ANÁLISE DE COMPONENTES PRINCIPAIS DOS DETETORES DE UM SISTEMA INTELIGENTE DE TRANSPORTES

Pedro C. E. Laranjeira
Departamento de Engenharia Civil, Instituto Superior Técnico
Universidade de Lisboa
José Pedro M. P. Tavares
Departamento de Engenharia Civil, Faculdade de Engenharia
Universidade do Porto
João A. Abreu e Silva
Departamento de Engenharia Civil, Instituto Superior Técnico
Universidade de Lisboa

RESUMO
A quantidade de dados gerados pela sociedade tem vindo a crescer rapidamente e a área de Engenharia de Tráfego não escapa a esta tendência. Vários sistemas inteligentes de transporte (ITS) produzem, registam e armazenam dados de tráfego há já alguns anos, sem que, muitas vezes, essa quantidade de dados tenha sido tratada, analisada ou utilizada. O presente trabalho desenvolve ferramentas específicas para o tratamento, visualização e análise dos dados provenientes do ITS do Município do Porto. Os sensores desse ITS serão classificados pela qualidade dos seus atributos, designadamente identificando os detetores que mais contribuem para a explicação do comportamento do tráfego. As metodologias desenvolvidas recorrem às técnicas de deformação dinâmica temporal, algoritmo Euclidiano e análise de componentes principais, de modo a realizar, sem custos, atividades da Engenharia de Tráfego que tradicionalmente mobilizam avultados recursos e tempo.

ABSTRACT
The amount of data generated by the present information society has been increasing rapidly and Traffic Engineering follows this trend. In doing this, several ITS have been producing, recording and storing traffic data for quite a few years. However, this information has had little treatment, analysis or practical use. This research develops specific tools for the treatment, visualization and analysis of data from Porto ITS. The ITS sensors will be classified according to the quality of their attributes, thus identifying the sensors that contribute most to explain traffic behavior. The developed methodologies use technics such as dynamic time warping, Euclidean algorithm and principal component analysis to perform, virtually cost free, traffic engineering related operations that would traditionally involve mobilizing large resources and time consuming. In addition, this work sets the fundamentals basis for creating real time traffic prediction models, based on machine learning techniques.

1. INTRODUÇÃO

O objetivo principal do trabalho é classificar os detetores do ITS da Cidade do Porto, designado por SIGA, quanto à sua importância e qualidade, no âmbito da caracterização do comportamento do tráfego rodoviário no município. Os dados de tráfego do SIGA são obtidos a partir de 111 detetores (*loop detectors*) agrupados em 10 zonas de macro-regulação. Para classificar apropriadamente cada um dos detetores do SIGA, é vital caracterizar o comportamento do tráfego associado a cada *loop*, em particular, assegurando que a informação numérica que caracteriza o perfil dos fluxos rodoviários tem a qualidade suficiente para garantir resultados matemática e estatisticamente fidedignos e robustos. Este tema e respetivos trabalhos foram amplamente desenvolvidos e aprofundados na dissertação Análise de Componentes Principais dos Detetores de um ITS (Laranjeira, 2018).

Os dados recolhidos pelo SIGA, durante mais de 10 anos, nunca foram integralmente utilizados, tratados ou corrigidos, contendo, evidentemente, informação essencial para a caracterização dos padrões de tráfego rodoviário urbano da cidade do Porto. Por outro lado, esses mesmos dados apresentam também uma grande probabilidade de estarem corrompidos pelas circunstâncias associadas à sua operação, nomeadamente devido ao congestionamento, estacionamentos em segunda fila, condições atmosféricas adversas, avarias, etc. Contudo,

tendo-se identificando o potencial intrínseco contido nos dados não sujeitos a qualquer tratamento e posterior análise, é esta informação não utilizada que constitui o verdadeiro combustível do presente trabalho de investigação e que, em última análise, concretiza o objetivo de classificar os detetores principais do ITS.

O primeiro passo a dar foi, portanto, garantir que a informação que caracteriza cada um dos detetores tem o formato correto e que o perfil numérico do tráfego de cada detetor é apropriado, do ponto de vista científico. Para eliminar lapsos, *outliers* e informação corrompida, as rotinas de tratamento de dados contemplam a análise de formato, limiares máximos e mínimos e recurso às técnicas Deformação Dinâmica Temporal (DTW) (Keogh e Ratanamahatana, 2002) e distância Euclidiana (Wikipedia 2018c) para determinar a semelhança entre séries e, deste modo identificando potenciais séries de dados utilizáveis na correção de erros em séries de dados incompletas ou que contêm *outliers*. O segundo passo foi a classificação dos detetores do ITS por meio da técnica estatística Análise de Componentes Principais (PCA) (Stack Overflow, 2016). A PCA é precedida pela implementação de um Filtro de Detetores (FD) que corresponde a um método de apoio à seleção dos detores cujas características numéricas e gráficas do perfil de tráfego são mais representativas da importância e coerência do tráfego registado pelo ITS. É importante referir que o termo *série* se refere a um período diário, ou seja, 288 intervalos de 5 minutos, correspondentes a 24 horas.

