Paulo Infante,
Centro de Investigação em Matemática e Aplicações, IIFA, Universidade de Évora;
Departamento de Matemática, ECT, Universidade de Évora
pinfante@uevora.pt

DADOS QUE PODEM SALVAR VIDAS: MODELAÇÃO E PREDIÇÃO DE ACIDENTES DE VIAÇÃO PARA UMA SEGURANÇA RODOVIÁRIA MAIS EFICAZ

A sinistralidade rodoviária é um dos grandes problemas da nossa sociedade, tendo consequências sociais relevantes, quer na vida e na saúde das vítimas e dos seus familiares, quer no impacto em outras dimensões da vida em sociedade. O projeto Modelação e Predição de Acidentes de Viação no distrito de Setúbal (MOPREVIS) surgiu com o objetivo fundamental de contribuir para a redução da sinistralidade grave neste distrito. Utilizando alguns dados e resultados obtidos no projeto, este artigo mostra como a aplicação de ferramentas de base matemática num contexto de transdisciplinaridade pode conduzir a resultados muito importantes para apoiar cientificamente a tomada de decisão, contribuindo para tornar mais eficaz a segurança rodoviária.

1. INTRODUÇÃO

As vias rodoviárias constituem uma rede dinâmica de mobilidade que incorpora uma preocupação emergente: os acidentes de trânsito (colisões, despistes ou atropelamentos). Um acidente pode causar, para além dos danos materiais, perdas humanas e/ou danos físicos e psicológicos irreparáveis a muitas das vítimas. Segundo a Organização Mundial da Saúde [1], os acidentes de viação causam aproximadamente 1.3 milhões de vítimas mortais anualmente e entre 20 a 50 milhões de feridos, sendo a principal causa de morte de crianças e adultos jovens entre os 5 e os 29 anos. Segundo a Comissão Europeia, Portugal é o oitavo país da União Europeia com mais vítimas mortais por milhão de habitantes (54, ou seja, mais 9 que a média dos 27 países) [2].

A dimensão económica é também muito relevante, tendo em conta os custos humanos, perda de produtividade, custos médicos, danos de propriedade, custos administrativos, entre outros. De acordo com um estudo divulgado pela Autoridade Nacional da Segurança Rodoviária (ANSR) [3], os acidentes de viação registados em Portugal no ano de 2019 tiveram um custo económico e social para o país estimado em 6422.9 milhões de euros, um valor que representou 3,03% do PIB nesse ano.

Neste contexto, a procura por soluções eficazes que possam prevenir acidentes de viação assume uma importância inquestionável. Aqui a Matemática, intrínseca à Ciência de Dados e à Inteligência Artificial, assume um papel fundamental que pode salvar vidas, nomeadamente através da capacidade de analisar e modelar dados, que constitui uma das maiores esperanças na luta contra

acidentes de trânsito. Descobrindo padrões ocultos e analisando informação histórica, é possível identificar fatores de risco e, mais importante, predizer acidentes antes que ocorram.

O Projeto MOPREVIS foi concebido para dar resposta a uma necessidade de uma Força de Segurança, a Guarda Nacional Republicana (GNR), no distrito de Setúbal, onde a sinistralidade rodoviária grave (definida como aquela em que resultam feridos graves e/ou mortos) era muito elevada. Este distrito, em particular, em 2018 registou o maior número de vítimas mortais entre todos os distritos do país, apesar de não ser dos distritos com maior número de acidentes. O projeto teve como objetivo fundamental contribuir para a redução da sinistralidade grave no distrito de Setúbal. Para tal foram traçados 5 objetivos específicos: 1) encontrar determinantes para a ocorrência e gravidade dos acidentes; 2) construir um sistema de informação espacial; 3) traçar o perfil dos intervenientes; 4) obter modelos preditivos para a ocorrência de acidentes por troço de estrada; 5) dotar a GNR com uma ferramenta digital de apoio à tomada de decisão em tempo real, de modo a permitir a otimização e a gestão dos recursos para a prevenção. Para atingir estes objetivos, foi formada uma equipa transdisciplinar, juntando investigadores dos departamentos de matemática, informática, geociências e sociologia da Universidade de Évora, com membros do Comando Territorial da GNR de Setúbal. Uma parte dos resultados obtidos no Projeto foi apresentada de uma forma essencialmente visual, suportada em infografias e na respetiva interpretação, privilegiando uma abordagem mais generalista e menos técnica, em [4]. Outros resultados importantes podem ser vistos em [5-9].

