada de formar a ser integrado num dos sistemas acima mencionados. Para esse efeito, são aplicadas técnicas de Machine Learning e usados dados históricos recolhidos por pórticos eletrónicos instalados ao longo de uma autoestrada.

1.3 Descrição do Problema

Para a elaboração da presente dissertação, em parceria com o projeto *OPTIMUM*[2] - *Research and Innovation Action* - financiado no âmbito do Horizonte 2020 - Programa - Quadro Comunitário de Investigação e Inovação da União Europeia, foi disponibilizado pela Infraestruturas de Portugal (IP)[3] um conjunto de dados provenientes de trinta e dois pórticos eletrónicos instalados ao longo da autoestradas A25, em Portugal.

[2] http://www.optimumproject.eu/
[3] http://www.infraestruturasdeportugal.pt/

A escolha desta autoestrada, que liga Aveiro a Espanha, reside no facto de ser uma das principais autoestrada de Portugal por onde passam diariamente milhares de veículos ligeiros e pesados com destino ao estrangeiro, sendo a principal porta rodoviário de entrada e saída para o resto da Europa.

Os dados fornecidos são referentes a Janeiro de 2012 a Dezembro de 2016 e contêm informações sobre o fluxo rodoviário nas diferentes vias de acesso à autoestrada A25. Através destes pretende-se analisar e extrair diferentes perfis de utilização da autoestrada A25 em diferentes intervalos de tempo (perfis diários, perfis semanais, etc.) a fim de identificar quais os troços da autoestrada que exibem perfis de utilização semelhante durante os dias úteis e durante os fins de semana.

Após executada essa análise de dados, pretende-se desenvolver modelos preditivos que estimem o fluxo rodoviário, para cada troço da autoestrada A25, com base nos dados históricos fornecidos. Os modelos desenvolvidos deverão ser capazes de prever o fluxo rodoviário em tempo real e com base no estado atual da rede.

Na elaboração de trabalhos desta natureza existem três questões principais a ter em conta: **Qualidade dos Dados**, **Eventos Esperados e Inesperados** e **Padrões Sazonais**.

- **Qualidade dos Dados:** os dados são recolhidos em tempo real, em intervalos de tempo de 5 minutos, através de pórticos eletrónicos posicionados ao longo da autoestrada A25. Os pórticos são sistemas formados por componentes eletrónicos que se encontram em campo aberto, sujeitos a condições atmosféricas adversas, falta de manutenção, entre outros fatores, como tal são propícios a ocorrência de falhas inesperadas. Como os dados são coletados em tempo real e em intervalos de tempo muito curtos, esses valores atípicos ou incomuns irão ser introduzidos no conjunto de dados, representativos do estado do tráfego rodoviário na área em que se encontra o pórtico eletrónico. Posto isto, deve-se detetar essas falhas presentes no conjunto de dados e converter em valores que façam sentido para o problema ou mesmo excluí-los e tratá-los como indisponíveis.

- **Eventos Esperados e Inesperados:** os dados fornecidos contêm informações sobre o fluxo rodoviário nos vários troços da autoestrada A25, ao longo do tempo em que foram recolhidos. Contudo, esses dados não chegam para prever o estado do tráfego com grande rigor, uma vez que eventos esperados e inesperados podem ocorrer tornando as previsões pouco precisas. Como exemplo, poderemos considerar a ocorrência de chuva ou a ocorrência de um evento (ex. jogo de futebol) como uma causa do aumento do número de viaturas nas redes rodoviárias. Desta forma, eventos esperados são mais fáceis de abordar usando diferentes fontes de dados, como dados da meteorologia, dados de eventos culturais, etc.

- **Padrões Sazonais:** um dos pontos importantes a ter em conta na elaboração dos diferentes perfis de utilização da autoestrada A25 são os padrões de sazonalidade. As diferentes estradas podem apresentar comportamentos distintos nas diferentes estações do ano. Como exemplo, poderemos considerar que autoestradas têm mais afluência em épocas de Verão, visto que as pessoas se encontram em período de férias, ao invés das estradas urbanas que apresentam mais afluência no período de inverno quando estas se encontram em período laboral. Para estas situações serão analisados e fornecidos contextos de modo a explicar esses mesmos comportamentos.

1.4 Abordagem Proposta

Conforme foi descrito na Secção 1.3, este trabalho possui dois objetivos concretos de descoberta de conhecimento em base de dados:

- **1º objetivo:** Identificar vários tipos de perfis de utilização da rede viária, tendo em conta tratar-se de um dia de semana ou fim-de-semana.
- **2º objetivo:** Implementação de vários modelos de previsão de fluxo rodoviário, tendo por base dados capturados através de pórticos eletrónicos;

De forma a garantir uma abordagem estruturada na análise de dados, esta dissertação seguirá a metodologia Cross Industry Standard Process for Data Mining (CRISP-DM), amplamente utilizada por especialistas na área de Data Mining (DM), como orientação em projetos de descoberta de conhecimento em base de dados [6].

Esta metodologia descreve as etapas típicas na elaboração de um projeto de DM, bem como as tarefas envolvidas em cada etapa do projeto. As suas principais vantagens são o facto de poder ser aplicada a qualquer tipo de negócio, como análise de dados comerciais ou financeiros, de ser independente do tipo de ferramenta a utilizar e de se assemelhar com os modelos de processo de Knowledge Discovery from Databases (KDD) [7].

A Figura 1.2 apresenta todas as etapas desta metodologia e suas dependências.

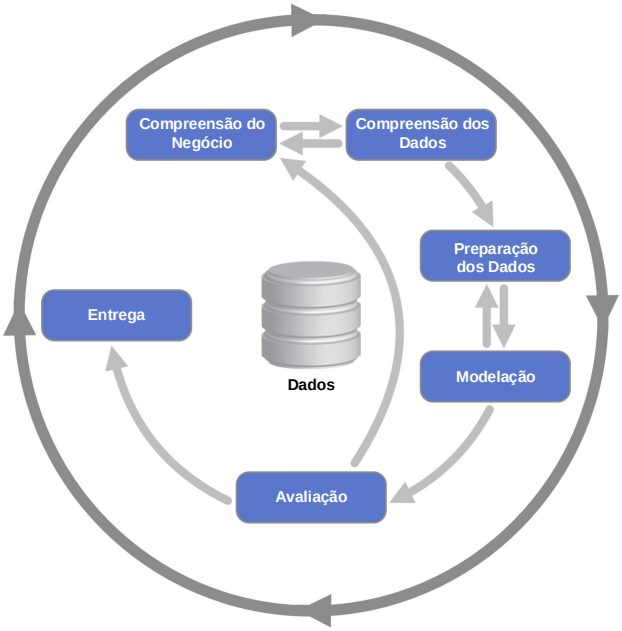


Figura 1.2: Metodologia CRISP-DM, adaptado a partir de [7]

Conforme é possível observar na Figura 1.2, o processo assemelha-se a um ciclo de vida iterativo assente em seis fases com setas que indicam as dependências mais relevantes e frequentes entre as fases. É importante perceber que a sequência entre as fases não é rígida e que a maioria dos projetos se movimenta entre as fases sempre que necessário.

Seguidamente são explicadas as várias fases da metodologia CRISP-DM.

- **Compreensão do Negócio**: Nesta fase inicial do processo é feita uma análise do problema do ponto de vista de negócio ou funcional e são definidos os objetivos que se pretende atingir com a descoberta de conhecimento em base de dados. Depois de estabelecidos os objetivos é elaborado um plano claro e objetivo das ações a serem tomadas que valide a satisfação dos objetivos propostos pelo projeto. O objetivo geral desta etapa é averiguar fatores importantes que podem ter impacto no resultado final.

- **Compreensão dos Dados**: Esta etapa consiste na recolha e subsequentemente na exploração dos dados com vista à sua compreensão e análise. É de extrema importância compreender as diferentes fontes de dados disponíveis e quais os tipos de dados que serão encontrados nessas fontes de dados, como sua localização, estrutura, formato e tipos de valores que podem assumir, de modo a averiguar quais os dados relevantes para o problema. Esta fase do processo pode ser divida em cinco partes:

  - Recolha de dados iniciais;
  - Compreensão dos dados;

  - Descrição dos dados;
  - Exploração e verificação da qualidade e disponibilidade dos dados;
  - Análise de subconjuntos interesse de dados;

  Posto isto, nesta fase do processo é realizada uma recolha dos dados relativos ao comportamento do tráfego rodoviário na autoestrada A25 e efetuada uma descrição, exploração e verificação da qualidade dos mesmos. Posteriormente, é também executada uma análise exploratória de perfis de utilização da autoestrada A25.

- **Preparação dos Dados**: Nesta etapa do processo são realizadas diversas tarefas a fim de construir um conjunto final de dados para que estes possam ser objeto de estudo pelas ferramentas de modelação de DM. A preparação dos dados é tipicamente dividida em cinco partes:

  - Seleção de dados;
  - Limpeza dos dados;
  - Transformação dos dados;
  - Integração dos dados;
  - Formatação dos dados;

  Visto que este trabalho apresenta dois objetivos diferentes de descoberta de conhecimento em base de dados, sempre que necessário deverá retroceder-se a fase de preparação dos dados.

- **Modelação**: Nesta etapa do processo são selecionadas e aplicadas várias técnicas de modelação de dados que permitem resolver os problemas identificados na primeira fase do processo. Nesta dissertação são aplicadas técnicas de agrupamento a fim de identificar troços da autoestrada A25 com perfis de utilização semelhantes para os dias úteis e para os fins de semana. Para o desenvolvimento de modelos preditivos que estimem o fluxo rodoviário em cada troço da autoestrada A25, são empregues modelos *naive* e modelos não-paramétricos.

- **Avaliação**: Nesta fase do processo é feita uma avaliação e revisão das atividades realizadas na construção do(s) modelo(s) através de métricas de desempenho. O principal objetivo é verificar se as metas pretendidas na primeira fase do processo são alcançados com alta qualidade para a perspetiva de análise de dados.

- **Entrega** (*Deployment*): Nesta última etapa do processo é produzido um relatório final contendo o resumo de todo o processo realizado, de maneira a tornar o conhecimento obtido sobre os dados em informação útil. É também elaborado um plano de monitorização e manutenção do projeto.

1.5 Contribuições

Esta dissertação é desenvolvida no âmbito do projeto *OPTIMUM - Research and Innovation Action* - financiado no âmbito do Horizonte 2020 - Programa - Quadro Comunitário de Investigação e Inovação da União Europeia cujo objetivo principal é explorar soluções inovadoras para colmatar o congestionamento das redes de transportes, designadamente a transferência de fluxos para vias menos congestionadas e reduzir assim os problemas de mobilidade que os cidadãos enfrentam no dia a dia.

As principais contribuições da presente dissertação passam por:

- Analisar e caracterizar diferentes perfis de utilização da autoestrada A25;
- Desenvolver modelos descritivos capazes de identificar troços de estrada com perfis de utilização semelhante para os dias úteis e para os fins de semana;
- Desenvolver modelos preditivos capazes de estimar o fluxo rodoviário em cada troço da autoestrada A25;

1.6 Estrutura do documento

Esta dissertação encontra-se organizada em seis capítulos distintos.

No presente capítulo é feito uma contextualização do tema em estudo e descrito a motivação e os objetivos que levaram à realização desta dissertação.

No segundo capítulo são abordados vários conceitos fundamentais que auxiliam à realização da presente dissertação. É feita uma revisão bibliográfica relativa à previsão de tráfego rodoviário e às abordagens utilizadas. De seguida, é abordado o conceito de Data Mining e sua relação em processos de descoberta de conhecimento em bases de dados, bem como algumas técnicas de DM relevantes para elaboração de trabalhos desta natureza.

No terceiro capítulo é efetuada uma descrição detalhada sobre as fontes de dados disponíveis para este trabalho e excetuada uma análise da disponibilidade e qualidade dos mesmos. Neste mesmo capítulo é também realizado um pré-processamento inicial dos dados a fim de obter um conjunto de dados uniformizados e preparados para o processamento de técnicas de DM, bem como uma análise exploratória de perfis de utilização da autoestrada A25 presente no conjunto de dados.

No quarto capítulo é descrito o processo adotado nesta dissertação relativamente à identificação de troços da Autoestrada A25 com perfis de utilização semelhante para os dias úteis e para os fins de semana.

No quinto capítulo são apresentados os métodos utilizados na criação de modelos de previsão do fluxo rodoviário para cada troço da autoestrada A25 e apresentados os resultados óbitos através de métricas de desempenho.

No sexto capítulo é elaborado um resumo de todo o trabalho desenvolvido na presente dissertação e retiradas as principais conclusões obtidas. Além disso é também apresentada uma reflexão do que poderia ser melhorado no futuro.

No final apresentam-se as referências bibliográficas utilizadas para o desenvolvimento desta dissertação.

Capítulo 2

Estado de Arte

Neste capítulo são apresentados vários conceitos fundamentais que auxiliam à realização da presente dissertação. Numa fase inicial, é efetuada uma revisão bibliográfica relativa à previsão de tráfego rodoviário e às abordagens utilizadas. De seguida, é abordado o conceito de Data Mining e a sua relação em processos de descoberta de conhecimento em bases de dados, bem como algumas técnicas de Machine Learning utilizadas na área.

2.1 Previsão de Tráfego

A previsão de tráfego é hoje parte integrante da maioria dos Sistemas Inteligentes de Transporte (SIT), nomeadamente em Sistemas Avançados de Gestão de Tráfego (ATMS) e em Sistemas Avançados de Informação ao Viajante (ATIS), porque os mesmos carecem de informações em tempo real e precisas sobre como as condições de tráfego irão evoluir ao longo do tempo [8].

Os ATMS são essencialmente utilizados para reduzir o congestionamento do tráfego, aumentar a segurança nas estradas e a melhorar o fluxo de tráfego de veículos, principalmente em centros urbanos, através de softwares que monitorizam as condições do tráfego. Os ATIS destinam-se a fornecer informações aos usuários do sistema de transporte como, por exemplo, opções de rota ou tempos de viagem estimados [5].

Na literatura científica, a fonte de dados mais utilizada na área de previsão de tráfego rodoviário são os sensores de *inductive-loop*, também designados por detetores de veículos (VD) [9]. Estes sensores de baixo custo, são constituídos por bobines instaladas no pavimento de estadas. A sua função é capturar dados, tais como o número de veículos que passam sobre as espiras, durante uma unidade de tempo [10].

A cada um destes sensores é associado uma localização e uma data e hora da captura efetuada. A Figura 2.1 representa o funcionamento desse sistema.

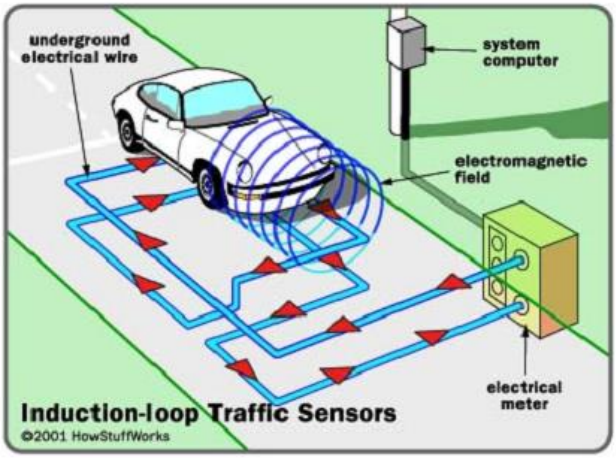


Figura 2.1: Esquema representativo de um sistema de *inductive-loop*. Fonte:[11]

O constante desenvolvimento na área tecnológica, aliada a uma era cada vez mais digital, permitiu a utilização de novas fontes de dados, tais como dados obtidos a partir de câmaras de vídeo-vigilância, veículos com GPS (*Global Positioning System*), dispositivos móveis, informações de eventos, possibilitando aos investigadores desenvolver novas metodologias de previsão usando diferentes parâmetros de entrada e saída, diferentes fontes de dados e diferentes horizontes temporais.

Apesar de atualmente existir uma enorme variedade de modelos e técnicas aplicadas à previsão de tráfego, o objetivo geral de todos os modelos é analisar dados estatísticos e dinâmicos de forma a gerar informação útil que possa ser utilizada no planeamento de gestão de tráfego e na tomada de decisão para viajantes.

Na literatura, a previsão de tráfego rodoviário pode ser classificada em dois horizontes temporais: previsões a curto prazo e previsões a longo prazo.

- **Previsões a curto prazo**
  São previsões feitas para horizontes temporais até duas horas [12]. No entanto, a sua definição exata pode variar consoante os métodos aplicados. Este tipo de previsão é essencialmente utilizado para estimar as condições de tráfego rodoviário em tempo real e com base em dados recolhidos por sensores de forma continua.

- **Previsões a longo prazo**
  São previsões realizadas em intervalos de tempo de meses a alguns anos [12]. Este tipo de previsões são principalmente utilizados no planeamento de sistemas de transportes e não tanto para operações de gestão de trânsito.

2.2 Modelos de Previsão

Os primeiros modelos de previsão de tráfego surgiram na década de setenta e tinham como principal foco a previsão do estado das redes de transportes em horizontes temporais a curto prazo (Ahmed & Cook, 1979), longo prazo (Lingras & Adamo, 1996) e no cálculo do tempo de viagem (Hiesz, 1974) [13]. Após os primeiros avanços na área da previsão de tráfego, diversas pesquisas foram realizadas resultando na formação de um vasto número de metodologias, usando uma extensa variedade de especificações matemáticas.

Atualmente essas abordagens vão desde ao uso de métodos estatísticos "clássicos", ao uso de métodos de inteligência computacional e devem-se em grande parte ao desenvolvimento de computadores mais rápidos e de métodos matemáticos mais flexíveis desenvolvidos nas últimas décadas [8].

De forma a organizar essa extensa literatura, Van Hinsbergen, Van Lint e Sanders propuseram uma taxonomia onde agrupam os diferentes modelos de previsão de tráfego em três grupos principais: modelos *naive*, modelos paramétricos e modelos não-paramétricos [14]. A figura 2.2 resume os principais modelos de previsão de tráfego referido pelos autores.

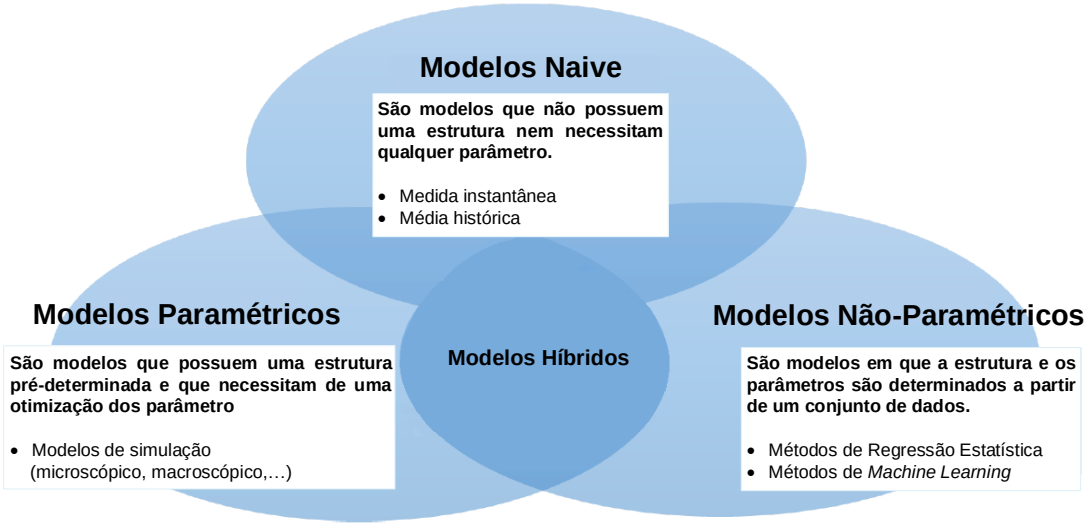


Figura 2.2: Taxonomia de modelos de previsão de tráfego, adaptado a partir de [14]

2.2.1 Modelos Naive

Os modelos *naive* são abordagens de previsão pouco sofisticados que não possuem uma estrutura modelo nem requerem qualquer parâmetro [15]. Estes modelos utilizam os dados fornecidos e servem-se de relações físicas exatas para estimar características do tráfego (por exemplo, distância = velocidade x tempo) [14].