A DTW nunca foi utilizada para correção de séries temporais de tráfego pelo que é novidade e permite avaliar cientificamente a utilidade desta técnica não só no âmbito da Engenharia de Tráfego, mas também para outras aplicações futuras. Na prática, Laranjeira (2018) compara a eficiência da DTW com a da distância Euclidiana. A PCA é um algoritmo muito utilizado para a redução de dimensionalidade de dados multivariados, simplificação da visualização de informação e agrupamento de dados. A utilização da PCA foi anteriormente proposta por Foerster (2008), para a determinação da localização ótima de equipamentos de ITS, numa rede rodoviária, tais como painéis de informação de tráfego, e também por Djukic *et al.* (2012), para a redução da dimensionalidade dos dados de tráfego descritos por matrizes origem-destino.

Os resultados do presente estudo vão muito além do objetivo principal do trabalho que foi, efetivamente, identificar os principais detetores que caracterizam o tráfego na cidade do Porto. A informação tratada pode ser utilizada para criar e revelar conteúdos e padrões que de outra forma permaneceriam escondidos nos dados. Na prática, o escopo do trabalho realizado e desenvolvido é vasto e não pode ser facilmente elencado. Por um lado, a própria informação e conteúdo produzidos, podem ser utilizados noutras aplicações e âmbitos, como os mencionados no Capítulo 4, a título de exemplo. Por outro lado, é evidente que a metodologia proposta pode ser implementada em qualquer sistema inteligente de transportes, assumindo particular relevância nos ITS onde a informação tem vindo a ser armazenada e não utilizada.

Após um breve enquadramento, o presente trabalho descreve as metodologias teóricas propostas para a resolução dos problemas atrás focados, nomeadamente das definições das técnicas DTW, distância Euclidiana e PCA e ainda os procedimentos de tratamento dos dados de tráfego do SIGA. Seguidamente, descrevem-se os procedimentos de recolha e tratamento de dados e os resultados gráficos e tabulares dessas operações. Com base nos dados tratados, procede-se à realização da PCA em múltiplos cenários, obtendo-se assim a classificação clara

dos principais detetores e a representação gráfica dessa seriação. Por último, a conclusão sintetiza os resultados obtidos e identifica várias possibilidades de trabalhos futuros.

2. METODOLOGIA

Os dados recolhidos pelo SIGA são armazenados em ficheiros de texto simples que contêm o débito de tráfego acumulado em intervalos de 5 minutos, tendo-se mantido este intervalo para todo o trabalho, modificando-se a escala temporal em função de diferentes objetivos. O tratamento de dados verifica, formata, agrega e compila os dados de tráfego, identificando os erros contidos nos ficheiros originais e, em função da tipologia do erro, atuando de forma preservar a quantidade máxima de dados, com a maior qualidade e integridade possíveis. Os erros identificados nas rotinas de tratamento de dados correspondem a falhas de formatação, número admissível de registos, lapsos de registo e *outliers* pontuais ou em sequência.

As séries são classificadas como séries aceitáveis ou de *outliers*, servindo as primeiras para corrigir as segundas. A correspondência entre a série de *outliers* e a série aceitável é estabelecida com recurso à DTW e à distância Euclidiana, mantendo o maior número possível de registos e, logicamente, salvaguardando a qualidade dos dados substituídos. A visualização dos dados é utilizada para validação da qualidade dos procedimentos de tratamento de dados e para simplificar a representação dos dados de modo a facilitar a compreensão da informação numérica. A linearidade de representação dos dados de tráfego é especialmente relevante na disseminação de conteúdos e na análise de padrões de tráfego.

Toda a informação tratada é armazenada em ficheiros de texto formatados que contêm o instante em que um dado registo foi realizado e o respetivo registo. O principal atributo dos dados formatados é o *timestamp* e a sua correta formatação é fundamental para levar a cabo as operações de escalonamento temporal que de outra forma seriam muito demoradas.

2.1. Comparação de séries temporais

A comparação de séries temporais é uma sub-rotina da substituição de séries de *outliers* e foi efetuada com recurso à DTW e à distância Euclidiana. O critério de semelhança espaciotemporal (Lopes, 2012) adotado assegura os melhores resultados no preenchimento de lacunas. Determina-se o valor da correlação de Pearson (r) (Wikipedia, 2018b), entre a série de *outliers* e as curvas com maior semelhança, para comparação da magnitude e qualidade da semelhança estabelecida por cada uma das técnicas. A semelhança entre curvas corresponde à minimização das funções objetivo da DTW e da distância Euclidiana. Conclui-se que a DTW obtém resultados de semelhança gráfica superiores à distância Euclidiana e que tal não é evidenciado pelo coeficiente de correlação r. Na figura seguinte podem visualizar-se as representações geométricas da distância Euclidiana, à esquerda, e da DTW, à direita.