Neste artigo, tendo como pano de fundo os resultados obtidos através do projeto MOPREVIS, ilustra-se como a aplicação de algumas ferramentas, de base matemática, na análise e modelação de dados de acidentes de viação pode conduzir a resultados muito importantes para apoiar cientificamente a tomada de decisão, permitindo tornar mais eficaz a segurança rodoviária. Na secção 2 apresentam-se duas metodologias estatísticas utilizadas e faz-se uma pequena incursão nos modelos de aprendizagem automática. Na secção 3 são apresentados, de uma forma crítica, alguns resultados obtidos com as referidas metodologias estatísticas. Na secção 4 explica-se a construção dos modelos de predição implementados na ferramenta digital atualmente a ser utilizada pela GNR de Setúbal. Termina-se, na secção 5, com algumas considerações de natureza geral.

2. ALGUMAS METODOLOGIAS ESTATÍSTICAS UTILIZADAS

A Carta de Controlo de Qualidade

De uma forma muito simples pode definir-se uma carta de controlo como uma representação gráfica de valores de uma estatística amostral (por exemplo, média, desvio padrão, amplitude, proporção, número de não conformidades) em função do tempo. A estatística mede uma determinada característica da qualidade com base numa amostra aleatoriamente selecionada. A característica da qualidade pode assumir uma índole quantitativa, isto é, pode ser medida e expressa por um número (controlo por variáveis) ou uma índole qualitativa (controlo por atributos).

As cartas de controlo mais usuais foram desenvolvidas por Walter A. Shewhart em finais da década de 1920, constituídas por uma linha central que representa o valor médio da característica da qualidade, no caso em que o processo se encontra sob controlo estatístico, e por duas linhas simetricamente colocadas acima e abaixo da linha central, designadas por limites de controlo. Designando por W uma estatística amostral que mede uma determinada característica da qualidade com média μ_W e desvio padrão σ_W, a carta ficará definida com uma linha central (LC) igual a μ_W, um limite superior de controlo (LSC) igual a $\mu_W + L \cdot \sigma_W$ e um limite inferior de controlo (LIC) igual a $\mu_W - L \cdot \sigma_W$. O coeficiente L representa a distância dos limites de controlo à linha central medida em unidades do desvio padrão da estatística. Usualmente, toma-se $L = 3$, sendo os limites conhecidos por limites "3-sigma". Desta forma, se a distribuição da estatística amostral for aproximadamente normal, existindo apenas causas aleatórias (de variação inerente ao processo), caso em que se refere que o processo está sob controlo estatístico, em média apenas 27 valores surgem fora dos limites de controlo em cada 10000. O efeito de uma alteração no processo, como consequência do aparecimento de uma causa externa (usualmente designada por causa assinalável), traduz-se numa alteração no(s) parâmetros(s) dessa distribuição de probabilidade que modela a variabilidade aleatória do processo, com o consequente aparecimento de valores da estatística fora dos limites de controlo e/ou de padrões não aleatórios na carta de controlo. Na aplicação apresentada na próxima secção utiliza-se uma *carta c*, cujos limites de controlo e linha central são dados por:

$$LIC = \bar{c} - 3\sqrt{\bar{c}}; LC = \bar{c}; LSC = \bar{c} + 3\sqrt{\bar{c}},$$

onde $\bar{c}$ representa a média do número de ocorrências por unidade de tempo ou espaço. Também se utiliza uma *carta*

u, cujos limites de controlo e linha central são dados por:

$$LIC = \overline{u} - 3\frac{\sqrt{\overline{u}}}{\sqrt{n_i}}; LC = \overline{u}; LSC = \overline{u} + 3\frac{\sqrt{\overline{u}}}{\sqrt{n_i}},$$

onde $\overline{u}$ representa a média do número de ocorrências por unidade e n_i a dimensão da unidade i.

Quando se pensa que poderá existir variação entre as categorias de uma dada variável categórica, deve optar-se por representar os dados para as categorias na mesma carta de controlo usando *rational ordering* (RO) ou *rational subgrouping* (RS). RO significa que os dados para todas as categorias são apresentados sequencialmente usando uma ordem temporal (por exemplo, semanas ou meses) de modo a ter pontos suficientes para uma carta sólida (usualmente entre 20 e 30, podendo em alguns casos ser ligeiramente inferior). RS significa que os dados para cada uma das categorias são agregados de modo a ter uma comparação transversal das várias categorias, usando apenas um ponto para cada categoria. Quando se aplica RS, os pontos deixam de estar ligados, pois já não se observam os dados ao longo do tempo. Tal permitirá avaliar se alguma categoria é uma causa assinalável relativamente às restantes. Neste caso, a obtenção de cartas de controlo separadas para cada categoria permitirá averiguar se há ou não estabilidade do processo dentro de cada categoria, podendo conduzir a intervenções sobre o próprio processo para uma melhoria. Vários exemplos pormenorizados de aplicação desta metodologia na área da saúde podem ser vistos, por exemplo, em [10]. Note-se que, tendo em conta a evolução do processo ao longo do tempo, a abordagem RO é muito mais informativa e potente do que a avaliação realizada por uma comparação transversal entre as várias categorias.