No campo da previsão de tráfego, estes modelos são aplicados na prática devido à sua fácil implementação e ao seu baixo esforço computacional [15]. Estes modelos são também utilizados como linha de base para comparações com outros modelos mais sofisticados. Exemplos de modelos *naive* são a medida instantânea e as médias históricas.

É de destacar ainda que existem abordagens de previsão desenvolvidas por autores, que combinam modelos *naive* com modelos não-paramétricas.

2.2.1.1 Medida Instantânea

Neste tipo de abordagem pressupõe-se que o estado atual do tráfego (fluxo, densidade, velocidade, etc.) permanecerá constante indefinidamente [15]. Este tipo de abordagem não requerer um poder computacional de elevada complexidade dado que nenhum cálculo é efetuado. Ao invés, o valor atual do tráfego é lido e estimado com base no mesmo o valor.

Na literatura científica, a aplicação desta técnica tem demonstrado resultados bastantes razoáveis para previsões a curto-prazo, nomeadamente em zonas de autoestrada e vias equiparadas devido ao facto de o comportamento do tráfego se apresentar de forma muito homogénea durante um longo período de tempo [12].

No entanto, a sua aplicação em zonas urbanas tem exibido resultados pouco consistentes devido à dinâmica inerente nesses locais (semáforos, rotundas, filas, etc.).

2.2.1.2 Médias Históricas

Esta técnica, também de baixo esforço computacional e de fácil implementação, consiste no uso de médias históricas a partir dos dados para uma determinada hora do dia, dias da semana ou até mesmo períodos de sazonalidade, com o propósito de estimar o comportamento do tráfego.

Este tipo de abordagem é indicado para previsões a longo prazo, onde o valor atual do tráfego não é tão importante, uma vez que não tem muita influência a longo prazo [12]. Porém, para previsões a curto prazo, este método não é muito adequado devido à sua incapacidade para lidar com eventos e incidentes inesperados.

2.2.1.3 Combinação de Medidas Instantâneas e Médias Históricas

São várias as abordagens que combinam estes dois métodos, nomeadamente uma bastante usual que deriva da suposição de que a relação entre os dados atuais e as médias históricas é bom indicador de como as condições de trânsito no próximo intervalo de tempo se vão distanciar das condições históricas [12].

São vários os estudos realizados por diversos autores que confirmaram que a aplicação deste tipo de abordagem permitiu aumentar o desempenho do modelo de previsão [14].

2.2.2 Modelos Paramétricos

Os modelos paramétricos, também designados por *model-driven*, são caracterizados por uma estrutura pré-determinada derivada do conhecimento do processo de tráfego [15]. Estes modelos apenas necessitam de uma otimização dos parâmetros com recurso a dados reais. Estes modelos aplicam funções analíticas como por exemplo, funções do tempo de viagem ou funções de filas, e modelos de simulação do tráfego (macroscópicos ou microscópicos) com o intuito de estimar determinadas características do tráfego [14].

Uma das fortes vantagens do uso deste tipo de modelos é que estes necessitam de menos dados quando comparados com modelos não-paramétricos [12]. Por sua vez, estes modelos são difíceis de construir devido à natureza do tráfego apresentar um comportamento com picos extremos e flutuações rápidas o que implica um longo e exaustivo processo de calibração dos parâmetros.

2.2.3 Modelos Não-Paramétricos

Os modelos não-paramétricos, também designados por *data-driven*, são vistos como abordagens de previsão, onde a estrutura do modelo e os parâmetros do mesmo são determinados a partir de um conjunto de dados [15].

Uma das fortes vantagens do uso deste tipo de abordagem é que esta requer menor domínio de conhecimento do processo do tráfego em comparação com os modelos paramétricos[12]. Contudo, para que a sua implementação seja bem-sucedida é necessário uma enorme quantidade dados.

No geral os modelos não-paramétricos podem ser divididos em dois tipos de abordagens:

- Métodos de Regressão Estatística Clássica
- Métodos de Machine Learning

2.2.3.1 Métodos de Regressão Estatística Clássica

Os métodos de regressão estatística clássica foram uma das primeiras abordagens aplicadas pela maioria – se não toda – a comunidade científica no desenvolvimento de modelos de previsão de tráfego assente em dados [8].

No campo da previsão de tráfego, estes métodos são amplamente utilizados para prever características do tráfego em um único ponto devido ao facto de retornarem resultados bastante satisfatórios [12]. O motivo da sua aplicação deve-se ao facto de a previsão de tráfego em único ponto ser vista como um problema de modelação de séries temporais que se insere no domínio dos diversos métodos estatísticos [12].

Exemplos de métodos de regressão estática são os modelos da família ARIMA (Auto-Regressivo Integrado de Médias Móveis), que se podem estender a diversas variâncias como ARMA, SARIMA, ARIMAX, etc. e métodos de Regressão Linear.

2.2.3.2 Métodos de Machine Learning

Os modelos de previsão de tráfego usando técnicas de ML tem vindo sistematicamente crescendo, devido em grande parte aos recentes avanços na área de programação orientada a objetos e nas suas aplicações em tempo real de recolha, armazenamento e gestão de grandes bases de dados de vários pontos de uma rede de transportes [8].

Na literatura científica, este tipo de abordagem tem comprovado resultados bastante bons no campo da previsão de tráfego a curto prazo devido ao facto de o comportamento dinâmico e não-linearidade apresentado pelo tráfego rodoviário poder ser modelado apenas recorrendo a dados [12, 14]. No entanto, uma das desvantagens do uso deste tipo de abordagem é que situações de tráfego inesperadas, como congestionamentos incomuns ou acidentes não são tidos em conta no modelo, o que pode representar um problema quando se quer prever características do tráfego rodoviário.

Exemplos de modelos de previsão de tráfego usando técnicas de ML são as Redes Neuronais Artificiais (RNAs) [16], Árvores de Decisão (AD) [17], Redes Bayesianas (RB) [18] e métodos compostos, como por exemplo, Random Forest (RF) [19].

2.2.4 Discussão

A partir da literatura, verifica-se a existência de um grande número de abordagens usando métodos de regressão estatística "clássica". Dentro dos métodos de regressão estatística "clássica", o algoritmo de regressão linear foi um dos métodos que apresentou resultados bastante satisfatórios tanto a nível de rapidez como a nível de precisão, especialmente em dados referentes a autoestradas [20]. Estes resultados devem-se ao facto do tráfego nas autoestradas ser menos dinâmico do que nas cidades.

Face à disponibilidade de dados sem precedentes, ao desenvolvimento de computadores mais rápidos e a capacidade de processar mais rapidamente estes dados, a maioria dos investigadores nos últimos anos tem adotado técnicas de ML [8].

Uma das abordagens mais utilizado no campo de previsão de tráfego são as Redes Neuronais Artificiais, devido à sua capacidade de trabalhar com enorme quantidade de dados multi-dimensionais e devido à sua boa capacidade preditiva [12]. No entanto, um dos grandes problemas das RNAs é que esta necessitam de um longo período de treino de forma a ajustar os pesos da rede.

Nesta dissertação o principal foco serão os modelos não-paramétricos, mais concretamente os modelos usando técnicas de Machine Learning, devido à enorme quantidade de abordagens implementadas nos últimos anos e também devido ao facto de retornarem bons resultados no campo da previsão de tráfego rodoviário a curto prazo.

2.3 Descoberta de Conhecimento em Base de Dados e *Data Mining*

O processo de Descoberta de Conhecimento em Base de Dados, ou Knowledge Discovery from Databases, tem visto o seu nível de importância aumentar nos últimos anos, devido em grande parte ao aumento exponencial do volume de bases de dados mas também porque existe o princípio de que essas bases de dados muito grandes podem ser fonte de conhecimento útil e de grande importância e aplicação nas mais diversas áreas.

Este processo tem como finalidade gerar conhecimento novo e potencialmente útil, através da procura e da identificação de padrões através da análise de dados, armazenados em bases de dados muitas vezes dispersos e inexplorados [21]. Essa análise é realizada através de inúmeras técnicas que intercetam disciplinas como estatística, sistemas de base de dados, ML, inteligência artificial, entre outras.

Existem autores que definem o termo Data Mining, ou prospeção de dados, como o processo de extração não trivial de informações implícitas, anteriormente desconhecidas e potencialmente úteis a partir de dados [22]. No entanto, existem outros autores que referem DM como uma etapa específica no processo de KDD, através do qual são aplicados algoritmos específicos para extrair padrões de dados, como se ilustra na figura 2.3.

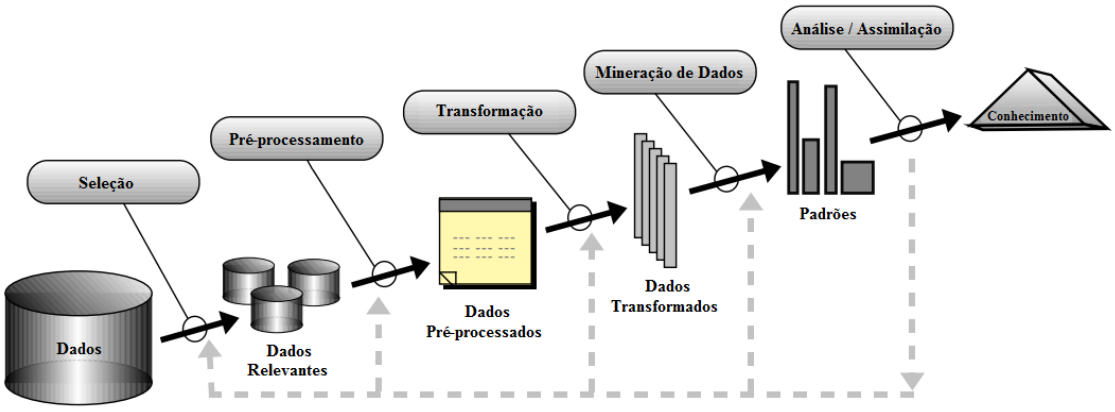


Figura 2.3: Figura representando o processo KDD [23]

De acordo com Fayyad, Shapiro e Smyth, este processo é altamente dinâmico, iterativo e interativo que envolve vários passos, com várias decisões a serem tomadas pelo utilizador e enunciam alguns passos básicos no processo KDD [23].

O primeiro passo passa por compreender, identificar e selecionar os dados relevantes no domínio da aplicação e dos objetivos que se pretendem atingir no projeto. Após feita uma colheita dos dados considerados importantes, estes devem ser armazenados em uma base de dados de forma a dar início ao processo de descoberta de conhecimento.

O passo seguinte consiste em realizar um pré-processamento dos dados. Os dados do mundo real tendem a ser incompletos, com ruído e inconsistentes. Nesse sentido, é necessário efetuar várias tarefas de modo a assegurar a qualidade dos dados envolvidos no processo de KDD.

Após realizado um pré-processamento dos dados, estes devem ser transformados para que possam ser objeto de estudo por técnicas de Data Mining. Deste modo, nesta fase do processo são selecionados os atributos dos dados que realmente têm importância e é feita uma transformação dos mesmos.

Depois de efetuados todos os passos anteriormente mencionados, o passo seguinte passa pela aplicação de técnicas de Data Mining. As técnicas de DM são de extrema importância no processo de descoberta de conhecimento em base de dados porque permitem extrair conhecimento útil, tais como anomalias, padrões e correlações em grandes conjuntos de dados e de representar esse conhecimento descoberto de forma percetível para o utilizador final, ou seja, o analista.

O último passo consiste em realizar uma interpretação e avaliação dos resultados alcançados na fase anterior. No final é documentado e reportado o conhecimento descoberto às partes interessadas ou incorporado esse conhecimento em sistemas tecnológicos.

2.4 Técnicas de Data Mining

O processo de DM engloba diversas técnicas, das mais variadas áreas, para análise de dados. Essas técnicas tem capacidade de realizar determinadas tarefas no processo de descoberta de conhecimento, sendo as tarefas mais comuns as seguintes [24]:

- **Descrição –** esta tarefa, muitas vezes utilizada em conjunto com técnicas de análise exploratória de dados, visa descrever de forma concisa os padrões e tendências revelados pelos dados.

- **Predição –** esta tarefa consiste na construção de modelos que permitam prever o valor de uma determinada variável em função de um conjunto de outras variáveis. O seu desenvolvimento pode ser dividido em dois níveis:

  - **Classificação –** esta tarefa visa classificar um item de dados em várias classes pré-definidas. Para tal, recorre-se a um conjunto de casos de treino, em que cada caso corresponde a uma determinada classe, e de seguida, com base nesses casos classificam-se novas instâncias.

  - **Regressão –** esta tarefa visa a construção de uma função que permita determinar uma relação entre variáveis dependentes e uma ou mais variáveis independentes. Este tipo de tarefa é utilizado quando os dados se encontram com características numéricas em vez de características categóricas.

- **Agrupamento –** esta tarefa visa identificar e agrupar objetos de dados com características semelhantes. Esta tarefa difere da classificação e da regressão pois não necessita que os registos sejam previamente categorizados.
- **Associação –** esta tarefa tem com objetivo encontrar um modelo que descreva as dependências significantes entre variáveis.

Associado aos sistemas computacionais aparece o conceito de Aprendizagem Automática ou Machine Learning.

ML é uma área de investigação, dentro da área da Inteligência Artificial (AI), relacionada com o desenvolvimento de algoritmos que permitem que os computadores aprendam a reconhecer padrões e a tomar decisões através de dados [25]. Estas técnicas podem ser divididas em duas categorias: **técnicas de aprendizagem supervisionada** e **técnicas de aprendizagem não-supervisionada**.

2.4.1 Técnicas de Aprendizagem Supervisionada

Na aprendizagem supervisionada os dados já se encontram rotulados. Neste tipo de aprendizagem é fornecido ao algoritmo um conjunto de dados de treino, onde os dados já se encontram pré-classificados através de um atributo desejado, e mediante a observação dos restantes atributos o algoritmo aprende a definir quais os atributos com maior relevância para a classificação [25].

Este tipo de aprendizagem é essencialmente usado em tarefas de classificação e regressão. Seguidamente, são descritos alguns métodos de aprendizagem supervisionada.

- **Árvores de Decisão** - Este método consiste na reprodução de um fluxograma em forma de árvore, onde cada nó indica um teste efetuado sobre um valor de atributo [25]. A ligação entre os nós, também designada por ramos, representam os valores possíveis do nó superior e as folhas da árvore representam a classe a que pertence o registo [25]. Após elaborada a árvore de decisão é possível classificar um novo registo, começando no nó raiz até chegar a uma folha. As AD são técnicas bastante simples e possuem grande precisão na classificação dos dados. Na Figura 2.4 pode ser visto um exemplo de árvore de decisão para a compra de um computador.

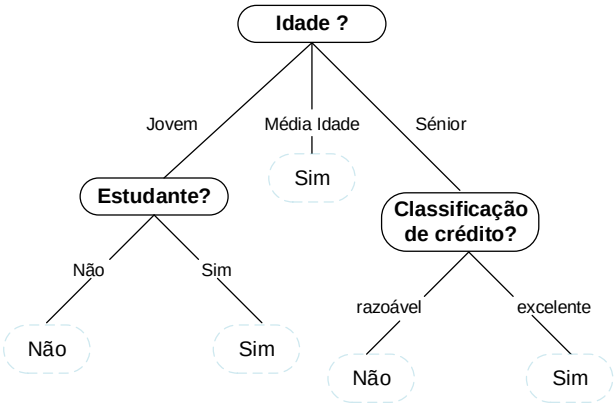


Figura 2.4: Árvore de decisão, adaptado a partir de [25]

- **Redes Bayesianas** - Este método baseia-se no teorema de Thomas Bayes de probabilidade condicional [25]. Este teorema descreve a probabilidade de um certo evento ocorrer, dada a probabilidade de um outro evento já ter ocorrido:

$$Probabilidade(A|B) = \frac{Probabilidade(B|A) * Probabilidade(A)}{Probabilidade(B)}$$

  Este método assume uma independência de classe condicional, ou seja, o valor de um determinado atributo numa classe é independente dos valores de outros atributos [25]. A sua utilização é útil quando se pretende analisar um sistema onde existem dependências estatísticas ou relações causais entre variáveis. As redes Bayesianas apresentam um nível de precisão alta e um tempo computacional baixo quando aplicado a grandes bancos de dados [25].

- **Redes Neuronais Artificiais** - As RNAs são métodos inspirados no funcionamento do cérebro biológico e nas suas conexões sinápticas, onde cada conexão tem associado um peso [25]. Durante o processo de aprendizagem, a rede ajusta estes pesos através de diversas iterações para conseguir classificar corretamente um objeto. Uma rede neuronal artificial é constituída por neurónios artificias interligados, onde cada neurónio representa uma função matemática que computa os pesos sinápticos da rede[25]. Estas encontram-se estruturadas em três tipos de camadas: camada de entrada, camada oculta e camada de saída. A camada de entrada recebe um conjunto de dados de exemplo que queremos classificar, a camada oculta executa vários cálculos internos e a camada de saída fornece um valor que representa uma classe. As principais vantagens das redes neuronais é a forte tolerância a ruído nos dados, bem como a capacidade para classificar padrões que ainda não tenham sido treinados [25]. O grande problema das redes neuronais é que estas necessitam de um longo período de treino de forma a ajustar os pesos da rede. Na Figura 2.5 pode observar-se um exemplo de uma rede neural.

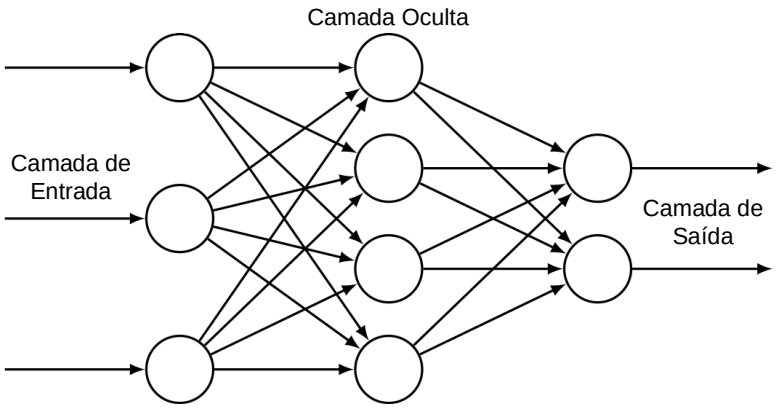


Figura 2.5: Rede Neuronal

2.4.2 Técnicas de Aprendizagem Não-Supervisionada

Na aprendizagem não-supervisionada, os dados não se encontram categorizados. Este tipo de aprendizagem consiste essencialmente em descobrir padrões ou similaridade entre os dados. Técnicas de agrupamento, de associação ou de descrição são métodos de aprendizagem não-supervisionada.