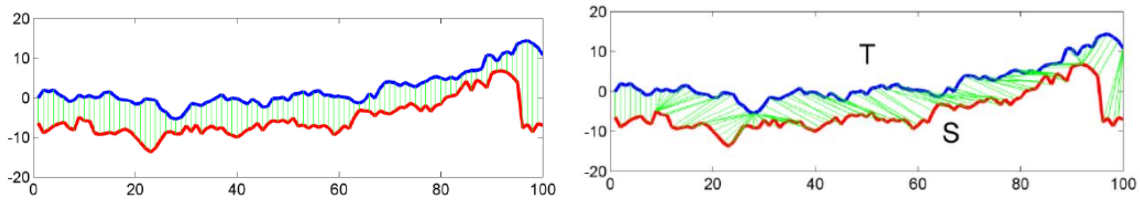

Figura 1: Representações geométricas dos algoritmos Euclidiano e DTW (Cassissi *et al.*, 2012)

A DTW é uma técnica de cálculo não linear que oferece flexibilidade no ajuste entre funções. Os coeficientes *r* obtidos para a DTW são inferiores aos obtidos com a distância Euclidiana pois o funcionamento da primeira assenta na determinação do caminho de menor custo, ao contrário da minimização da distância entre pares de dados homólogos, característico da distância Euclidiana e da correlação de Pearson. A DTW atende à natureza estocástica do tráfego, fenómeno em que se verificam variações de volume em instantes ou períodos homólogos. Apesar da obtenção de resultados de maior qualidade, a complexidade O(*mxn*) da DTW (Stack Overflow, 2016) implica tempos de processamento até 100 vezes superiores aos da distância Euclidiana (Ratanamahatana e Keogh, 2004). De modo a mitigar o problema, utilizaram-se técnicas de *early abandoning* e *lower bounding*, respetivamente as Bandas de Sakoe-Chiba (Sakoe e Chiba, 1978) e o LB_Keogh (Keogh, 2012), que podem ser utilizadas em conjunto (Minnaar, 2014), como se pode verificar nos gráficos da Figura 2.

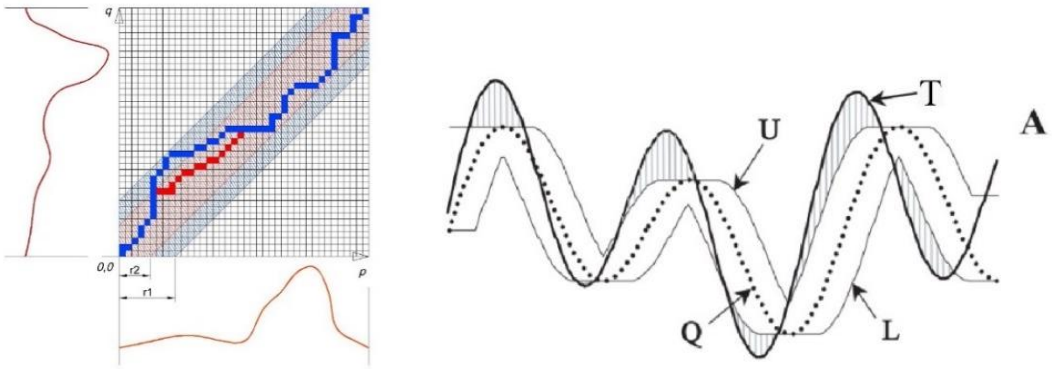

Figura 2: Banda de Sakoe-Chiba (à esquerda) e aplicação do algoritmo LB_Keogh à direita (adaptado de Keogh e Ratanamahatana, 2002)

Os valores de *reach* e do *window* fixaram-se recorrendo aos critérios normalmente adotados em trabalhos de investigação (Ratanamahatana e Keogh, 2004) para as bandas de Sakoe-Chiba e do envelope do LB_Keogh, respetivamente, e correspondem a 10% do número de registos, isto é, o valor de 30 para ambos os parâmetros. Para além destas técnicas de restrição de soluções, os conjuntos de séries aceitáveis e de *outliers* foram separados em dias úteis e dias não úteis, reduzindo o tempo de pesquisa por séries semelhantes em 28,5%, para as séries de *outliers* em dias úteis, e em 71,4% para os fins-de-semana.

2.2. Metodologia de substituição em séries de *outliers*

As séries diárias de *outliers* identificam-se pelos registos superiores ao débito máximo de veículos por intervalo de 5 minutos (120 veículos por via), pela existência de débitos nulos e, ainda, por uma quantidade máxima de registos anormais (superiores ou nulos), correspondentes a 33% do número de registos diários (288). Após a identificação da série

aceitável mais semelhante à série de *outliers*, procede-se à substituição dos valores anormais pelos valores normais da série aceitável.

No manual da CCDR-N (CCDR-N, 2008) preconiza-se um débito máximo teórico de 1900 uve/h/via, para um nível de serviço E com velocidade máxima de 70 km/h, em estradas de vias múltiplas e indica, também, que a capacidade para uma estrada de duas vias é de 1700 uve/h/via. A generalidade dos detetores do SIGA recolhem dados em múltiplas vias, admitindo-se, portanto, uma capacidade aproximada de 1800 uve/h/via numa rua urbana. Todos os detetores estão associados à regulação de intersecções semaforizadas, facto que reduz o débito máximo em cerca de 30%. Considerando ainda um fator de correção empírico de 1,15, admitem-se finalmente 120 veículos por via e por 5 minutos para o limiar máximo de débito, conforme a expressão matemática (1):

$$Vp_{5min} = \frac{1800 \times 0.70}{60} \times 5 \times f_c \approx 120 \quad (1)$$

Em que V_{p5min} representa o débito acumulado por via e por 5 minutos e f_c representa um fator de correção empírico de 1.15 [adimensional].

Caso sejam identificados *outliers* pontuais, o valor irregular é substituído pela média dos valores vizinhos. No caso dos *outliers* sequenciais, apenas os valores anormais são substituídos pelos valores homólogos da série aceitável mais semelhante identificada pela técnica de semelhança. Em ambos os casos, assume-se que a substituição de valores nulos durante o período entre as 00:00 e as 06:00 não enviesa análises futuras dado que o débito de tráfego neste período é normalmente reduzido. Após a correção das séries, procedeu-se à compilação dos dados tratados em cada detetor, que contêm, finalmente, a série temporal, devidamente formatada, corrigida, sem lacunas e sem valores anormais.

2.3. Recolha e tratamento de dados

Os detetores foram caracterizados quanto à sua localização geográfica e, especialmente, quanto ao número e tipo de vias monitorizadas. Além dos resultados tabulares, são também elaborados vários gráficos que permitem a visualização da informação de forma agregada e confirmar graficamente a semelhança estabelecida entre séries de *outliers* e aceitáveis, a existência de lacunas e a consistência da recolha dos dados. A informação corrigida e formatada é agregada em 5, 15 e 60 minutos e produziram-se gráficos com dados diários, semanais e mensais para todos os detetores (Figura 3).

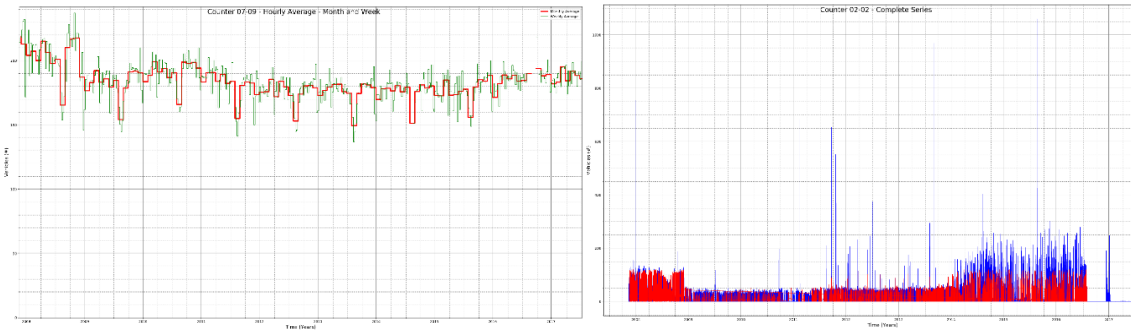

Figura 3: Volume de tráfego médio horário, por mês e semana (à esquerda) e verificação da extensão temporal total da série temporal (à direita)

Determinaram-se as horas de ponta de cinco intervalos horários, por cada dia da semana, e para todos os detetores, nas formas tabular e gráfica. Verificou-se que o tratamento automático de dados substituiu os erros de registo e *outliers* com sequências temporais provenientes de séries aceitáveis. No caso particular do presente estudo, a distância Euclidiana é cerca de 14 vezes mais rápida do que a DTW, garantindo, apesar de tudo, resultados apropriados.

2.4. Metodologia da análise de componentes principais

A PCA tem duas aplicações principais. A primeira está relacionada com a visualização da informação e a segunda com a redução dimensional, simplificações de dados multivariados e a classificação e agrupamento de informação. O presente estudo utilizou ambas as aplicações.

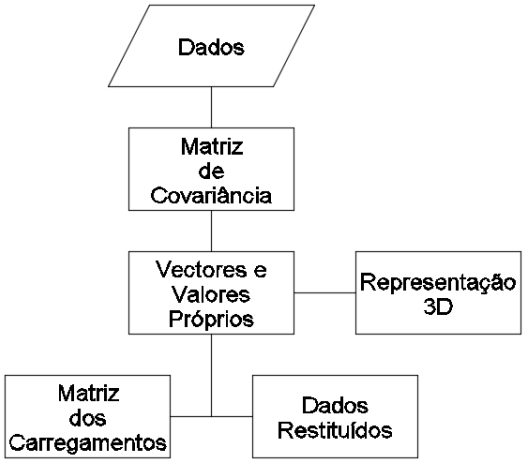

Figura 4: Fluxograma do algoritmo de análise de componentes principais

Na Figura 3, podem identificar-se as principais fases de uma análise de componentes principais. Fazendo corresponder uma única variável a um único detetor, a dimensionalidade inerente à caracterização do tráfego no Porto materializa um espaço vetorial com 111 dimensões. A quantidade elevada de variáveis dificulta a avaliação dos dados e o ruído presente nos registos condiciona a utilização eficiente da informação. Estas condições fazem com que a PCA seja uma técnica apropriada para determinar os detetores principais, em particular por meio da contração dimensional.

2.4.1. Visualização de dados

A PCA facilita a visualização dos dados partindo da construção dos chamados vetores próprios (Wikipedia, 2018d). O número total de vetores próprios coincide com o número de variáveis que descrevem o fenómeno analisado. Os vetores próprios são calculados recorrendo à variância intrínseca dos dados, contida em cada uma das variáveis. Por exemplo, selecionando os três vetores próprios caracterizados por valores próprios máximos, é possível formar um sistema de coordenadas tridimensional que possibilita, por exemplo, a visualização dos dados com mais de três dimensões (Figura 5).