2.4.2.1 Técnicas de Agrupamento

As técnicas de agrupamento têm como finalidade a divisão de um conjunto de dados em grupos de objetos com características semelhantes [25]. O objetivo é determinar quais os objetos presentes num conjunto de dados apresentam um conjunto de propriedades semelhantes que os diferencie dos elementos de outros grupos.

O resultado final da aplicação deste tipo de técnicas é a formação de grupos com alta homogeneidade interna e alta heterogeneidades externa. De seguida são descritos vários métodos de agrupamento.

- **Métodos baseados em repartição** - Os métodos baseados em repartição consistem em organizar n objetos em k grupos, onde cada grupo contém pelo menos um objeto e onde $k <= n$ [25]. A maioria destes métodos têm como critério de separação a distância entre os objetos. Isto faz com que a separação dos objetos em grupos apresente um formato esférico. Os algoritmos mais comuns de repartição são: *k-Means* e *k-Medoids*.

- **Métodos Hierárquicos** - Neste tipo de método de agrupamento, os dados são decompostos hierarquicamente em grupos [25]. Os grupos criados por este tipo de método podem ser visualizados por meio de um dendograma, conforme se pode observar na Figura 2.6.

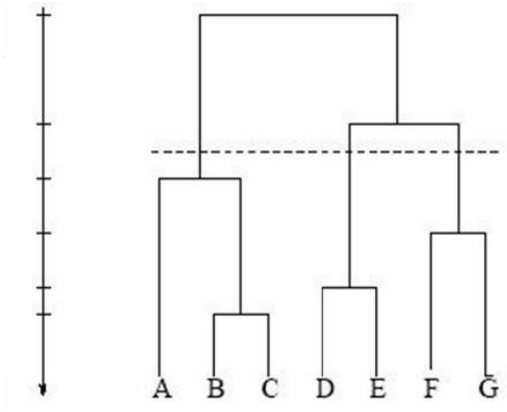


Figura 2.6: Exemplo de um dendrograma

Os métodos hierárquicos podem ser classificados em dois tipos: aglomerativos e divisivos.

- **Métodos Aglomerativos :** Nos métodos aglomerativos, cada objeto representa um grupo separado e a cada iteração que é feita, os grupos vão-se unindo em grupos maiores, até que todos os grupos formem um só grupo, ou até que atingem um determinado limiar. Os algoritmos AGNES (*AGglomerative NESting*) e CURE (*Clustering Using Representatives*) são exemplos de algoritmos aglomerativos.

- **Métodos Divisivos :** Nos métodos divisivos todos os objetos começam num só grupo e a cada iteração que é feita, os grupos vão se separando em grupos menores, até que existam grupos com apenas um objeto ou até que atinjam um determinado limiar. O algoritmo DIANA (*DIvisive ANAlysis*) é um exemplo de algoritmo divisivo.

- **Métodos Baseados em Densidade** - Dada a necessidade de encontrar aglomerados com formas arbitrarias, foram desenvolvidos métodos com base na noção de densidade de um aglomerado [25]. Exemplos de algoritmos deste método são: DBSCAN (*Density-Based Spatial Clustering of Applications with Noise*) , OPTICS (*Ordering Points to Identify the Clustering Structure*) e DENCLUE (*Density-based Clustering*).

- **Métodos Baseados em Grelha** - Nos métodos baseados em grelha, os dados são quantificados em um número finito de células que formam uma estrutura em grelha na qual todas as operações de agrupamento são executadas [25]. O algoritmo STING (*STatistical INformation Grid*) é um exemplo de algoritmo de grelha.

2.4.3 Análise de Séries Temporais

Uma série de temporal é um conjunto de observações, medidas ao longo do tempo, ordenadas segundo o instante em que ocorreram e, normalmente, registadas em períodos de tempo regulares.

Uma série temporal pode ser classificada de contínua ou discreta. Em tempo contínuo, as observações da série temporal são medidas a cada instante de tempo, enquanto que em tempo discreto as observações da série são espaçadas em intervalos de tempo, como horas, dias, semanas, meses ou anos.

A maneira tradicional de analisar uma série temporal é através da decomposição das suas características:

- **Tendência** – É um movimento que traduz a influência de fatores que fazem com que a série aumente ou diminua o seu valor com o passar do tempo. Esta componente caracteriza-se como um movimento ascendente ou descendente de longa duração. Quando a série temporal não apresenta qualquer tendência ascendente ou descendente ela é denominada de série estacionária.

- **Cíclica** - são movimentos caracterizados pelas oscilações de subida e de queda nas séries, de forma suave e repetida, ao longo da componente de tendência, que pode ou não ser periódica.

- **Sazonal** - Os movimentos sazonais corresponde às oscilações de subida e de descida que ocorrem num determinado período do ano, do mês, da semana ou do dia. A diferença entre as componentes sazonais e cíclica é que a componente sazonal possui movimentos facilmente previsíveis, ocorrendo em intervalos regulares de tempo, enquanto movimentos cíclicos tendem a ser irregulares.

Através da análise das características de uma série temporal é possível extrair componentes relevantes para elaboração de um modelo de previsão de forma a estimar determinados parâmetros da série. Na previsão de um valor futuro de uma série temporal pretende-se determinar um valor $\hat{X}_{t+m}$ a partir de um instante t para um horizonte temporal m , considerando as observações passadas.

Capítulo 3

Fontes de Dados

Neste capítulo são analisadas as fontes de dados disponíveis para realização da presente dissertação. Inicialmente é feita uma descrição detalhada da forma como os dados se encontram organizados e dos tipos de valores que podem assumir. Posteriormente, é realizada uma verificação da disponibilidade e qualidade dos dados, com o intuito de avaliar a consistência das informações facultadas (exemplo: dados duplicados, valores em falta, valores discrepantes, etc.).

Após efetuada uma compreensão dos dados, é executado um pré-processamento inicial dos mesmos de modo a remover as inconsistências detetadas anteriormente. Para além disso, é também criado um subcapítulo inteiramente dedicado à análise de perfis de utilização da autoestrada em questão, onde são destacados observações de subconjuntos relevantes.

Este capítulo obedece a duas etapas da metodologia CRISP-DM: compreensão e preparação dos dados. Toda a análise de dados elaborada neste capítulo tem como base duas ferramentas de visualização de dados designados por, Tableau Desktop[1] e Rstudio[2].

[1] https://www.tableau.com/
[2] https://www.rstudio.com/

3.1 Compreensão dos Dados

Para realização deste trabalho foram disponibilizados pela Infraestruturas de Portugal vários conjuntos de dados provenientes da computação de registos de trinta e dois pórticos eletrónicos localizados ao longo da autoestradas A25, em Portugal.

Os pórticos eletrónicos são sistemas munidos de câmaras e detetores laser transmissores que captam informações dos automobilistas nas estradas portuguesas no momento da sua passagem. Essas informações são posteriormente utilizadas para faturação do uso da autoestrada e para gestão da própria rede rodoviária. A Figura 3.1 ilustra um desses sistemas eletrónicos utilizados na recolha de informações nas estradas portuguesas.

Figura 3.1: Sistema eletrónico utilizado pela Ascendi[3]

Quanto à localização e distribuição geográfica dos 32 pórticos eletrónicos disponíveis para este trabalho, a Figura 3.2 revela o posicionamento geográfico dos mesmos por toda a extensão da autoestrada A25. É de salientar que 16 pórticos eletrónicos se encontram num determinado sentido e os restantes no sentido inverso, dai não serem visíveis 32 pórticos eletrónicos na Figura 3.2.

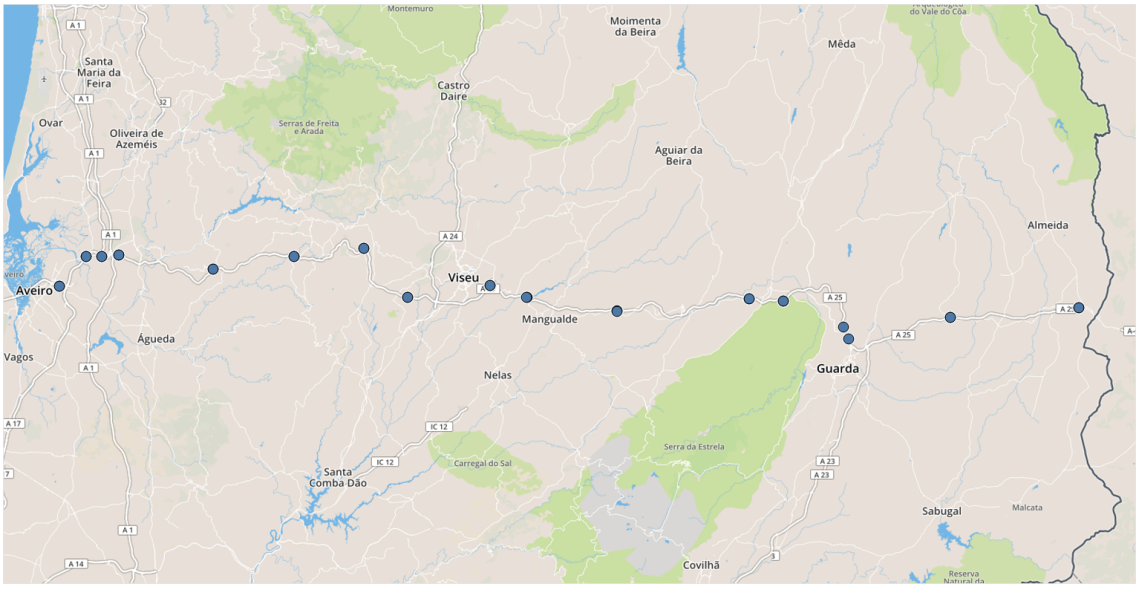


Figura 3.2: Localização geográfica dos pórticos eletrónicos na autoestrada A25

[3]Ascendi: Operadora de infraestruturas de transporte em Portugal

A representação gráfica deste mapa só foi possível devido ao facto de se encontrar presente no conjunto de dados fornecidos, informações relativas à localização geográfica dos pórticos eletrónicos (latitude e longitude).

De seguida, é feita uma descrição detalhada das características dos vários conjuntos de dados de forma ao leitor se familiarizar e ter uma melhor perceção dos mesmos.

3.1.1 Descrição dos Dados

Um dos ficheiros fornecidos, designado por *metadata.json*, contém vários registos identificados por um código único correspondente ao campo *_id*.

Cada um desses registos é constituído por um vasto conjunto de informações alusivas a cada pórtico eletrónico, nomeadamente a sua localização geográfica exata (latitude e longitude), o nome da concessão detentora, a sua direção ou o nome do troço de estrada onde se encontra. É de frisar que, para efeitos de identificação de um pórtico eletrónico existe um campo, designado por *sensor_id_holder*, que atribui um código específico para cada sensor.

Na Tabela 3.1 são apresentados todos os parâmetros presentes no ficheiro *metadata.json*, assim como a sua descrição e o tipo de medida associado.

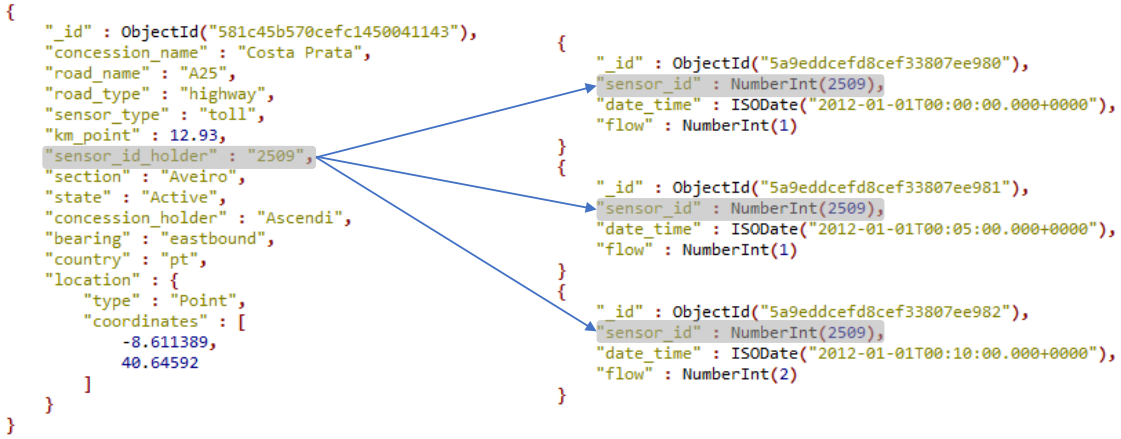
Tabela 3.1: Parâmetros do ficheiro *metadata.json*

Nome da variável		Descrição	Tipo de medida
_id		Código correspondente a cada registo	Cadeia de caracteres
sensor_id_holder		Código de identificação do pórtico	Cadeia de caracteres
concession_name		Indica o nome da Concessão detentora pelo pórtico	Cadeia de caracteres
concession_holder		Indica o nome da Concessão responsável pela manutenção do pórtico	Cadeia de caracteres
road_name		Indica o nome da estrada	Cadeia de caracteres
road_type		Indica o perfil de estrada	Cadeia de caracteres
sensor_type		Indica o tipo de sensor	Cadeia de caracteres
section		Indica o nome do troço da estrada	Cadeia de caracteres
state		Indica o estado do sensor	Cadeia de caracteres
bearing		Indica a direção do sensor	Cadeia de caracteres
country		Indica o código do país	Cadeia de caracteres
km_point		Indica o quilometro onde se encontra o pórtico	Numérico
location	type	Indica o tipo de localização	Cadeia de caracteres
	coordinates	Indica as coordenadas geográficas	Numérico

Outro dos ficheiros fornecidos, e sobre o qual assenta a maioria do trabalho realizado, designa-se por *sensor_values_A25.json*. Este ficheiro é constituído por um total de 16.578.036 registos adquiridos durante o período de Janeiro 2012 a Dezembro de 2016.

Cada registo deste conjunto de dados é caracterizado por quatro variáveis: *_id*, *sensor_id*, *date_time* e *flow*. O campo *_id*, tal como referido anteriormente, representa um código único destinado a identificar cada registo presente no conjunto de dados. O parâmetro *sensor_id* permite associar cada um desses registos ao respetivo pórtico eletrónico. Os parâmetros *flow* e *date_time* representam, respetivamente, o somatório do número de veículos que intersetou o pórtico eletrónico em intervalos de 5 minutos, numa determinada data e hora.

Na Figura 3.3 podemos observar a formatação dos diferentes registos existentes nos ficheiros *metadata.json* (a) e *sensor_values_A25.json* (b), assim como a visualização de três registos referentes ao pórtico eletrónico com o *sensor_id* e *sensor_id_holder* "2509".



(a) *Metadata.json* (b) *Sensor_values_A25.json*

Figura 3.3: Estrutura dos diferentes registos existentes nos ficheiros

3.1.2 Disponibilidade e Qualidade dos Dados

Uma das etapas fundamentais na análise de dados é a verificação da disponibilidade e qualidade dos dados, visto que estes podem comprometer seriamente as soluções propostas no final do projeto.

Tal como é descrito na Secção 3.1, os dados utilizados são referentes a trinta e dois pórticos eletrónicos instalados ao longo da autoestrada A25. Estes sistemas, formados por componentes eletrónicos, encontram-se sujeitos a condições atmosféricas adversas, falta de manutenção, atos de vandalismo, entre outros fatores, como tal são propícios a ocorrência de falhas inesperadas.

Como os dados são coletados em tempo real e em intervalos de tempo muito curtos, esses valores atípicos ou incomuns irão ser introduzidos no conjunto de dados, representativos do estado do tráfego rodoviário na área onde se encontra o pórtico eletrónico. Posto isto, esta dissertação apresenta as seguintes questões:

- **Existem dados em falta durante um intervalo de tempo aleatório?**
- **Existem erros de medição nos dados?**

No que diz respeito à questão da disponibilidade dos dados, esta pode ser compreendida como o coeficiente entre o número de dados observados e o volume de dados esperados. A seguinte equação expressa o coeficiente de disponibilidade de dados para cada um dos pórticos eletrónicos:

$$Disponibilidade = \frac{1}{N} \frac{\sum_{i}^{N} N° \; de \; Registo \; Observados_{\; i}}{N° \; de \; Registo \; Esperados} \tag{3.1}$$

onde i se refere a um dia de N ($i \in N$) e N o número total de dias do conjunto de dados, sendo que neste contexto, N é equivalente a 1827 dias. O número de registos esperados em um dia i corresponde a 288, devido ao facto de os dados se encontrarem agregados em intervalos de 5 minutos e de um dia ter 24 horas (1440 minutos).

Depois de efetuados todos os cálculos, verifica-se que cada um dos pórticos eletrónicos apresenta um coeficiente de disponibilidade de dados acima dos 95%, como é possível observar na Figura 3.4.

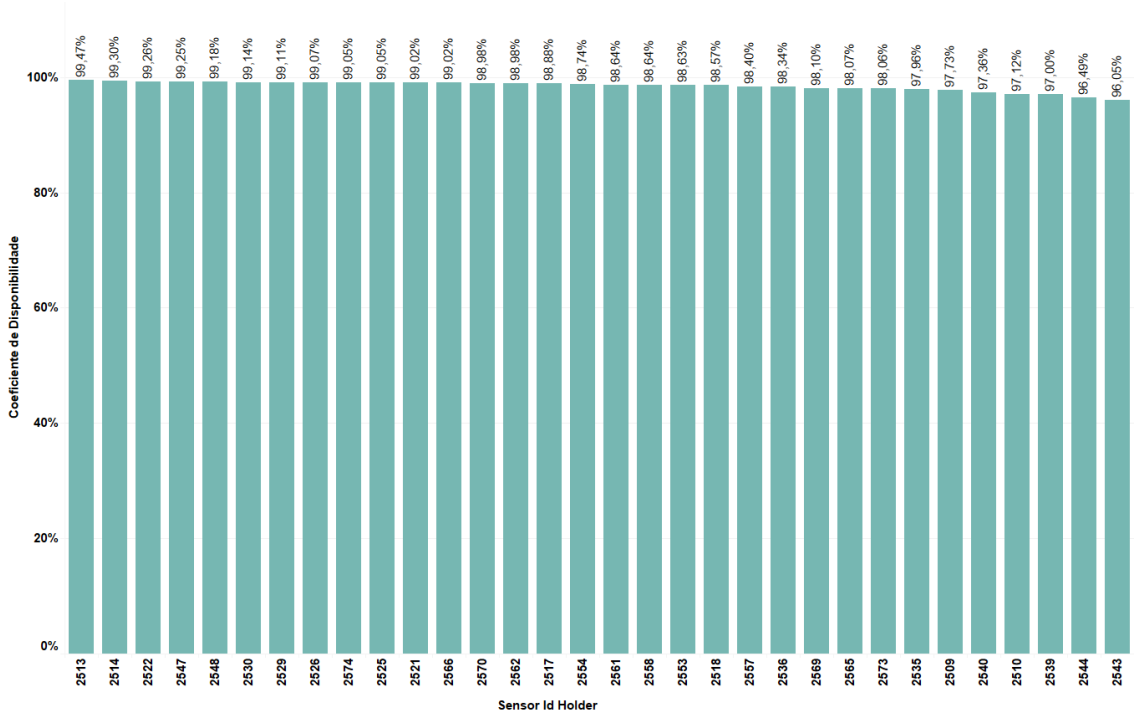


Figura 3.4: Disponibilidade de dados para cada pórtico eletrónico

Estes valores eram expectáveis, visto que a circulação nestas vias se encontra sujeita a aplicação de uma taxa com recurso ao sistema exclusivamente eletrónico sem possibilidade de pagamento manual no local e qualquer avaria durante um longo período de tempo poderia representar um prejuízo para a Concessão responsável pela autoestrada.