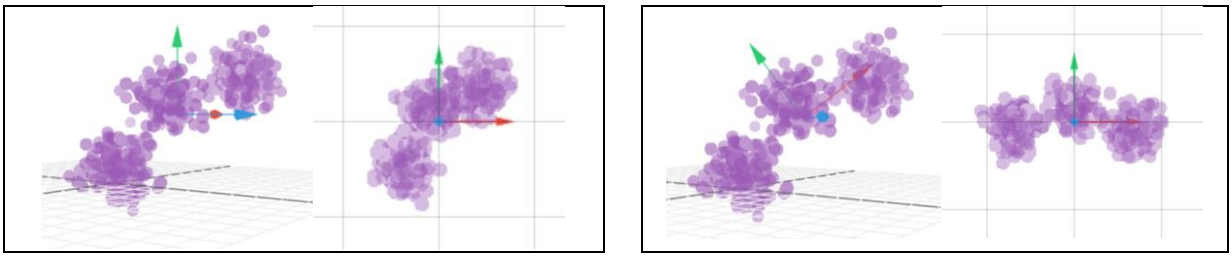

Figura 5: Visualização de dados num referencial típico (à esquerda) e no referencial de componentes principais (à direita) (Powell, 2015)

Na figura anterior, partindo da observação do referencial de componentes principais, é também possível distinguir subconjuntos, ao passo que, no referencial original, tal distinção é visualmente difícil ou mesmo impossível.

2.4.2. Redução de dimensionalidade

Reduzir a dimensionalidade significa explicar um fenómeno recorrendo a um menor número de variáveis do que as originalmente necessárias, sem perda de precisão e rigor. Ao contrário da visualização, a redução de dimensionalidade tem em conta o número de vetores próprios necessários à explicação do fenómeno. A contração dimensional pode também ser utilizada para a restituição dos dados originais, ou seja, utilizando um número inferior de vetores próprios é possível restituir os dados expressos numa dimensionalidade inferior para a dimensionalidade original. Os dados restituídos podem, por exemplo, ser utilizados em modelos preditivos, com a vantagem de apresentarem uma natureza menos estocástica e maior uniformidade.

2.4.3. Matriz primária e matriz de covariância

Partindo de dados verificados e tratados, a PCA tem início com a construção da matriz primária que é constituída pelos valores de tráfego horário acumulado, em que às linhas corresponde o *timestamp* e às colunas da correspondem detetores. A intersecção entre cada linha e coluna corresponde ao volume de tráfego horário acumulado. Não se admitem linhas ou colunas completamente nulas nesta matriz, sob pena de enviesamento dos resultados da análise de componentes principais.

A matriz da covariância é uma matriz simétrica que resume a covariância entre as variáveis que constituem a matriz primária. Já que a cada detetor corresponde uma variável, a matriz da covariância estabelece a correlação de Pearson entre os detetores.

2.4.4. Determinação do número de componentes principais

A partir da matriz de covariância é possível definir os valores e vetores próprios (Wikipedia, 2018d). Os vetores próprios são vetores particulares de uma matriz, cuja principal direção é caracteristicamente constante e evidente na presença de uma transformação linear algébrica. A magnitude, ou norma, do vetor próprio corresponde a um escalar designado por valor próprio, pelo que o número de valores próprios coincide com o número de vetores próprios, ou, neste caso, detetores.

A contração dimensional significa, simplesmente, selecionar um número inferior de vetores próprios cuja importância relativa, implicitamente representada pelo seu respetivo valor próprio, explica a maior quantidade de variância presente nos dados. A quantidade relativa de variância é expressa pela magnitude do valor próprio, passando os vetores próprios a designar-se por componentes principais. Os testes usualmente utilizados para determinação do número de componentes principais são a regra de Kaiser, teste de Cattell e a percentagem de variância explicada acumulada (PVE) (Wikipedia, 2018a). No presente trabalho adotou-se o limiar de 95% da PVE para determinação do número de componentes principais.

2.4.5. Matrizes dos carregamentos e dos dados restituídos

A matriz dos carregamentos indica o contributo de cada detetor para cada componente principal. Nesta matriz, as linhas representam os componentes principais, as colunas representam os detetores e os valores, ou carregamentos, da matriz representam a importância de cada detetor em cada componente principal.

A matriz dos dados restituídos obtém-se a partir dos dados existentes no espaço vetorial dos componentes principais. A restituição dos dados recorre a uma contração dimensional e pode conter resultados apresentando volumes de tráfego “negativos”, condição resultante do cálculo estatístico. De facto, valores negativos na matriz dos dados restituídos são comuns e devem ser tidos em conta, caso se pretenda realizar análises com os dados contraídos.

3. ANÁLISE DE COMPONENTES PRINCIPAIS

A análise dos dados resultantes da aplicação das metodologias descritas revela a existência de vários detetores cujas séries temporais podem enviesar a análise de componentes principais. De modo a garantir que a PCA produz bons resultados, utilizou-se o FD para identificação dos detetores mais adequados a utilizar na análise de componentes principais. O FD observa quatro critérios: a regra dos 50%, a sobreposição transversal, a análise gráfica das séries e o volume de tráfego médio horário.

3.1. Filtro de Detetores

A regra dos 50% relaciona a quantidade de séries diárias aceitáveis com a quantidade de séries de *outliers* de um detetor. Nos casos em que o número de séries de *outliers* é superior ao número de séries aceitáveis, é evidente que as séries de *outliers* foram corrigidas à custa de um número reduzido de séries aceitáveis, facto que pode resultar no enviesamento da PCA. Por este motivo, a PCA só admite detetores em que o número de séries aceitáveis é superior ao número de séries de *outliers*. Esta regra exclui, *a priori*, 25 detetores.