No que se refere aos erros de medição, a única variável que é medida ao longo do tempo é o fluxo. Esta variável, representa o somatório do número de viaturas que intersectou um determinado pórtico eletrónico em intervalos de tempo de 5 minutos. Deste modo, os erros de medição que poderão estar presentes no conjunto de dados são:

- **Dados corrompidos:** Existir registos com valores incorretos ou atípicos, como por exemplo fluxo abaixo de zero.

- **Registos duplicados**

Após feita uma análise dos dados, para cada um dos pórticos eletrónicos, constatou-se que estes não apresentam qualquer erro de medição dos acima mencionados.

3.2 Preparação Inicial dos Dados

Nesta fase do processo são feitos alguns ajustes nos conjuntos de dados de forma a melhorar a eficiência e fiabilidade dos mesmos.

Este processo teve início com a importação dos ficheiros *metadata.json* e *sensor_values A25.json* para os softwares de programação Tableau e R, sobre o qual seriam analisados e modelados os conjuntos de dados. Optou-se por converter os ficheiros no formato *JSON* em *CSV*, no sentido de ser mais facilmente gerido e analisado o conjunto de dados fornecidos pelos softwares acima referidos.

Outro dos processos realizados, prende-se com a dimensão do ficheiro *sensor_values A25.json*. Este ficheiro contém todas as medições efetuadas pelos pórticos eletrónicos, perfazendo no total cerca de 20 *Gygabytes* de informação. De modo a aumentar a eficiência na análise e modelação dos dados, procedeu-se à criação de vários ficheiros, em que cada ficheiro corresponde ao nome do código de identificação do pórtico eletrónico (*sensor_id_holder*), contendo as respetivas medições efetuadas.

Face aos coeficientes de disponibilidade apresentados na Secção 3.1.2, para cada um dos pórticos eletrónicos entendeu-se uniformizar a série temporal, considerando a frequência adotada (5 minutos), de modo a ter 288 registos diários. Para tal, procedeu-se à interpolação dos valores em falta utilizando a função "*na.approx()*"da biblioteca "*zoo*"[5] no programa Rstudio, a fim de completar as respetivas séries temporais.

Na Figura 3.5 é possível observar, a título de exemplo geral para um dia aleatório de um pórtico eletrónico, os valores em falta.

[5]https://cran.r-project.org/web/packages/zoo/zoo.pdf

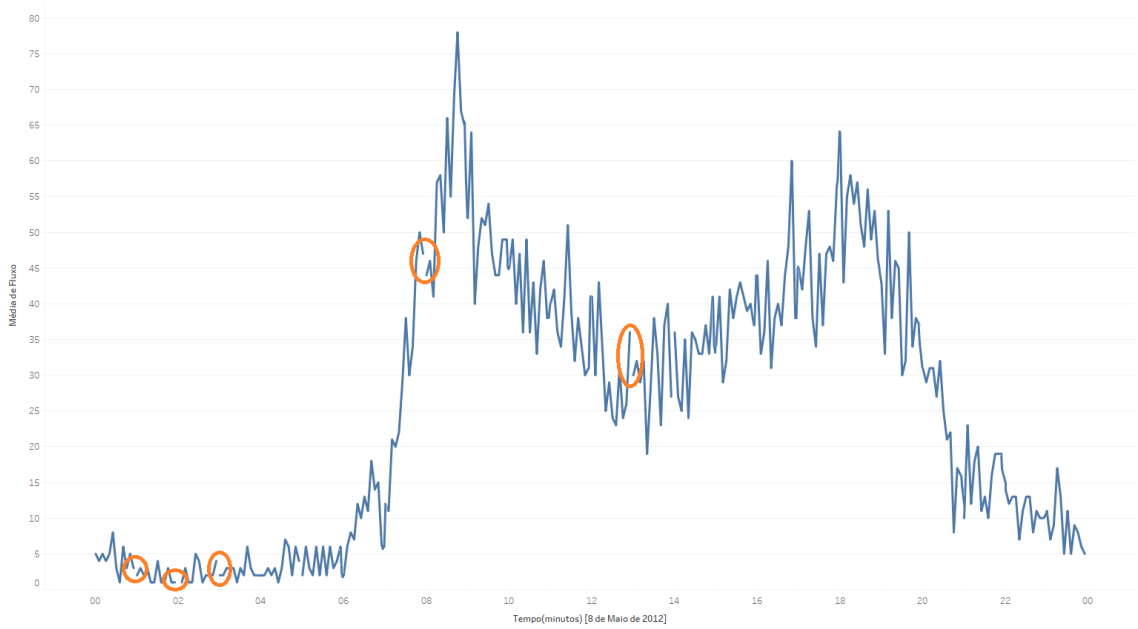


Figura 3.5: Valores em falta

3.3 Análise de Perfis de Utilização da Autoestrada A25

Nesta secção é feita uma investigação sobre diversos fatores que podem influenciar o fluxo rodoviário na autoestrada A25. Esta análise assenta em diversos fatores, tais como a localização da estrada, o horário laboral em vigor no país, padrões de sazonalidade, bem como o próprio dia da semana.

Todos estes fatores tem uma influência significativa no comportamento do fluxo rodoviário, sendo determinantes na caracterização do perfil de uma autoestrada. Por este motivo, a análise aqui desenvolvida, mostra-se fulcral na deteção de perfis de comportamento que podem ser aplicados na estimação das condições do tráfego rodoviário.

Para esse efeito, é realizada uma análise exploratória aos dados históricos com recurso à ferramenta de visualização de dados Tableau em diferentes granularidades de tempo.

3.3.1 Perfil Diário

Quando analisamos o fluxo rodoviário num dia útil, torna-se bastante previsível que o comportamento durante o dia seja diferente do período noturno. Durante o dia, é expectável que o fluxo rodoviário apresente dois picos máximos causados pelo horário laboral. Pelo contrário, durante a noite o fluxo aproxima-se de zero porque existe uma menor utilização das vias.

A Figura 3.7 , referente aos pórticos eletrónicos "2509"e "2544", localizados próximo de zonas com grandes dimensões populacionais (Ver localização na Figura 3.6), é representativa desse mesmo comportamento.

Figura 3.6: Localização dos Pórticos "2509"e "2544"

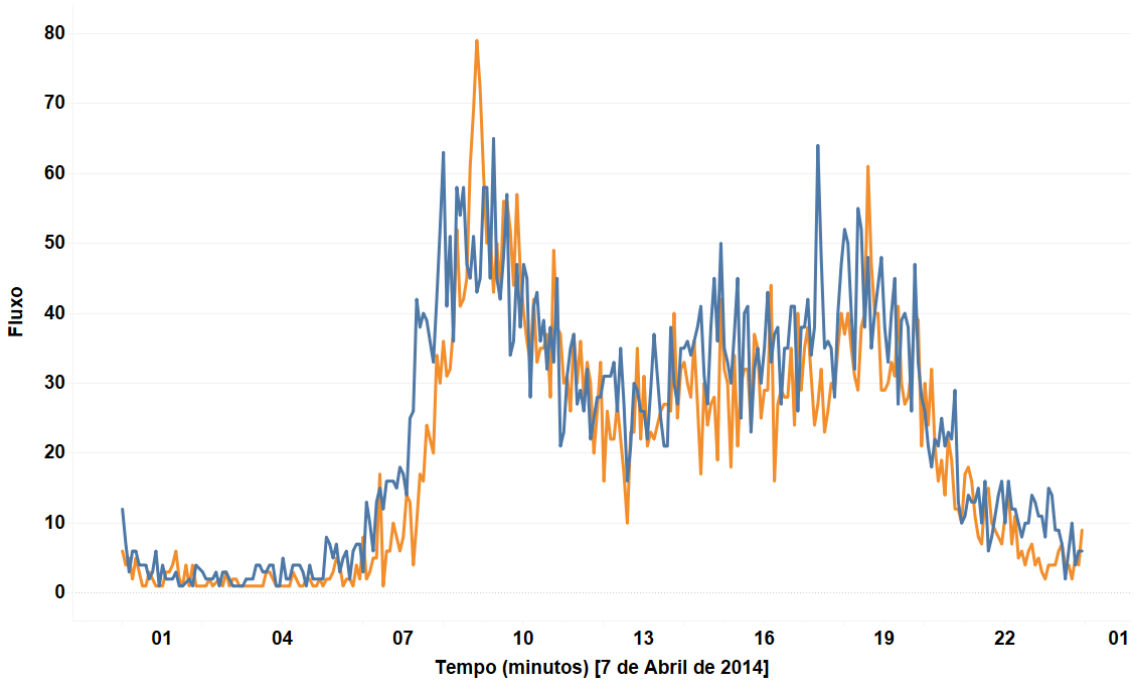


Figura 3.7: Fluxo rodoviário num dia útil, próximo de zonas com grandes dimensões populacionais

Como se pode observar na Figura 3.7, existe um aumento de fluxo rodoviário entre as 8 e as 10 horas da manhã, período durante o qual as pessoas se deslocam para os respetivos postos de trabalho. Da mesma forma, como seria de esperar, ao final da tarde verifica-se um novo pico, por volta das 18 horas, que corresponde ao regresso a casa.

Durante este intervalo, apesar da redução considerável no fluxo rodoviário, continua a existir um número de viaturas significativo nas estradas, devido ao facto deste período corresponder às horas de maior atividade diária.

Desta análise, destaca-se outro fator determinante que se prende com a distribuição

e localização geográfica dos pórticos eletrónicos. Quando comparados com os pórticos eletrónicos localizados junto a zonas rurais, aqueles que se localizam próximo de grandes centros populacionais manifestam comportamentos do fluxo rodoviário mais expressivos, com picos mais exuberantes e maiores amplitudes.

Em oposição a estes, a Figura 3.9 refere-se aos pórticos eletrónicos "2558"e "2573", localizados numa zona habitacional de menor população (Ver localização na Figura 3.8), onde o comportamento do fluxo rodoviário é claramente distinto. Nesta identifica-se uma curva com oscilações mais bruscas mas com máximos menos expressivos.

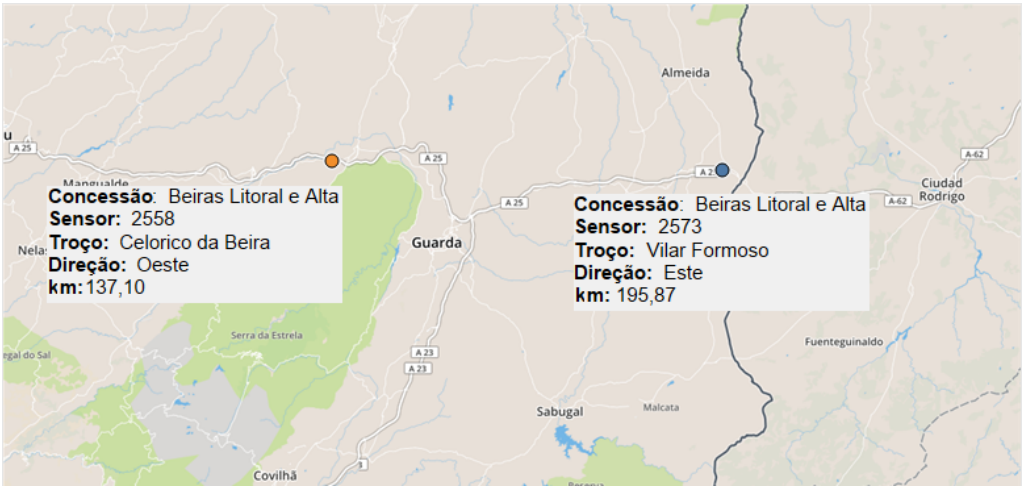


Figura 3.8: Localização dos Pórticos "2558"e "2573"

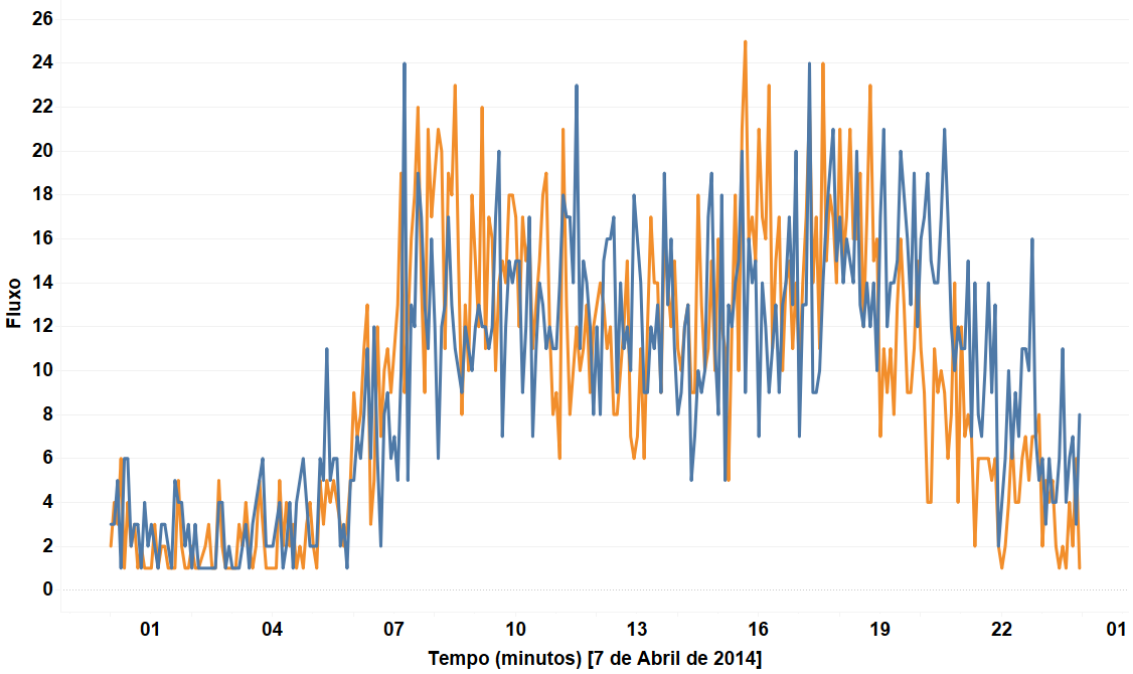


Figura 3.9: Fluxo rodoviário num dia útil, numa zona habitacional de menor população

3.3.2 Perfil Semanal

Outro dos fatores analisados está relacionado com a diminuição do fluxo rodoviário durante os fins de semana. Enquanto que nos dias úteis o comportamento do fluxo rodoviário mantém-se praticamente igual, durante o fim de semana existe um decréscimo significativo no número de carros. Esta diminuição na circulação de veículos deve-se à interrupção do normal horário laboral.

A Figura 3.11, que corresponde à semana de 20 a 26 de Outubro de 2014, referente aos pórticos eletrónicos "2509"e "2510"((Este) e (Oeste) respetivamente) localizados perto da cidade de Aveiro (Ver Figura 3.10), com os dados agregados em intervalos de uma hora, é exemplo disso.

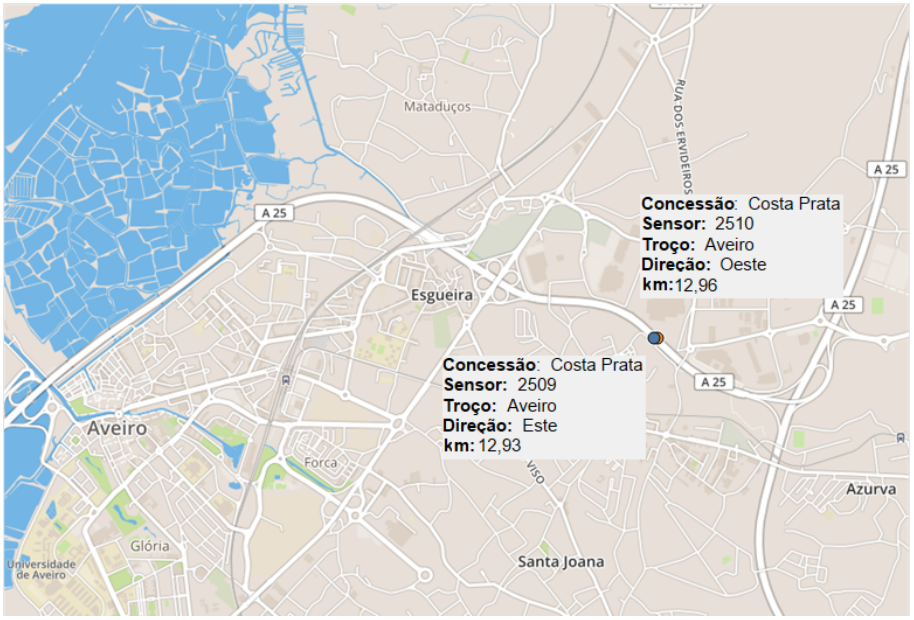


Figura 3.10: Localização dos Pórticos "2509"e "2510"

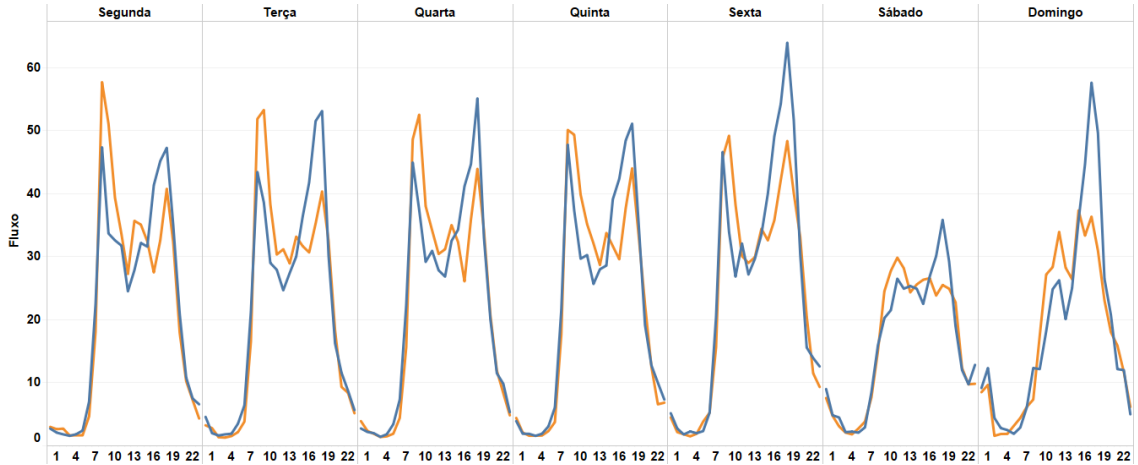


Figura 3.11: Comportamento do fluxo rodoviário semanal

Ao analisar esta figura verifica-se um aumento significativo do fluxo rodoviário durante o período da tarde, na sexta-feira e no domingo. Este comportamento mostrou-se constante durante todos os meses. São várias as razões que podem justificar este padrão, sendo a sexta-feira o ultimo dia útil do horário laboral em vigor no país, é o dia em que as pessoas optam por realizar atividades lúdicas, como a ida a restaurantes, atividades ao ar livre, entre outras. É também neste dia, que muitas pessoas optam por se deslocar para fora das cidades para um fim de semana lúdico ou de visita a familiares. Em consequência da deslocação das pessoas para fora das cidades à sexta-feira, é de esperar que ao domingo se registe um aumento da circulação de veículos durante o período da tarde, devido ao retorno dos cidadãos às suas habitações.

Apesar das várias características desenvolvidas nesta análise, existem situações em que estes padrões não se verificam. Algumas das variações observadas nos comportamentos podem ser explicadas por: feriados, condições atmosféricas adversas, greves e festividades pontuais, como concertos ou comemorações.

3.3.3 Perfil Mensal

Outro fator igualmente importante na caracterização do comportamento do fluxo rodoviário de uma autoestrada, é a sua evolução mensal ao longo do ano. Isto é, durante o ano podem existir alguns meses, durante os quais se verificam uma concentração maior do número de viaturas a circular na autoestrada. Esta distribuição desigual ao longo do ano conduz a períodos de grande procura (a chamada época alta) e a períodos de procura reduzida (época baixa), efeito denominado pela sazonalidade.

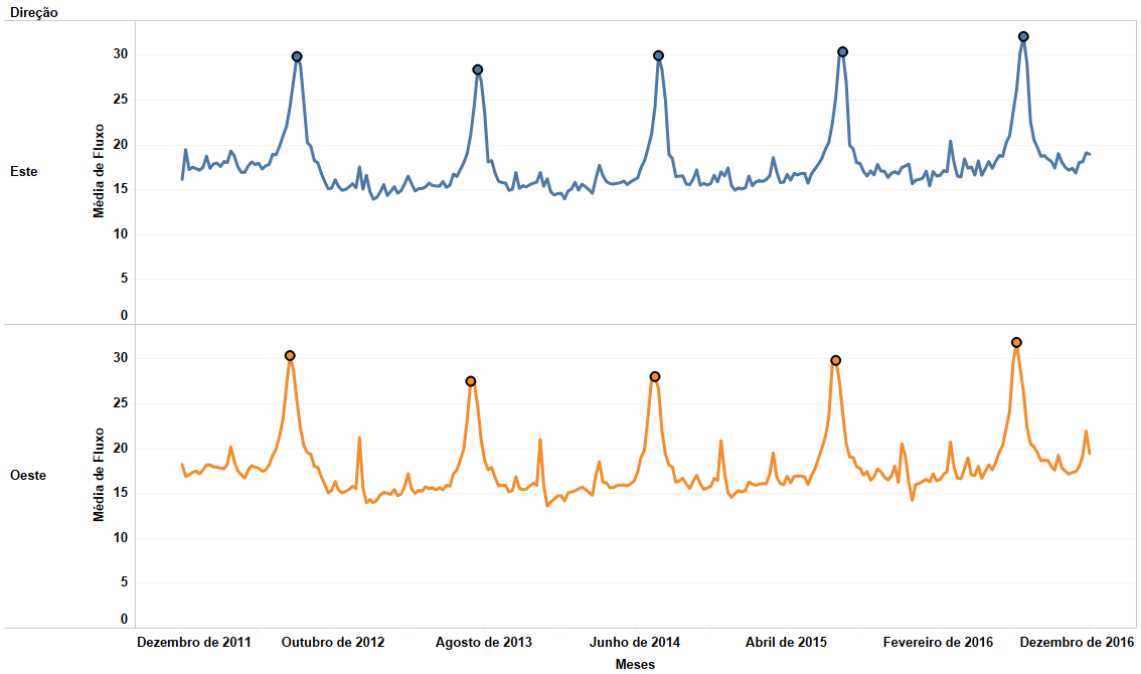


Figura 3.12: Comportamento do fluxo rodoviário Mensal

Após análise dos cinco anos de dados, verifica-se a existência de um padrão sazonal durante os meses de Verão (Julho, Agosto e Setembro), como se pode observar na Figura 3.12 com os dados agregados por mês. A explicação para esse aumento de fluxo rodoviário na autoestrada A25 deve-se ao facto desses meses corresponderem ao período de férias laborais e escolares.

Capítulo 4

Identificação de troços da autoestrada A25 com perfis de utilização semelhante

Neste capítulo é feita uma descrição do processo adotado nesta dissertação relativamente à identificação de troços da Autoestrada A25 com perfis de utilização semelhante para os dias úteis e para os fins de semana. De forma a organizar esse processo, é criada a seguinte metodologia:

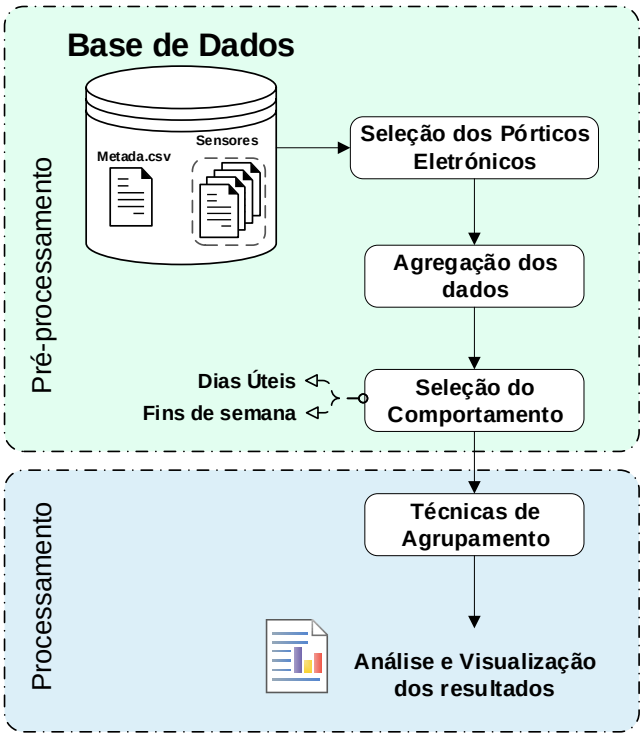


Figura 4.1: Metodologia desenvolvida na identificação de troços da autoestrada A25 com perfis de utilização semelhante

4.1 Pré-Processamento dos dados

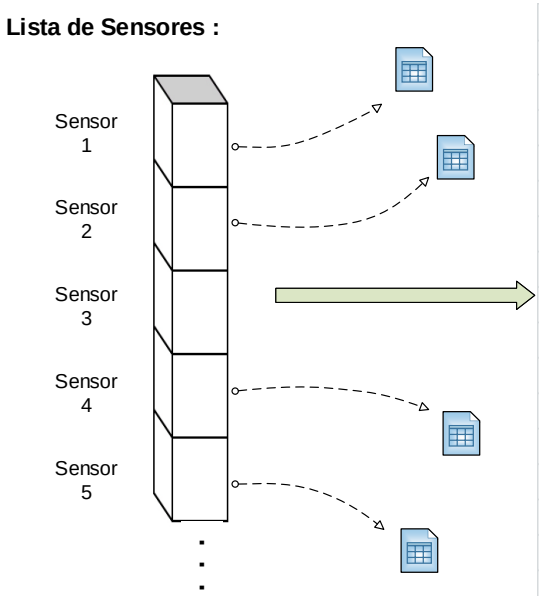
Numa primeira fase, tendo em conta que uma autoestrada apresenta duas direções opostas de circulação de veículos, são escolhidos os pórticos eletrónicos presentes na base de dados de acordo com uma direção pré-definida (Este ou Oeste).

Esta seleção é feita através da leitura dos atributos *sensor_id_holder* e *direction*, contidos no ficheiro *metadata.csv*, em associação com os nomes dos ficheiros que se encontram na diretoria "Sensores". O resultado final deste processo é uma lista contendo 16 pórticos eletrónicos numa dada direção e as suas respetivas medições.

De seguida, é efetuada uma agregação individual dos dados em médias históricas para os dias úteis e para os fins de semana. Ou seja, para cada pórtico presente na lista é realizado a média do fluxo dos dias úteis (Segunda a Sexta), entre Janeiro 2012 a Dezembro de 2016 e em intervalos de tempo de 5 minutos. O mesmo se aplica para o comportamento dos fins de semana (Sábado a Domingo).

Esta agregação tem como finalidade o estudo do comportamento semanal dos diversos troço da autoestrada A25 onde se encontram os respetivos pórticos eletrónicos, a fim de identificar quais os pórticos que apresentam o mesmo comportamento durante os dias úteis ou durante os fins de semana.

A Figura 4.2 representa o resultado final do processo da agregação dos dados aplicado a cada um dos pórticos existente na lista, sendo de destacar a decomposição do atributo *date_time* em *Hora* e *Minuto*, bem como a criação de um novo atributo designado por *Perfil_de_semana* contendo dois valores distintos: 1 e 2. O valor 1 corresponde à média do fluxo dos dias úteis enquanto que o valor 2 corresponde à média do fluxo dos fins de semana.



Perfil_de_semana	Hora	Minuto	Media_Fluxo
...	...	...	...
1	23	10	10
1	23	15	10
1	23	20	10
1	23	25	9
1	23	30	9
1	23	35	8
1	23	40	8
1	23	45	8
1	23	50	8
1	23	55	8
2	0	0	12
2	0	5	12
2	0	10	12
2	0	15	12
...	...	...	...

Figura 4.2: Resultado final do processo de agregação dos dados

Depois de realizada a agregação dos dados, é feita uma extração do tipo de perfil semanal que se pretende analisar. Ou seja, caso se pretenda analisar o comportamento dos dias úteis nos vários troços da autoestrada A25 são extraídas todas as séries temporais relativas aos pórticos eletrónicos presentes na lista cujo o *Perfil_de_semana* seja igual a 1. No final, obtém-se um conjunto de dados contendo várias séries temporais relativas ao tipo de perfil semanal selecionado.

Após executado este pré-processamento de dados, são então aplicadas técnicas de agrupamento com o intuito de agrupar os pórticos eletrónicos que apresentem o mesmo padrão comportamental. Esta análise permite de certo modo identificar quais os troços da autoestrada A25 que apresentam características semelhantes durante os dias úteis ou durante os fins de semana.

4.2 Processamento dos dados

Nesta secção são descritas as técnicas aplicadas por este trabalho na identificação de troços da autoestrada A25 com perfis de utilização semelhante para os dias úteis, bem como para os fins de semana.

Conforme foi abordado no Capítulo 2, na Secção 2.4.2.1, existe na literatura um vasto número de algoritmos de agrupamento. A escolha de um algoritmo ou de outro depende essencialmente do tipo de dados com que se está a trabalhar e dos objetivos que se pretendem alcançar.

Para o processamento de dados, a presente dissertação optou por métodos de agrupamento baseado em repartição. Este tipo de método tem por base a divisão de um conjunto de dados contendo n objetos, padrões ou pontos, em k grupos de dados com características semelhantes, onde cada grupo representa um *cluster* (C_k). Ou seja, os dados são classificados em k grupos, que satisfazem os seguintes requisitos:

- Cada grupo deve conter pelo menos um objeto;
- Cada objeto deve pertencer apenas a um grupo;

Para cada *cluster* C_k existe um ponto $\overline{x}^{(k)}$,designado por centróide, que representa o centro do seu grupo. Neste trabalho, esse ponto é definido como a média de todos os objetos pertencentes ao grupo, ou seja:

$$\overline{x}^{(k)} = \frac{1}{n_k} \sum_{x_i \in C_k} X_i \tag{4.1}$$

onde n_k é o número de objetos pertencentes ao *cluster* C_k.

Para criação dos agrupamentos, os algoritmos de repartição utilizam na sua maioria um critério de agrupamento designado por somatório dos erros quadrados como forma de garantir a formação de grupos compactos e coesos.

Este critério consiste em avaliar a qualidade dos agrupamentos criados através do cálculo das distâncias de cada objeto aos seus centróides. O objetivo essencial é encontrar uma partição tal que, para um número fixo de *clusters*, minimiza a soma dos erros quadrados. Então, a soma dos erros quadrados para um agrupamento contento k grupos é dada por:

$$E = \sum_{j=1}^{k} \sum_{x_i \in C_k} d\left(x_i, \overline{x}^{(j)}\right)^2 \tag{4.2}$$

onde x_i representa o objeto i e $\overline{x}^{(k)}$ o centróide do agrupamento C_k.

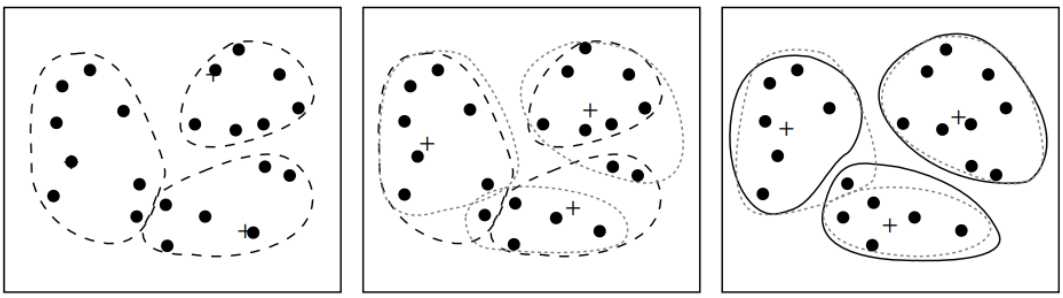
4.2.1 K-Means

Na literatura científica, o *K-Means* é um dos algoritmos mais utilizados dentro dos métodos de agrupamento baseados em repartição. Este algoritmo tem por base o conceito de centróide e aplica o critério do somatório do erro quadrado para agrupar um conjunto de objetos com características semelhantes.

O algoritmo *K-Means*, inicialmente, recebe como parâmetros de entrada um conjunto de dados contendo n objetos e um valor de k agrupamentos a formar. De seguida, o algoritmo atribui de forma aleatória k objetos do conjunto de dados como pontos centrais dos agrupamentos (centróides). Para cada um dos restantes objetos, o algoritmo calcula a distância euclidiana entre o objeto e os centróides, de forma a atribuir cada objeto ao centro de grupo mais próximo.

Depois de feitas as atribuições, são calculados novos pontos centrais dos agrupamentos através da média dos objetos pertencentes ao grupo e o processo volta a repetir a atribuição dos objetos aos novos centros de grupo criados até que seja atingido um critério de convergência ou até que o número máximo iterações seja atingido.

A Figura 4.3 ilustra todo este processo de agrupamento para um conjunto de objetos usando o método *K-Means*.



(a) Agrupamentos Inicial (b) Iterações (c) Agrupamento Final

Figura 4.3: Agrupamento de um conjunto de objetos usando o método *K-Means*, adaptado a partir de [25].

As principais vantagens e desvantagens deste algoritmo são [25]:

- **Vantagens:**
  - Fácil implementação;
  - Eficiente e robusto no agrupamento de dados;
  - Rápido e sem grandes custos a nível computacional;
- **Desvantagens:**
  - Necessita que o número de *cluster* (k) seja previamente inserido no algoritmo;
  - Os resultados finais vão sempre depender dos pontos de inicialização dos centróides;
  - Limita-se apenas a encontrar *clusters* em formatos esféricos;
  - Pontos sem interesse podem levar ao desvio dos centros dos *clusters*;

Em seguida são descritos os principais passos do algoritmo *K-Means*, em formato de código, por forma a resumir tudo o que foi dito anteriormente, de uma forma mais concisa:

Algoritmo 1: K-Means

Entrada:
k: número de *cluster*
D: conjunto de dados contendo n objetos
Saída: Um conjunto de k *clusters*

```
1 início
2     Escolha arbitrária de k objetos do conjunto de dados D como centro inicial dos
        clusters (centróides);
3     repita
4         Atribuir cada objeto ao seu centróide mais próximo, com base na distância
            euclidiana entre o objeto e o centróide;
5         Recalcular o centro de cada grupo;
6     até que as atribuições do cluster parem de mudar ou o número máximo de iterações
        seja atingido;
7 fim
```

4.2.1.1 Determinar o Número de *Clusters*

Tendo em conta a necessidade de especificar o número de grupos a serem inseridos no algoritmo *K-Means*, o qual muitas vezes é considerado como uma desvantagem no uso deste algoritmo, esta dissertação faz uso do método de *Elbow* por forma a encontrar o número apropriado de agrupamentos a serem formados para os diversos conjuntos de dados.

O método *Elbow* consiste na visualização da variação da soma total dos erros quadrados em função do número de *clusters* (k) inserido no algoritmo. A ideia principal é escolher um valor de k no qual a soma total dos erros quadrados cai abruptamente e para o qual a adição de outro agrupamento, $k + 1$, em nada melhora a modelação dos dados significativamente.

Ou seja, é aplicado algoritmo *K-Means* no conjunto de dados para um intervalo de valores inteiros de k=1 até k=10 e para cada valor de k é calculado a soma total dos erros quadrados obtido pelo algoritmo.

De seguida, é produzido um gráfico de linhas contendo a soma total de erros quadrados em função dos valores de k, como se pode observar na Figura 4.4. A partir do valor indicado pelo "cotovelo" no gráfico, significa que não existe ganho em relação ao aumento de do número *clusters*, ou seja, neste caso em concreto o valor apropriado de agrupamentos é três porque adição de de outro grupo em nada altera o somatório do erro quadrado.

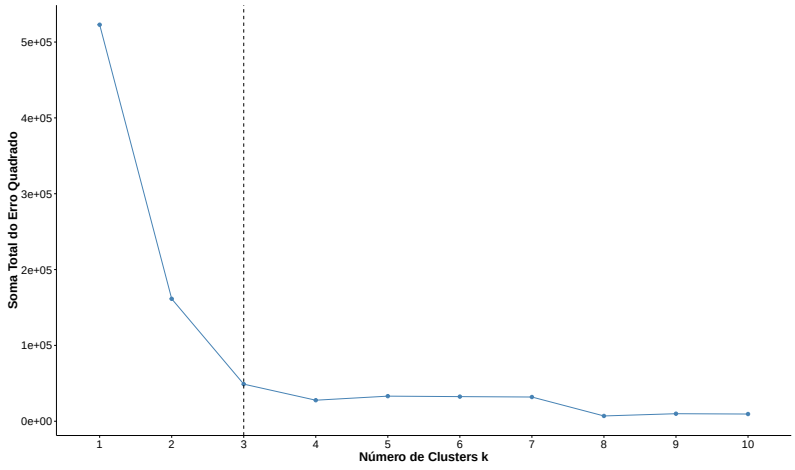


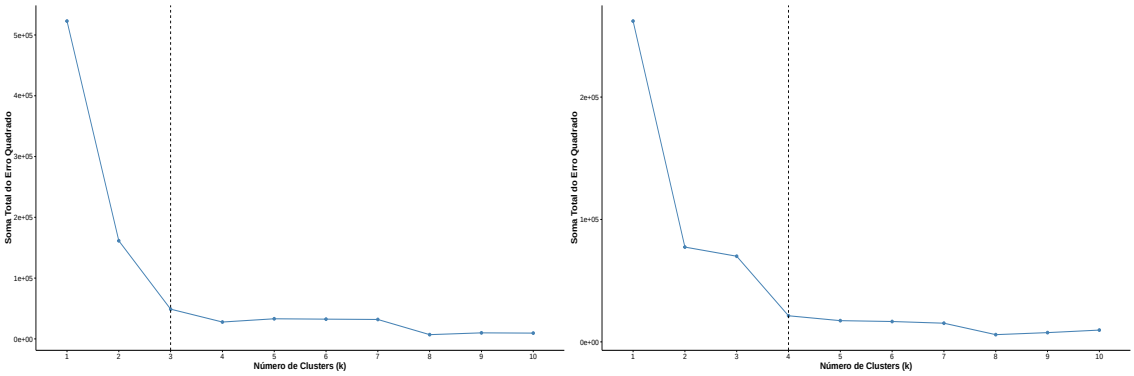
Figura 4.4: Representação gráfica do Método *Elbow*

4.3 Análise e Visualização dos Resultados

Nesta secção são utilizados os vários conjuntos de dados provenientes do pré-processamento realizado na Secção 4.1 e aplicadas as técnicas descritas na Secção 4.2, de forma a agrupar, numa dada direção, pórticos eletrónicos que apresentem o mesmo padrão do fluxo rodoviário durante os dias úteis ou durante os fins de semana. Por sua vez, isto traduz-se

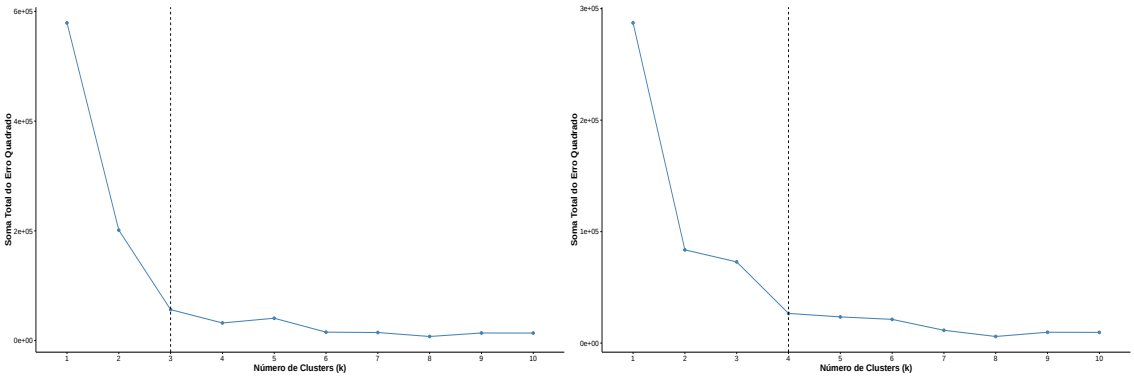
na identificação de troços da autoestrada A25 com perfis de utilização semelhante para os dias úteis e para os fins de semana. Para aplicação do algoritmo *K-Means* e do método *Elbow* são utilizados os pacotes *cluster*[8] e *factoextra*[9] do programa R.