A sobreposição transversal de séries garante a quantidade mínima de registos simultâneos entre os diferentes detetores incluídos na PCA. A sobreposição é avaliada em cada detetor, por meio da relação entre o número de séries diárias e número máximo de séries diárias. A avaliação da sobreposição transversal tornou evidente a relevância de realizar vários cenários

de PCA, nomeadamente, os cenários de 60%, 70%, 80% e 90% (Purge60, Purge70, Purge80 e Purge90, respetivamente). Laranjeira (2018) aprofunda com maior rigor a origem dos cenários de PCA, expressa sobretudo pela proporcionalidade inversa entre o número de registos e o número de detetores utilizados para realização de cada cenário de PCA.

A análise gráfica das séries avalia o comportamento do tráfego ao longo do tempo de registo do detetor. Em rigor, permite verificar graficamente se a série tem falhas de registo, a respetiva extensão de tais falhas, se apresenta *outliers* e se o tráfego registado é consistente. A análise é realizada com base nos gráficos semanais e mensais de volumes de tráfego horários. Após a classificação individual da análise gráfica, determina-se a classificação gráfica agregada, que relaciona os três parâmetros da seguinte forma:

$$AG = \begin{cases} FR + 3 \times C + NO, & C\ e\ NO \neq 0 \\ 0, & C\ ou\ NO = 0 \end{cases} \quad (2)$$

Onde FR se refere à classificação de falhas de registo, C corresponde à classificação da consistência e NO significa a classificação do número individual de *outliers*. O FD classifica então o detetor por meio da seguinte formulação:

$$C_{Detector} = \frac{S \times R \times AG \times D \times 100}{TMH} \quad (3)$$

Em que S corresponde à classificação da sobreposição de séries, R corresponde à classificação da regra dos 50%, AG corresponde à classificação agregada da análise gráfica, D corresponde ao número de dias com registos e TMH corresponde ao tráfego médio diário horário. O FD dispensa 68 dos 111 detetores disponíveis, para realização das PCA. Os cenários, ou classes, de PCA utilizam os seguintes números de detetores:

Tabela 1: Quantidade de detetores utilizados nos cenários de PCA

Designação Classe	Dias com registos simultâneos	Número de Zonas	Número de Detetores
Purge60	212	9	35
Purge70	421	8	30
Purge80	1070	6	19
Purge90	2677	3	3

3.2. Resultados

Os resultados das análises são armazenados nos ficheiros de texto elencados de seguida:

Tabela 2: Resumo dos ficheiros de *output* das PCA

Output (ficheiro)	Linha	x	Coluna
01_Raw_Dataframe	*Timestamp*	x	Detetores
02_Data_in_PCA_Space	*Timestamp*	x	Componentes Principais
03_PCA_E_vectors	Componentes Principais	x	Detetores
04_PCA_E_values	Componentes Principais	x	Valor próprio
05_PCA_Loadings_Matrix	Componentes Principais	x	Detetores
06_PCA_Explained_Variance	Componentes Principais	x	Variância explicada
07_Reconstructed_Data	*Timestamp*	x	Detetores

Na Figura 6 pode observar-se o mapeamento da variância dos registos de tráfego, utilizando 2 e 3 dimensões, num referencial de componentes principais definido pelos vetores próprios mais relevantes do cenário Purge80. Para além das dimensões físicas utilizadas, pode observar-se a classificação de cada registo enquadrado num de cinco intervalos diários, possibilitando a observação da consistência dos dados para cada momento do dia e a realização das chamadas análise de *clusters*.

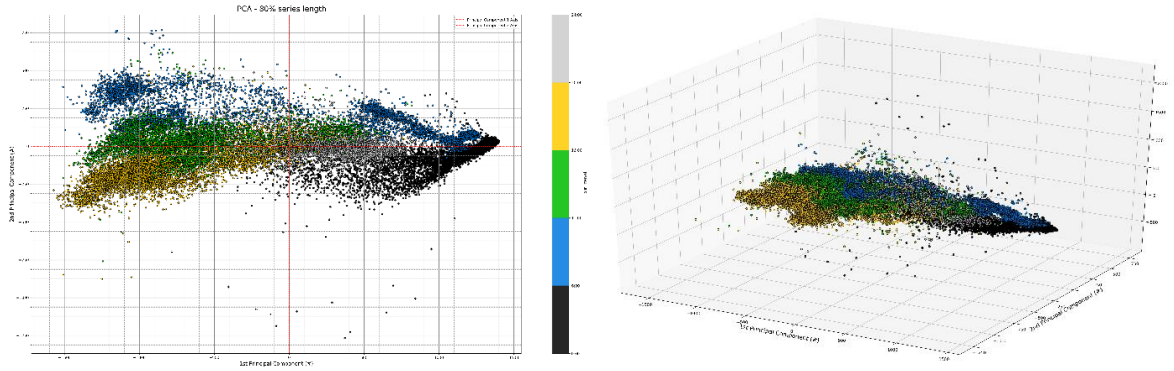

Figura 6: Visualização 2D e 3D dos resultados da Classe Purge80

Para a classe Purge60, são necessários 5 componentes principais; no caso das classes Purge70 e Purge80, são necessários apenas 4 componentes principais e, no caso da classe Purge90 é necessário apenas um único componente principal. O número de componentes principais explica, no mínimo, 95% da variância presente nos dados.

Para a classificação dos detetores tomaram-se os valores absolutos das variâncias e elencaram-se 579 dias de eventos que incluem por exemplo jogos no Estádio do Dragão, a semana da Queima das Fitas, as corridas de S. Silvestre e a *Red Bull Air Race*. Efetuaram-se dois conjuntos de PCA, um que inclui dias de eventos e outro sem os dias de eventos, tornando possível apreciar o impacto dos dias atípicos nos resultados da PCA.

Lembrando que a designação dos detetores tem a nomenclatura *zona-número*, e que a caracterização do território é proeminente, os resultados da ordenação da classe Purge60 são os que oferecem mais informação acerca do tráfego na cidade. Neste cenário, que exclui os dias de eventos, os 10 detetores principais que melhor caracterizam o tráfego são o 09-03, 08-06, 10-04, 03-09, 09-01, 02-17, 07-08, 03-07, 08-07 e o 09-09, por ordem decrescente de importância. A tabela seguinte lista os 4 primeiros detetores, de acordo com cada cenário de PCA, com e sem eventos.

Tabela 3: Seriação dos detetores principais, com e sem dias de eventos

Ordenação dos Detetores	Purge60		Purge70		Purge80		Purge90	
	Eventos		Eventos		Eventos		Eventos	
	Com	Sem	Com	Sem	Com	Sem	Com	Sem
1	09-03	09-03	02-12	02-12	03-09	03-18	03-03	03-03
2	08-06	08-06	02-06	02-06	03-18	03-09	03-18	03-18
3	09-01	10-04	05-03	03-18	02-13	02-13	02-12	02-12
4	10-04	03-09	03-18	05-03	03-06	03-06	-	-

Verifica-se que os resultados obtidos nos cenários com e sem eventos são semelhantes, havendo ligeiras reordenações.

3.3. Visualização de resultados

A visualização dos dados permite observar o comportamento do tráfego de forma estática e dinâmica. Distribuíram-se geograficamente os detetores do ITS do Porto aos quais foram associadas sequências semanais médias de tráfego, determinadas com recurso ao escalonamento de *timestamps*. O mapa serviu também para a construção dos vídeos de cada uma das PCA, que retrata a variação de tráfego simultânea em cada um dos detetores. Esta simulação dinâmica é a ferramenta que melhor resume e facilita a compreensão do fenómeno do tráfego na cidade, pois evidencia os pontos de pressão rodoviária na cidade tal como o comportamento relativo dos condutores nos arcos da rede.

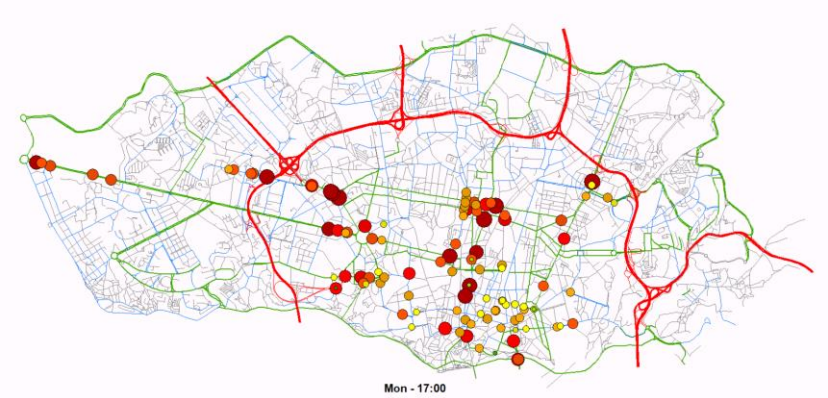

Figura 7: Captura de tela, filme com 111 detetores

4. CONCLUSÕES

Tendo-se verificado a existência de séries diárias com lapsos, utilizaram-se as técnicas DTW e de distância Euclidiana para identificar as séries aceitáveis mais semelhantes às séries de *outliers*. Para a deteção de *outliers* estabeleceram-se limiares máximos e mínimos de volume de tráfego em cada detetor, conjugando as variáveis macroscópicas de tráfego com o número de pistas monitorizadas pelo detetor. Para além dos valores omissos e dos valores anormalmente elevados, as técnicas algorítmicas de tratamento de dados utilizadas, eliminaram erros de formatação, construíram *timestamps*, rentabilizaram tempos de computação e maximizaram a otimização e aproveitamento dos dados gerados pelo ITS.

Conclui-se que a DTW é um método que produz bons resultados, mas que a sua complexidade impede a sua utilização nos casos em que a rapidez necessária à obtenção de resultados é um fator essencial. Para alcançar resultados igualmente fiáveis, mas com tempos

inferiores de computação, a distância Euclidiana é mais apropriada e, no caso particular do presente estudo, 14 vezes mais rápido do que a DTW.