Esta análise encontra-se divida em dois subgrupos, Direção Este e Direção Oeste, de modo a considerar a direção como um fator diferenciador na utilização da autoestrada A25. Inicialmente, é executado o método de *Elbow* nos vários conjuntos de dados, como forma de determinar o número de agrupamentos ideal a ser inserido como parâmetro de entrada no algoritmo *K-Means*. As Figuras 4.5 e 4.6 representam os gráficos obtidos a partir deste método para os vários conjuntos de dados.



(a) (b)

Figura 4.5: Análise do número de agrupamentos ótimos na direção Este para dias úteis (a) e fins de semana (b)



(a) (b)

Figura 4.6: Análise do número de agrupamentos ótimos na direção Oeste para dias úteis (a) e fins de semana (b)

[8] https://cran.r-project.org/web/packages/cluster/
[9] https://cran.r-project.org/web/packages/factoextra/index.html

Numa primeira análise é possível concluir que ambos os sentidos, Este e Oeste, da autoestrada A25 apresentam o mesmo número de agrupamentos ideais para os dias úteis e para os fins de semana. O número de agrupamentos ideais (k) traduz-se no número de grupos de objetos com características semelhantes ideal presente no conjunto de dados. Através da linha a tracejado em cada um dos gráficos acima apresentados, é possível observar os valores de k para os quais o valor do somatório do erro quadrático decresce significativamente e para os quais a adição de outro agrupamento, $k + 1$, pouco ou quase nada altera o soma dos erros quadrado.

Após identificados e definidos os valores k correspondentes ao número de grupos a serem formados, procede-se a aplicação do algoritmo *K-Means* nos vários conjuntos de dados, de forma agrupar pórticos eletrónicos com padrões de fluxo iguais para os dias úteis e para os fins de semana. As Figuras 4.7, 4.8, 4.9 e 4.10 mostram o resultado da aplicação do algoritmo *K-Means*.

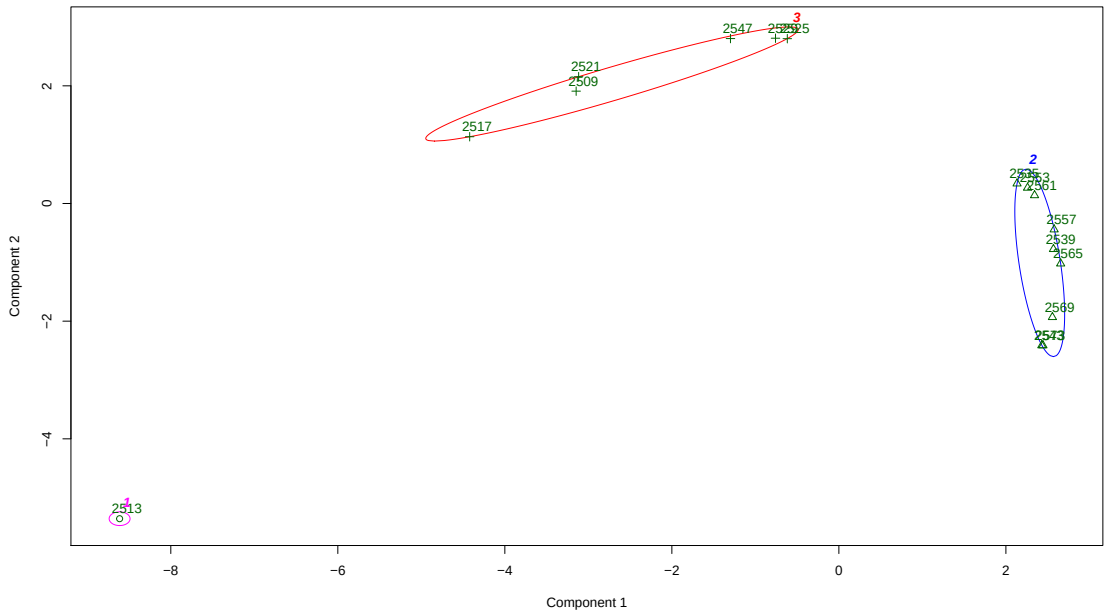


Figura 4.7: Resultado do *K-Means* com k=3 para os pórticos eletrónicos na direção Este e para os dias úteis

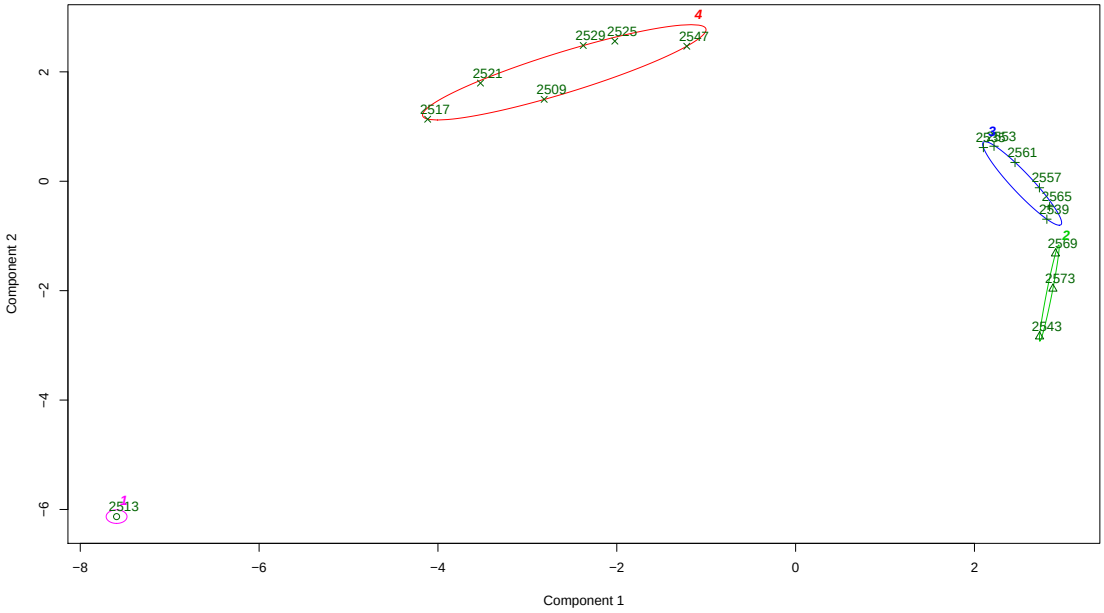


Figura 4.8: Resultado do *K-Means* com k=4 para os pórticos eletrónicos na direção Este e para os fins de semana

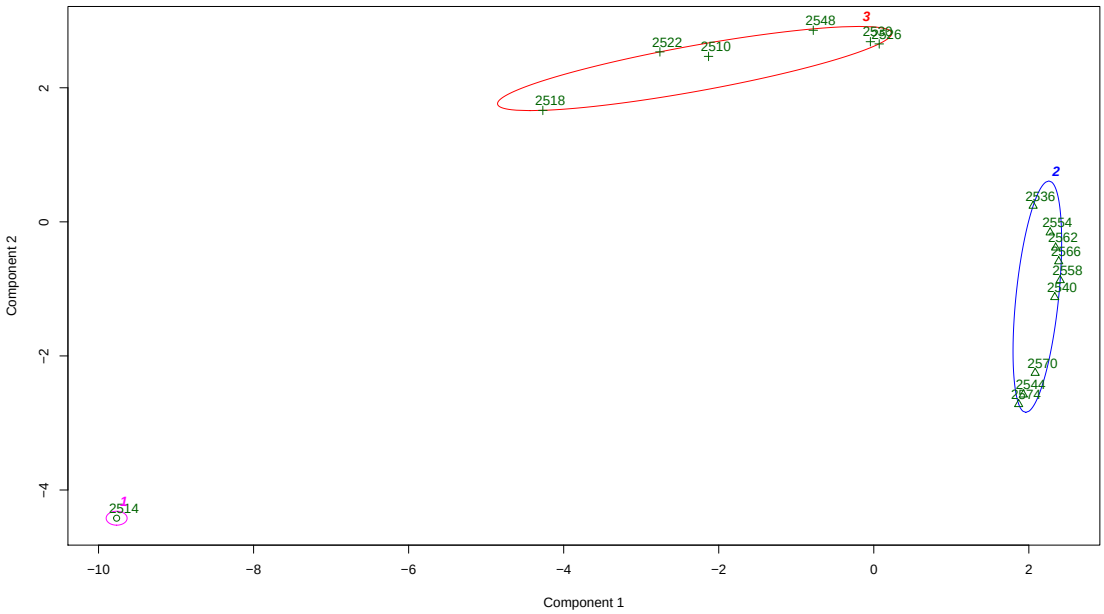


Figura 4.9: Resultado do *K-Means* com k=3 para os pórticos eletrónicos na direção Oeste e para os dias úteis

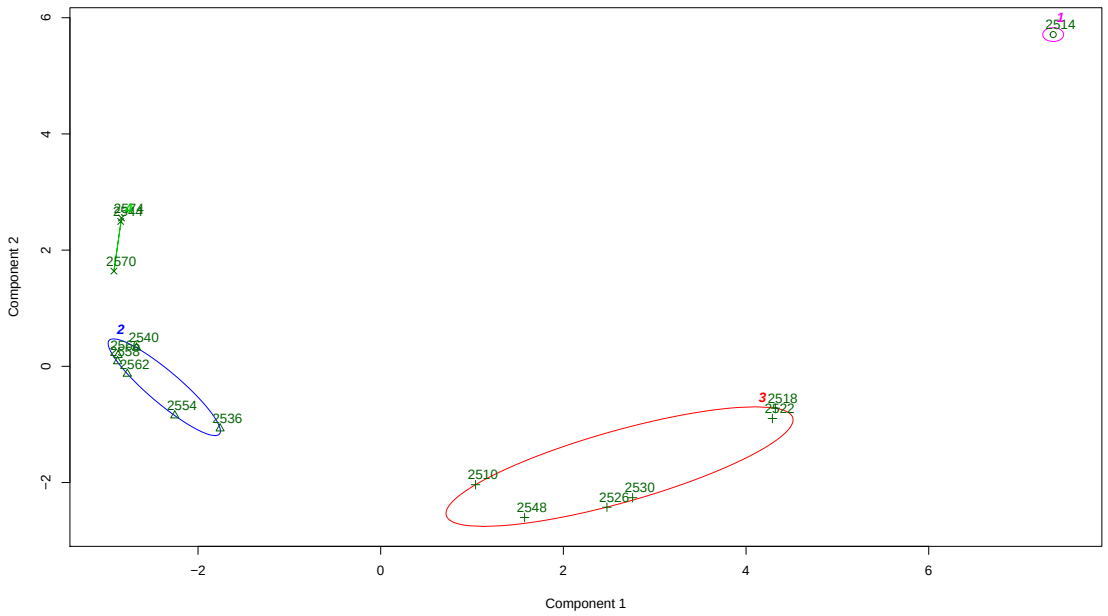


Figura 4.10: Resultado do *K-Means* com k=4 para os pórticos eletrónicos na direção Oeste e para os fins de semana

De modo a retirar conclusões sobre os grupos criados para os diferentes comportamentos analisados, elaborou-se a visualização geográfica dos pórticos eletrónicos consoante o grupo em que se inserem.

Analisando as Figuras 4.11 e 4.12, para o perfil de utilização durante os dias úteis de cada troço da autoestrada A25, tanto na direção Este como Oeste, verifica-se que os pórticos localizados perto do litoral correspondem a um perfil de utilização diferente de pórticos localizados mais no interior.

Figura 4.11: Visualização geográfica dos pórticos com o mesmo perfil comportamental durante os dias úteis na direção Este

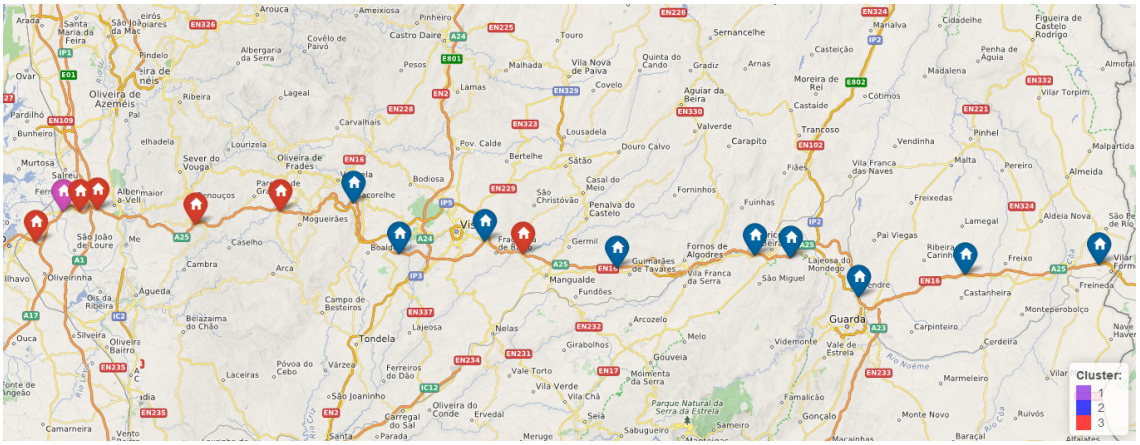


Figura 4.12: Visualização geográfica dos pórticos com o mesmo perfil comportamental durante os dias úteis na direção Oeste

Em relação ao troço Angeja, onde ficam situados os pórticos "2513"(Este) e "2514"(Oeste), verifica-se um comportamento especial (*cluster* 1) devido ao facto de este troço de estrada fazer interseção com várias autoestrada, A29, A17 e A25, o que por sua vez se traduz num aumento bastante considerável de fluxo de carros nessa via.

Relativamente ao comportamento do fluxo rodoviário durante os fins de semana, as figuras 4.13 e 4.14 ilustra os pórticos eletrónicos com o mesmo padrão comportamental. É possível verificar que existe uma maior diversidade ao fim de semana tanto em zonas rurais como zonas urbanas. Isto acontece porque durante este período de tempo as pessoas optam por realizar diversas atividades lúdicas, como ir à praia ou jantar fora, fazendo com que haja uma maior diversidade de utilização das vias.

Figura 4.13: Visualização geográfica dos pórticos com o mesmo perfil comportamental durante os fins de semana na direção Este

Figura 4.14: Visualização geográfica dos pórticos com o mesmo perfil comportamental durante os fins de semana na direção Oeste

4.4 Validação dos Modelos

De forma a validar os grupos criados pelo algoritmo *K-Means* nos vários conjuntos de dados, este trabalho aplica o método de *Silhouette*.

Este método consiste no cálculo dos índices de *Silhouette* ($s(i)$) a fim de avaliar o quão bem os grupos estão separados e compactados. Estes valores são calculados através da seguinte equação:

$$s(i) = \frac{b(i) - a(i)}{\max(a(i), b(i))} \tag{4.3}$$

onde, *a(i)* representa a distância média do objeto i em relação a todos os outros objetos pertencentes ao mesmo grupo do objeto i e *b(i)* a distância mínima do objeto i em relação aos restantes objetos que não pertencem a esse mesmo grupo.

Os índices de *Silhouette* de um objeto i variam entre -1 e $+1$. Quando o valor é positivo e muito próximo de um, significa que este apresenta um grau de correlação bastante elevado com os restantes objetos do próprio grupo. Ao invés, quando o valore de $s(i)$ é negativo, indica que o objeto está mal combinado com os grupos vizinhos. Ao aplicar-se a média sobre todos os índices de *silhoutte* $s(i)$ referentes a um determinado grupo, é possível avaliar o quão bem os *clusters* estão separados e compactados.

As Figuras 4.15, 4.16, 4.17, e 4.18 apresentam os valores de *Silhouette* para cada análise de dados efetuada. É possível verificar que os grupos criados pelo algoritmo *K-Means* apresentam um grau de separação e compactação bastante aceitável devido ao facto de estes apresentarem um índice de *Silhouette* positivo e muito próximo do valor 1, à exceção do *cluster* 1, que se encontra no limiar,formado por apenas um objeto.

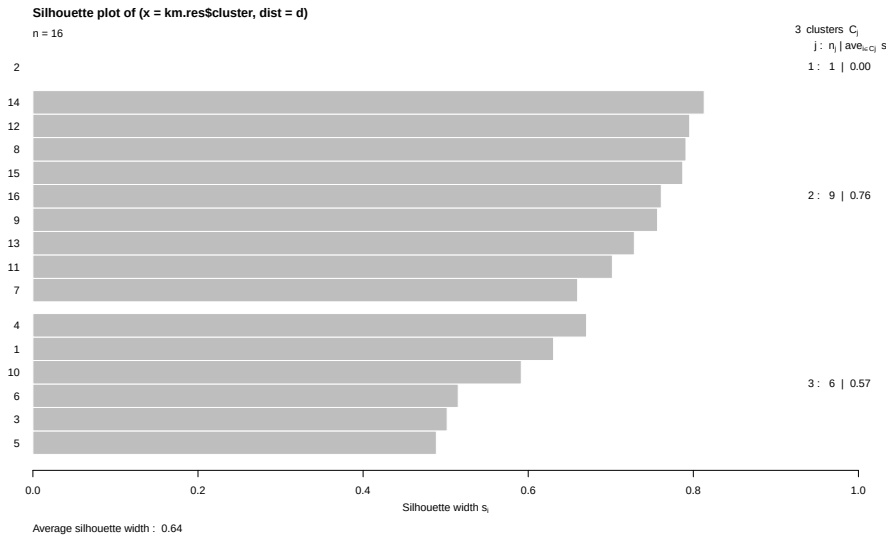


Figura 4.15: Valores de *Silhouette* para os pórticos eletrónicos na direção Este considerando o comportamento dos dias úteis

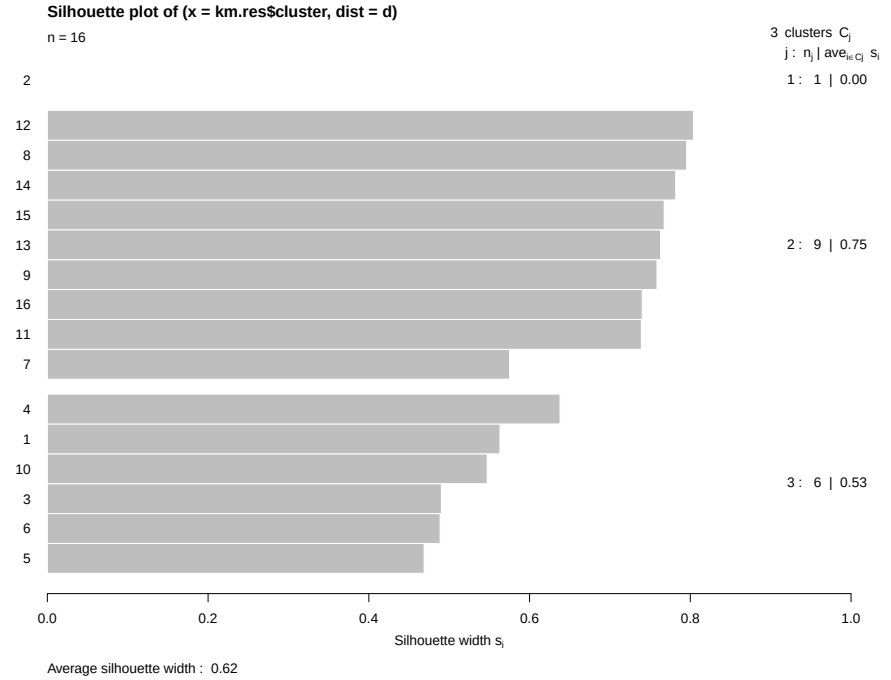


Figura 4.16: Valores de *Silhouette* para os pórticos eletrónicos na direção Oeste considerando o comportamento dos dias úteis

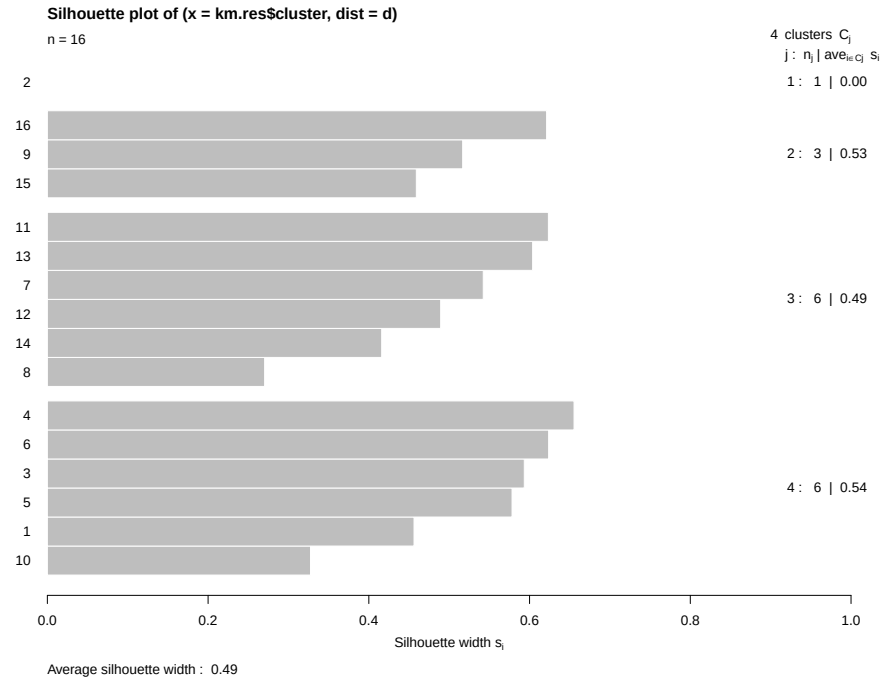


Figura 4.17: Valores de *Silhouette* para os pórticos eletrónicos na direção Este considerando o comportamento dos fins de semana

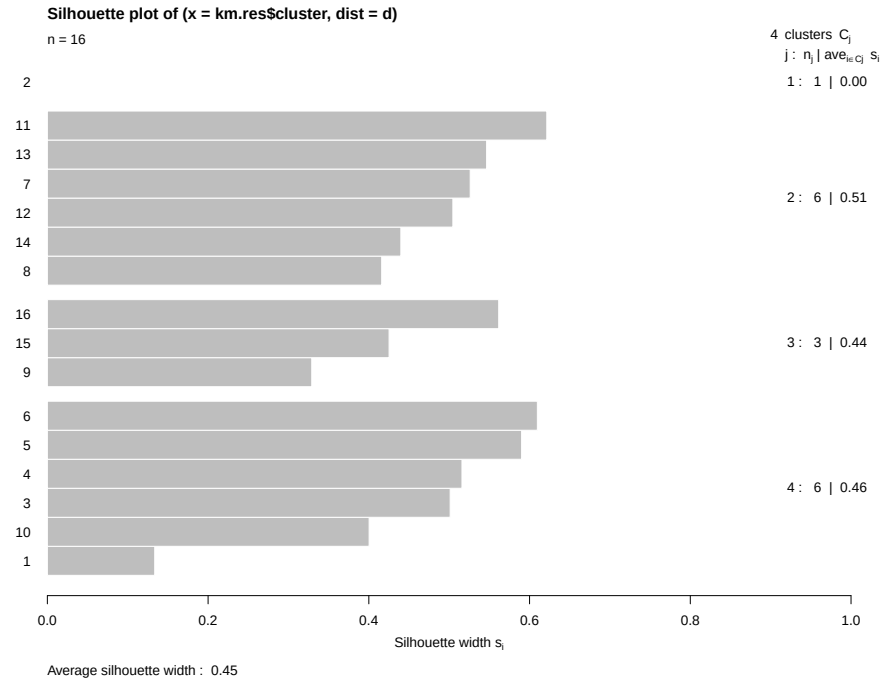


Figura 4.18: Valores de *Silhouette* para os pórticos eletrónicos na direção Oeste considerando o comportamento dos fins de semana

Capítulo 5

Previsão do Fluxo Rodoviário

Neste capitulo são apresentados os métodos utilizados na criação de modelos de previsão do fluxo rodoviário para cada troço da autoestrada A25. Numa fase inicial, é realizada uma descrição do processo de modelação de dados efetuado nos vários conjuntos de dados alusivos aos pórticos eletrónicos fornecidos, a fim de se proceder à aplicação das respetivas técnicas de previsão.

De seguida, são descritas as diferentes técnicas adotadas por este trabalho e apresentados os resultados obtidos para cada um dos modelos implementados através de métricas de desempenho. Os modelos desenvolvidos deverão ser capazes de prever o fluxo rodoviário com base no estado atual da rede e em tempo real.

Devido ao elevado número de dados e à incapacidade de processar os mesmos através de técnicas de previsão atualmente desenvolvidas, os dados considerados são relativos apenas a dois anos (2015 a 2017), o que corresponde a 210528 leituras para cada um dos pórticos eletrónicos. São também selecionados dos trinta e dois pórticos existentes na base de dados, oito pórticos eletrónicos aleatoriamente (observar figura 5.1), para os quais são criados os respetivos modelos de previsão do fluxo rodoviário e apresentados os resultados obtidos.

Este trabalho irá prever o fluxo rodoviário para horizontes temporais de: 15 minutos, 30 minutos e 1 hora.

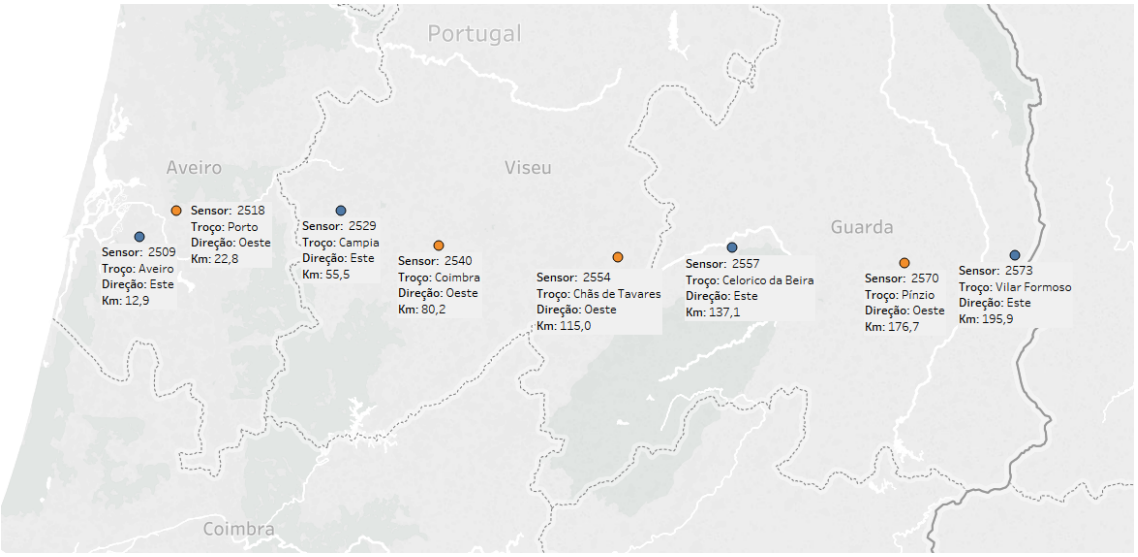


Figura 5.1: Pórticos eletrónicos selecionados

5.1 Modelação dos Dados

No Capitulo 3, Secção 3.3, são analisados vários perfis de utilização da autoestrada A25 em diferentes granularidades de tempo (Perfil diário, Perfil semanal e Perfil Mensal). Dessa análise, é possível destacar diversos fatores que caracterizam o comportamento do fluxo rodoviário em cada troço da autoestrada. Esses fatores são:

- Minutos [0-55]
- Hora [0-23]
- Dia da semana [1-7]
    1. Domingo;
    2. Segunda;
    3. Terça;
    4. Quarta;
    5. Quinta;
    6. Sexta;
    7. Sábado;
- Perfil de semana [1-2]
    1. Dias úteis;
    2. Dias de fim de semana;
- Mês [1-12]

Estes fatores são cruciais na elaboração de um modelo de previsão dado que permitem descrever o comportamento do fluxo rodoviário em cada troço da autoestrada. Através da leitura e da decomposição do parâmetro *date_time*, presente nos vários conjuntos de dados, procede-se à criação destes novos parâmetros.

De seguida, é criado outro parâmetro designado por *Valor_Observado*. Este parâmetro contem os valores da coluna fluxo desfasados consoante o horizonte temporal que se deseja prever, ou seja, se o horizonte temporal for 15 minutos o valor observado deverá corresponder ao valor do fluxo três observações à frente, isto porque os dados se encontram em intervalos de tempo de 5 minutos.

Caso se deseje alterar o horizonte temporal para 30 minutos ou 1 hora deverá-se aplicar respetivamente no *Valor_Observado* um avanço de seis ou doze observações em relação aos valores da coluna fluxo. Esta coluna irá ser usada como variável alvo para o qual os modelos iram aprender a prever através da leitura dos restantes atributos.

É também criado outro parâmetro designado por *Media_Histórica*. Este parâmetro contém a média dos Valores Observados para o mesmo dia da semana, mês e horário. Na Figura 5.2, pode observar-se a comparação entre o *Valor_Observado* e o valor da *Media_Histórica*, para a semana de 12 de janeiro de 2015 a 19 de janeiro de 2015, de um pórtico eletrónico aleatório.

Desta observação conclui-se que a média histórica do fluxo rodoviário já é um bom indicador de como tráfego rodoviário se irá comportar. Este simples prognóstico é usado como o linha de base para comparar diferentes métodos de previsão, bem como amplamente utilizado em várias aplicações práticas como, por exemplo, estimativa do tempo de viagem [14].

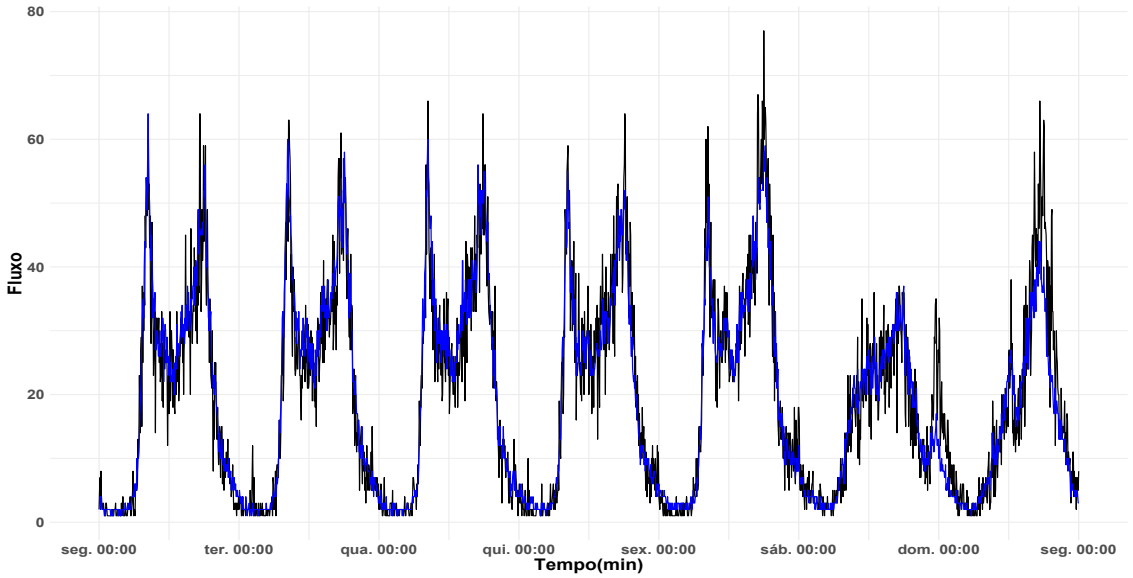


Figura 5.2: Comparação entre o parâmetro *Valor_Observado* (preto) e *Media_Histórica* (azul) para semana aleatória de um pórtico eletrónico

Depois de transformados os dados, obtém-se um conjunto de dados final apto para a aplicação de técnicas de previsão.

Na Figura 5.3 pode observar-se como exemplo geral para um pórtico eletrónico selecionado, a transformação do parâmetro *date_time* nos vários parâmetros anteriormente mencionados, bem como a criação do parâmetro *Valor_Observado* para um horizonte temporal de 15 minutos.

date_time	Fluxo	Valor_Observado	Média_Historica	Mes	Perfil_de_semana	Dia_de_Semana	Hora	Minuto
2015-01-01 00:00:00	2	6	4	1	1	5	0	0
2015-01-01 00:05:00	2	15	5	1	1	5	0	5
2015-01-01 00:10:00	2	10	4	1	1	5	0	10
2015-01-01 00:15:00	6	10	5	1	1	5	0	15
2015-01-01 00:20:00	15	18	6	1	1	5	0	20
2015-01-01 00:25:00	10	6	4	1	1	5	0	25
2015-01-01 00:30:00	10	14	4	1	1	5	0	30
2015-01-01 00:35:00	18	18	5	1	1	5	0	35
2015-01-01 00:40:00	6	8	3	1	1	5	0	40
2015-01-01 00:45:00	14	10	3	1	1	5	0	45
2015-01-01 00:50:00	18	6	3	1	1	5	0	50
2015-01-01 00:55:00	8	11	4	1	1	5	0	55
2015-01-01 01:00:00	10	11	4	1	1	5	1	0
2015-01-01 01:05:00	6	10	3	1	1	5	1	5
...	...	...	...	...	...	...	...	...

Figura 5.3: Conjunto de dados final referente a um pórtico eletrónico

Antes de se dar início à aplicação das respetivas técnicas adotadas neste trabalho, procede-se à divisão do conjunto de dados final em dois subconjuntos principais: um conjunto de dados para treino contendo 80% dos registos (168422 registos) e os restantes 20% (42106 registos) para fins de teste.

O conjunto de treino será utilizado para construir os modelos, enquanto que o conjunto de teste serve para testar os resultados óbitos pelos modelos gerados, de maneira a validar os mesmos.

5.2 Técnicas Aplicadas

Nesta trabalho são escolhidos dois tipos de modelos de previsão : modelos *naive* e modelos não-paramétricos. Os modelos *naive* são aplicados nesta dissertação devido ao facto de serem amplamente utilizados na área de previsão de tráfego a curto prazo como linha de base para comparações de outros modelos mais sofisticados. Convém notar que estes modelos são fáceis de implementar e comportam um baixo esforço computacional. Dentro dos modelos *naive* são aplicadas as técnicas do valor atual e da média histórica localizada.

Em relação aos modelos não-paramétricos é aplicada uma técnica de Machine Learning, designado por Random Forest. É selecionado este algoritmo por se apresentar como gerador de melhores resultados na área da previsão do fluxo rodoviário a curto prazo em [19], tanto ao nível de tempo de processamento como ao nível de precisão.

5.2.1 Valor Atual

O método do valor atual é usado neste trabalho como um método simples e prático na previsão do fluxo rodoviário a curto prazo.

Este método consiste em prever características do tráfego para um horizonte temporal pré-definido com base no estado atual da rede. Se no instante de tempo (t) o valor do fluxo rodoviário corresponder a sessenta viaturas, então para um horizonte temporal (n), a previsão (t+n) corresponderá ao valor do fluxo rodoviário no instante t, ou seja, sessenta viaturas. A Figura 5.4, ilustra a título de exemplo geral, o modelo de previsão baseado no método do valor atual para um horizonte temporal de 15 minutos.

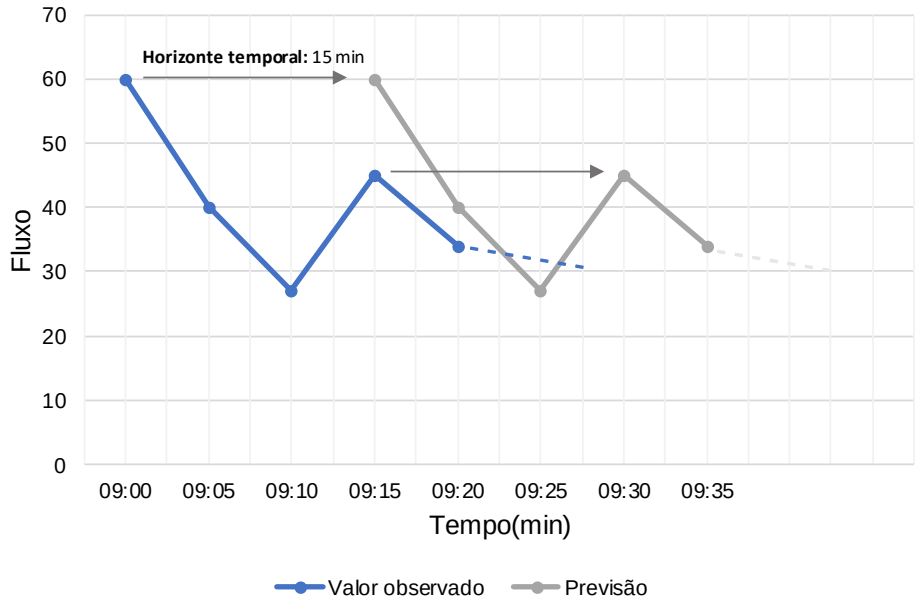


Figura 5.4: Exemplo do método do valor atual

Para previsões a curto prazo, espera-se que a utilização deste técnica apresente bons resultados. No entanto, para previsões a longo prazo é de esperar que esta técnica apresente resultados pouco plausíveis, devido ao facto de o comportamento do tráfego rodoviário ser bastante instável levando a que num horizonte temporal muito elevado o comportamento do tráfego se distancie do valor atualmente lido.