O FD conduziu à criação de 4 cenários distintos para realização da PCA e respetiva classificação dos detetores. As PCA realizadas assentam na sobreposição de registos temporais tanto para um número considerável de detetores como para um período extenso, suportando assim resultados estatisticamente fiáveis. As técnicas de visualização evidenciam os bons resultados obtidos e facilitam a seleção de detetores, em função da finalidade de aplicação, da coerência numérica dos seus registos e das suas características geográficas e resumem facilmente a qualidade, densidade e padrões existentes nos conteúdos criados.

As técnicas DTW e de distância Euclidiana permitem a criação de modelos de previsão de tráfego, em tempo real, por meio do cálculo da semelhança entre séries temporais. Tendo em conta que a complexidade da DTW é $O(n^2)$ e que da métrica Euclidiana é $O(n)$, conclui-se que a última produz resultados mais rapidamente e com precisão adequada à modelação preditiva.

Os dados registados pelo ITS e verificados pela presente metodologia podem ser utilizados na construção e atualização de matrizes origem-destino. Os dados tratados proporcionam também a possibilidade de análise contínua dos volumes de tráfego, informação essa dificilmente obtida por métodos tradicionais.

Sempre que uma organização possuir um ITS, é pertinente aplicar o modelo proposto por Foerster (2008) para otimizar a disposição dos equipamentos de recolha de dados ou localizar novos equipamentos a integrar no sistema.

Por último, a informação tratada pode ser utilizada para caracterizar as relações entre o uso dos solos e os padrões de tráfego, em diferentes escalas temporais e geográficas. A análise destas interdependências constitui a matéria essencial da disciplina de Transportes, dado que o conhecimento da relação entre o uso funcional dos solos e os padrões de tráfego está na origem da criação e implementação de diferentes políticas de Transporte.

Agradecimentos

À Câmara Municipal do Porto pela disponibilização dos dados de tráfego. Toda a informação tratada e produzida, designadamente os gráficos e vídeos, pode ser disponibilizada a pedido, com a autorização da Câmara Municipal do Porto sempre que aplicável.

REFERÊNCIAS BIBLIOGRÁFICAS

Cassissi, C. e e Montalto, P. e Cannata, A. e Aliotta, M.A. e Pulvirenti, A. (2012) Similarity measures and dimensionality reduction techniques for time series data mining. *Creative Commons Attribution License*.

CCDR-N (2008) *Engenharia de Tráfego: Níveis de Serviço em Estradas e Auto-Estradas.* Edição da Comissão de Coordenação e Desenvolvimento Regional do Norte, Portugal.

Djukic, T. e Van Lint, J.W.C. e Hoogendoorn, S.P. (2012) Application of principal component analysis to predict dynamic origin-destination matrices. *Journal of the Transportation Research Board, n. 2283.* Transportation Research Board of the National Academies, Washington, D.C., EUA, p. 81–89.

Foerster, G. (2008) *Traffic state estimation using hierarchical clustering and principal component analysis: a practical approach.* Fraunhofer Institute for Transportation and Infrastructure Systems (IVI), Alemanha.

Keogh, E. (2012) *LB_Keogh homepage*. Publicação na world wide web.

Keogh, E. e Ratanamahatana, C.A. (2002) Exact Indexing of Dynamic Time Warping. *28th Very Large Databases Conference (VLDB)*. Hong Kong, China, p. 406-417.

Laranjeira, P. (2018) *Análise de Componentes principais dos detetores de um sistema inteligente de transportes*. Dissertação de Mestrado. Instituto Superior Técnico, Lisboa, Portugal.

Lopes, J. (2012) *Traffic Prediction for Unplanned Events on Motorways*. Tese de Doutoramento, Instituto Superior Técnico, Lisboa, Portugal.
Minnaar, A. (2014) *Time series classification and clustering with Python*. Publicação na world wide web.
Powell, V. (2015) *Principal component analysis explained visually*. Publicação na world wide web.
Ratanamahatana, C. A. e Keogh, E. (2004) Making Time-series Classification More Accurate Using Learned Constraints. *SIAM International Conference on Data Mining* (SDM '04), Florida, EUA, p. 11-22.
Sakoe, H. e Chiba, S. (1978) Dynamic Programming Algorithm Optimization for Spoken Word Recognition. *IEEE Transactions on Acoustics, Speech, and Signal Processing, n. 26*, p. 43-49.
Stack Overflow (2016) *Principal Component Analysis (PCA) in Python*. Publicação na world wide web.
Wikipedia (2018a) *Factor analysis*. Publicação na world wide web.
Wikipedia (2018b) *Pearson correlation coefficient*. Publicação na world wide web.
Wikipedia (2018c) *Euclidean distance*. Publicação na world wide web.
Wikipedia (2018d) *Eigenvalues and eigenvectors*. Publicação na world wide web.

Pedro C.E. Laranjeira (pedro.laranjeira@tecnico.ulisboa.pt)
Departamento de Engenharia Civil, Instituto Superior Técnico, Universidade de Lisboa
Av. Rovisco Pais, 1049-001 Lisboa, Portugal

José Pedro M. P. Tavares (ptavares@fe.up.pt)
Departamento de Engenharia Civil, Faculdade de Engenharia, Universidade do Porto
Rua Dr. Roberto Frias, 4200-465 Porto, Portugal

João A. Abreu e Silva (joao.abreu@civil.ist.utl.pt)
Departamento de Engenharia Civil, Instituto Superior Técnico, Universidade de Lisboa
Av. Rovisco Pais, 1049-001 Lisboa, Portugal