5.2.2 Média Histórica Localizada

Para previsões a longo prazo, este trabalho aplica o método da média histórica localizada para prever características do tráfego rodoviário. Este técnica consiste em calcular a média do fluxo rodoviário com base nos dados históricos, referentes a um determinado mês, dia e horário da semana com o propósito de estimar o fluxo rodoviário.

Por exemplo, para prever o fluxo rodoviário às oito horas de uma segunda-feira do mês de Janeiro, esta técnica irá calcular a média do fluxo rodoviárias de todas segunda-feiras

do mês de janeiro às oito horas presente nos dados fornecidos.

5.2.3 *Random Forest*

Random Forest é um algoritmo de aprendizagem supervisionada que pertence a família dos chamados Métodos Combinados ou *Ensemble Method*. Estes métodos, ao invés de aplicarem apenas um modelo, usam um conjunto de modelos com o objetivo de conseguir melhores resultados através da agregação dos resultados obtidos de cada um dos modelos.

O algoritmo *Random Forest* consiste na criação de um conjunto de árvores de decisão treinadas a partir de um conjunto de amostras aleatória retiradas a partir do conjunto de dados de treino. Além disso, cada árvore é treinada a partir de um subconjunto de atributos presentes em cada amostra de dados.

No final, para a classificação de um exemplo, cada uma das árvores emite a sua decisão e através de um mecanismo de votação (*Bagging*) é elegida a decisão mais votada. Na Figura 5.5 encontra-se uma representação do algoritmo *Random Forest*.

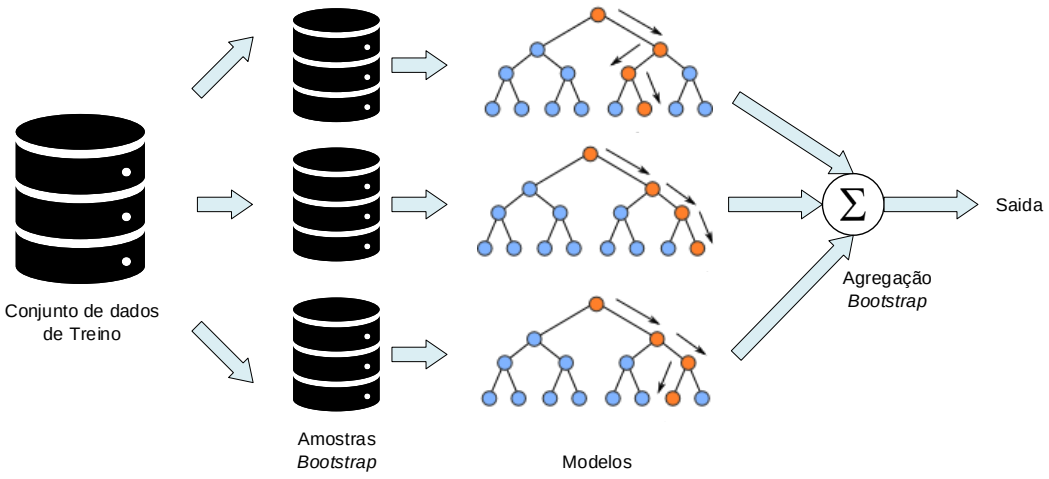


Figura 5.5: Representação do algoritmo *Random Forest*

Este algoritmo pode ser usado tanto para tarefas de classificação como de regressão. Quando aplicado em problemas de classificação, os votos de cada árvore são computados e a classe com mais votos é a resposta final do modelo. Quando aplicado em problemas de regressão, o resultado final é a média dos valores de saída de cada árvore.

As vantagens e desvantagens deste algoritmo são[26]:

- **Vantagens:**
  - Ótimo desempenho em termos de precisão;
  - Fácil configuração por usar um pequeno número de parâmetros;
  - Adequada a grandes quantidades de dados;

  - Funciona bem com dados ausentes;
  - Devido à natureza aleatória na construção de cada árvore, o problema de sobreajuste (*overfitting*) não ocorre tão facilmente;

- **Desvantagens:**
  - Quanto maior o número árvores de decisão a serem inseridas no algoritmo mais lento se torna o modelo;

Para aplicação do algoritmo *Random Forest* é utilizado o pacote *Ranger*[1] do programa R. Os parâmetros de entrada inseridos neste algoritmo são:

- Número de árvores na floresta (*ntree*) = 100;
- Número de atributos utilizados para construir cada árvore (*mtry*) = 2;
- Conjunto de dados para treino;

5.3 Validação dos Modelos de Previsão

Nesta secção são apresentados os resultados obtidos através dos modelos de previsão desenvolvidos usando o conjunto de dados de teste. De forma a aferir a qualidade dos mesmos, são utilizadas as seguintes métricas de erro:

- **Erro Médio Absoluto (Mean Absolute Error (MAE)):**

$$MAE = \frac{1}{n}\sum_{t=1}^{n}|y_t - \hat{y}_t|$$

- **Erro Médio Absoluto Percentual (Mean Absolute Percentage Error (MAPE)):**

$$MAPE = \frac{100\%}{n}\sum_{t=1}^{n}|\frac{y_t - \hat{y}_t}{y_t}|$$

Nas equações acima apresentadas, n representa o número de observações no conjunto de dados de teste, y_t o valor real no instante t e $\hat{y}_t$ o valor previsto para o instante t. Para aplicação destas métricas é utilizado o pacote *Metrics*[2] do programa R.

O Erro Médio Absoluto é uma das métricas mais utilizadas para cálculo de erro de previsão. Esta métrica representa a média dos erros cometidos pelo modelo de previsão ao longo de uma série de períodos de tempo. Outra métrica utilizada neste trabalho é Erro Médio Absoluto Percentual. Esta métrica expressa, em percentagem, a média da diferença absoluta entre os valores previstos e os valores atuais.

[1] https://cran.r-project.org/web/packages/ranger/index.html
[2] https://cran.r-project.org/web/packages/Metrics/index.html

Neste trabalho, a fim de prever o fluxo rodoviário em cada troço da autoestrada A25, são aplicados diferentes horizontes temporais: 15 minutos, 30 minutos e 1 hora. Seguidamente, são apresentados os erros obtidos pelos modelos desenvolvidos para cada um dos pórticos referidos na Figura 5.1.

Tabela 5.1: Resultado dos modelos desenvolvidos para o Pórtico eletrónico "2509"

Modelos	**Métricas**	**Horizonte Temporal**		
		15 min	**30 min**	**1 hora**
Valor Atual	**MAE**	5,74	6,57	13,04
	MAPE(%)	39,74	44,35	92,03
Média Histórica Localizada	**MAE**	5,16	5,12	5,14
	MAPE(%)	35,83	35,73	35,77
Random Forest	**MAE**	4,87	4,95	5,34
	MAPE(%)	33,83	34,19	36,47

Tabela 5.2: Resultado dos modelos desenvolvidos para o Pórtico eletrónico "2518"

Modelos	**Métricas**	**Horizonte Temporal**		
		15 min	**30 min**	**1 hora**
Valor Atual	**MAE**	6,28	7,08	9,25
	MAPE(%)	36,75	40,6	52,37
Média Histórica Localizada	**MAE**	5,73	5,69	5,75
	MAPE(%)	33,48	33,17	33,59
Random Forest	**MAE**	5,32	5,37	5,54
	MAPE(%)	32,26	31,85	32,61

Tabela 5.3: Resultado dos modelos desenvolvidos para o Pórtico eletrónico "2529"

Modelos	**Métricas**	**Horizonte Temporal**		
		15 min	**30 min**	**1 hora**
Valor Atual	**MAE**	5,51	6,01	7,41
	MAPE(%)	39,43	42,58	51,95
Média Histórica Localizada	**MAE**	4,78	4,79	4,81
	MAPE(%)	35,26	35,29	35,86
Random Forest	**MAE**	4,62	4,68	4,78
	MAPE(%)	34,44	34,76	35,49

Tabela 5.4: Resultado dos modelos desenvolvidos para o Pórtico eletrónico "2540"

Modelos	Métricas	Horizonte Temporal		
		15 min	30 min	1 hora
Valor Atual	MAE	3,95	4,14	4,73
	MAPE(%)	47,89	50,41	57,58
Média Histórica Localizada	MAE	3,38	3,38	3,39
	MAPE(%)	44,46	44,30	44,48
Random Forest	MAE	3,30	3,31	3,36
	MAPE(%)	43,47	43,49	44,12

Tabela 5.5: Resultado dos modelos desenvolvidos para o Pórtico eletrónico "2554"

Modelos	Métricas	Horizonte Temporal		
		15 min	30 min	1 hora
Valor Atual	MAE	4,35	4,53	5,18
	MAPE(%)	44,50	46,26	52,42
Média Histórica Localizada	MAE	3,82	3,80	3,81
	MAPE(%)	40,57	40,52	40,26
Random Forest	MAE	3,64	3,65	3,69
	MAPE(%)	36,74	38,91	39,13

Tabela 5.6: Resultado dos modelos desenvolvidos para o Pórtico eletrónico "2557"

Modelos	Métricas	Horizonte Temporal		
		15 min	30 min	1 hora
Valor Atual	MAE	4,37	4,52	5,08
	MAPE(%)	44,27	46,45	52,04
Média Histórica Localizada	MAE	3,54	3,53	3,54
	MAPE(%)	37,15	37,16	37,32
Random Forest	MAE	3,55	3,54	3,60
	MAPE(%)	37,36	37,40	37,76

Tabela 5.7: Resultado dos modelos desenvolvidos para o Pórtico eletrónico "2570"

Modelos	Métricas	Horizonte Temporal		
		15 min	30 min	1 hora
Valor Atual	MAE	3,85	3,92	4,19
	MAPE(%)	51,09	51,82	54,89
Média Histórica Localizada	MAE	3,68	3,67	3,68
	MAPE(%)	48,91	48,73	49,19
Random Forest	MAE	3,26	3,26	3,29
	MAPE(%)	44,01	43,76	44,12

Tabela 5.8: Resultado dos modelos desenvolvidos para o Pórtico eletrónico "2573"

Modelos	**Métricas**	**Horizonte Temporal**		
		15 min	**30 min**	**1 hora**
Valor Atual	**MAE**	3,76	3,84	4,09
	MAPE(%)	48,76	50,06	53,17
Média Histórica Localizada	**MAE**	3,28	3,27	3,28
	MAPE(%)	41,95	42,01	42,04
Random Forest	**MAE**	3,13	3,13	3,14
	MAPE(%)	41,61	41,76	41,57

Observando os valores de erro de cada um dos modelos desenvolvidos relativos a cada pórtico eletrónico da Figura 5.1, verifica-se que o algoritmo Random Forest foi o que apresentou melhores resultados a nível de precisão, à exceção do pórtico eletrónico "2557"alusivo ao troço Celorico da Beira, em direção Este.

Nestes modelos são usados como parâmetros de entrada 100 árvores de decisão e dois atributos aleatórios na construção de cada árvore. A previsão do fluxo é efetuada com base nos atributos *Fluxo*, *Mês*, *Dia_de_semana*, *Perfil_de_semana*, *Hora*, *Minuto* e *Média_Histórica*. É de frisar que estes modelos apresentam um erro médio absoluto bastante baixo, mas um erro médio absoluto percentual bastante elevado.

Como forma de baixar esse erro, uma melhoria importante seria a utilização e integração de outras fontes de dados como, dados de eventos, dados da meteorologia, etc. Outra melhoria importante seria alteração dos parâmetros de entrada do algoritmo RF. Quanto maior o número de árvores de decisão inseridas neste algoritmo melhor precisão o modelo obtém, bem como alteração do número de atributos utilizados na construção de cada árvore. No entanto, devido às limitações de processamento de dados, estes parâmetros não puderam ser alterados.

Relativamente aos modelos *naive* desenvolvidos, o modelo do valor atual demonstrou-se pouco eficaz em comparação com o modelo da Média Histórica Localizada para os vários horizontes temporais. A razão que pode justificar essa diferença é o facto de o comportamento do tráfego rodoviário ser bastante instável levando a que em horizontes temporais muito elevados o comportamento do tráfego se distancie do valor atualmente lido fazendo com que o valor da média histórica obtenha melhores resultados em comparação com o valor actual.

Capítulo 6

Conclusão e Trabalho Futuro

Neste capítulo são apresentadas as principais conclusões do trabalho realizado e algumas recomendações para trabalhos futuros.

6.1 Conclusão

Com a realização deste trabalho, e através de uma metódica análise de dados, foi possível caracterizar e prever diversos tipos de fluxo de tráfego rodoviário, sendo estes relativos a certos troços de autoestrada em Portugal.

Os dados utilizados neste estudo, os quais foram gentilmente fornecidos pela Infraestruturas de Portugal, foram recolhidos por pórticos eletrónicos localizados ao longo de certos troços da autoestrada A25. É de referir, que estes dados relativos à A25, foram selecionados especificamente desta autoestrada pelo facto de esta estar identificada como sendo uma das principais vias de acesso ao país vizinho, tendo também como elemento diferenciador nesta escolha, o elevado fluxo rodoviário, especialmente de veículos de transporte de mercadorias pesados. É de notar que o período de tempo a que se reportam estes dados, está definido entre Janeiro de 2012 e Dezembro de 2016.

De forma a caracterizar o comportamento do fluxo rodoviário de cada um dos troços previamente referidos, foram criados modelos de agrupamento através do algoritmo k-means, sendo assim possível obter diversos tipos de perfil de utilização, os quais, por sua vez, podem ser caracterizados e agrupados consoante a taxa de utilização da via, respetivamente em dois períodos específicos: os dias úteis e os fins de semana.

A implementação destas técnicas de agrupamento, que deu origem a uma série de resultados experimentais, verificou a presença compacta de grupos de pórticos que observaram o mesmo padrão comportamental, o que, do ponto de vista da análise de dados, pode ser devidamente verificado através dos índices de silhueta obtidos, os quais, por sua

vez, caracterizam a coesão dos grupos criados.

Para além do referido anteriormente, outro dos objetivos definidos para a realização desta tese de mestrado, consistia no desenvolvimento de modelos preditivos capazes de prever fluxos relativos a certos troços de autoestrada com base em estatísticas já existentes, considerando para tal, uma análise em tempo real do estado da rede.

Os resultados obtidos nesta análise não foram de encontro ao previamente esperado devido ao facto de o software utilizado se ter deparado com limitações de processamento, as quais, consequentemente, deram origem à criação de maus modelos de decisão. Para além destes inconvenientes, há que referir também, que os modelos selecionados para esta série de estudos, não foram os mais indicados devido à taxa de erro bastante elevada que foram evidenciando ao longo do tempo útil da realização deste trabalho.

6.2 Trabalho Futuro

Considerando as conclusões e limitações no trabalho desenvolvido, há aspetos que podem ser melhorados ou explorados, tais como a utilização de um sistema de processamento distribuído (ou paralelo) que possivelmente permitiria lidar de uma maneira melhor com o elevado volume de processamento de dados.

A inclusão de outras fontes de dados, tais como dados meteorológicos ou de eventos, com vista à melhoria das estimativas dos modelos desenvolvidos, bem como os ajustes necessários aos algoritmos que são aplicados, são também outros aspetos a superar para a obtenção global de melhores resultados.

A metodologia aplicada no estudo destes temas deve ser reformulada, bem como a investigação científica deve passar por novos estágios por forma a serem desenvolvidos melhores algoritmos, tendo em conta que estes são os fatores fulcrais para a realização de trabalhos desta natureza que envolvem uma grande análise estatística.

Bibliografia

[1] E. Commission. “Together towards competitive and resource-efficient urban mobility”. Em: (2013).

[2] E. Commission. “Cities of tomorrow - challenges, visions, ways forward”. Em: (2011).

[3] E. Commission. “Mobilising Intelligent Transport Systems for EU cities”. Em: (2013).

[4] E. T.S. I. (ETSI). *Systems for people on the move*. "https://www.etsi.org/technologies-clusters/clusters/transportation". 2012.

[5] Y. Lin, P. Wang e M. Ma. “Intelligent Transportation System(ITS): Concept, Challenge and Opportunity”. Em: 2017.

[6] G. Piatetsky. *CRISP-DM, still the top methodology for analytics, data mining, or data science projects*. 2014. URL: https://www.kdnuggets.com/2014/10/crisp-dm-top-methodology-analytics-data-mining-data-science-projects.html.

[7] C. Shearer. “The CRISP-DM model: the new blueprint for data mining”. Em: (2000).

[8] E. I. Vlahogianni, M. G. Karlaftis e J. C. Golias. “Short-term traffic forecasting: Where we are and where we’re going”. Em: (2014).

[9] C.-H. Wei e Y. Lee. “Development of freeway travel time forecasting models by integrating different sources of traffic data”. Em: (2007).

[10] L. A. Klein, M. K. Mills, D. Gibson e L. A. Klein. *Traffic detector handbook: Volume II*. Rel. téc. 2006.

[11] M. Shariff, O. C. Puan, N. Mashros e U. J. Bahru. “REVIEW OF TRAFFIC DATA COLLECTION METHODS FOR DRIVERS’CAR–FOLLOWING BEHAVIOUR UNDER VARIOUS WEATHER CONDITIONS”. Em: (1950).

[12] I Seminar e B. Kažič. “Predicting Traffic Flow Based on Heterogeneous Data Sources”. Em: (2014).

[13] M. K. Thivaios. *Multi-source Big Data Fusion Driven Proactivity for Intelligent Mobility*. 2015. URL: http://www.optimumproject.eu/uploads/documents/deliverables/D1.1.pdf.

[14] Hans Van Lint e Chris Van Hinsbergen. “Short-term traffic and travel time prediction models”. Em: (2012).

[15] C P Ij Van Hinsbergen, J. Lint e F M Sanders. "Short Term Traffic Prediction Models". Em: (2007).

[16] M. Moniruzzaman, H. Maoh e W. Anderson. "Short-term prediction of border crossing time and traffic volume for commercial trucks: A case study for the Ambassador Bridge". Em: (2016).

[17] B. Kažič, D. Mladenić e L. Bradeško. "Complex Event Detection and Prediction in Traffic". Em: ().

[18] S. Sun, C. Zhang e G. Yu. "A Bayesian network approach to traffic flow forecasting". Em: (2006).

[19] B. Kazic, D. Mladenić e A. Košmerlj. "Traffic Flow Prediction from Loop Counter Sensor Data using Machine Learning Methods". Em: (2015).

[20] J. Rice e E. Van Zwet. "A simple and effective method for predicting travel times on freeways". Em: (2004).

[21] C. N. Costa, J. V. Coutinho, L. H. de Magalhães e M. A. Arbex. "Descoberta de Conhecimento em Bases de Dados". Em: *Revista Eletrônica: Faculdade Santos Dumont* ().

[22] W. J. Frawley, G. Piatetsky-Shapiro e C. J. Matheus. "Knowledge discovery in databases: An overview". Em: (1992).

[23] U. Fayyad, G. Piatetsky-Shapiro e P. Smyth. "From Data Mining to Knowledge Discovery in Databases". Em: (1996).

[24] D. T. Larose e C. D. Larose. *Discovering knowledge in data: an introduction to data mining*. 2014.

[25] J. Han, J. Pei e M. Kamber. *Data mining: concepts and techniques*. Elsevier, 2011.

[26] M. Dmitrievsky. *Floresta de Decisão Aleatória na Aprendizagem por Reforço*. 2018. URL: https://www.mql5.com/pt/articles/3856.