O Modelo de Regressão Logística

Ao pretender modelar uma variável resposta binária, para a qual o resultado da resposta de cada indivíduo é um "sucesso" ou um "insucesso", o modelo de regressão logística é o mais aplicado. Este é um modelo linear generalizado com componente aleatória binomial [11]. Admita-se que se tem um conjunto de possíveis variáveis explicativas independentes, que podem ser representadas por um vector $\mathbf{x}^\intercal = (x_1, x_2, \ldots, x_p)$. Representando a probabilidade condicional do resultado ser um sucesso por $Pr(Y = 1|\mathbf{x}) = \pi(\mathbf{x})$, o modelo de regressão logística é definido por

$$\pi(\mathbf{x}) = \frac{e^{g(\mathbf{x})}}{1 + e^{g(\mathbf{x})}}.$$

A função g lineariza a resposta sendo designada por função *logit*

$$g(\mathbf{x}) = \ln\left(\frac{\pi(\mathbf{x})}{1 - \pi(\mathbf{x})}\right) = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \ldots + \beta_p x_p,$$

mostrando que $\pi(\mathbf{x})$ aumenta ou diminui conforme a função logística e permitindo interpretar mais facilmente o efeito de cada variável. Se algumas das variáveis independentes forem categóricas (como o sexo ou o tipo de via), utiliza-se uma coleção de variáveis *dummy*, sendo a estratégia mais usual a de criar k-1 destas variáveis quando existem k categorias. Assim, representando por δ_{ij}, $j = 1, 2, \ldots, k$-1 as variáveis *dummy* e por β_{ij} os seus coeficientes, então o *logit* do modelo para p variáveis, onde a i-ésima é categórica com k categorias, pode ser escrito como

$$g(\mathbf{x}) = \beta_0 + \beta_1 x_1 + \ldots + \sum_{j=1}^{k-1} \beta_{ij}\delta_{ij} + \ldots + \beta_p x_p.$$

A grande popularidade dos modelos de regressão logística está relacionada com a facilidade de interpretação dos coeficientes através do *Odds Ratio* (OR). Por exemplo, numa variável dicotómica em que $x = 1$ traduz a presença de uma dada caraterística e $x = 0$ a sua ausência, o OR consiste na razão entre a chance/possibilidade (*odds*) de ocorrer "sucesso" entre os indivíduos com $x = 1$, dado por $\pi(1)/[1 - \pi(1)]$ e o (*odds*) de ocorrer "sucesso" entre os indivíduos com $x = 0$, dada por $\pi(0)/[1 - \pi(0)]$. Esta razão de chances neste exemplo é dada pela exponencial do coeficiente de x no modelo de regressão logística e traduz o aumento (se o coeficiente é positivo) ou diminuição (se o coeficiente é negativo) das chances/possibilidades de ocorrer "sucesso" nos indivíduos com a característica presente relativamente aos indivíduos com a característica ausente. O "sucesso" é visto como a ocorrência de um qualquer evento (seja ele positivo ou negativo) que interessa modelar, podendo também ser o estar acima ou abaixo de um ponto de corte de uma variável contínua ou ordinal, tornando o modelo aplicável nas mais diversas áreas do conhecimento.

As estimativas dos parâmetros são obtidas por máxima verosimilhança. Desenvolvimentos deste modelo, estratégias de modelação, inferência sobre os parâmetros, bondade de ajustamento, validação de pressupostos e análise de resíduos podem ser consultados, por exemplo, em [12].

Modelos de Aprendizagem Automática

Em termos gerais, os modelos de aprendizagem automática (*machine learning* - ML) são sistemas computacionais

que a partir de algoritmos assimilam conhecimento com base em dados e utilizam esse conhecimento para realizar predições, apoiar a tomada de decisões informadas e executar tarefas complexas. É evidente uma forte base matemática subjacente a esses algoritmos, abrangendo, por exemplo, conceitos de álgebra linear, análise matemática e probabilidade e estatística. São tais ferramentas matemáticas que viabilizam a compreensão de como os algoritmos aprendem a partir dos dados, otimizam os parâmetros e se generalizam para novas situações. As redes neuronais (*Neural Networks* - NN), as máquinas de vetores de suporte (*Support Vector Machines* - SVM) e as árvores de decisão (*Decision Trees* - DT) são exemplos de algoritmos ML que têm uma base sólida em matemática. Além disso, a estatística desempenha um papel fundamental na seleção, avaliação e validação do desempenho dos modelos, bem como na obtenção de informação útil e fiável no apoio à tomada de decisão.

A escolha entre modelos estatísticos clássicos e ML deve ser pautada pela natureza do problema, pela quantidade e qualidade dos dados disponíveis, pela necessidade de interpretação específica dos coeficientes dos modelos e pela complexidade das relações subjacentes. Os modelos ML oferecem vantagens, por exemplo, em situações em que as relações entre variáveis são muito complexas, em que se pretende explorar grandes volumes de dados e/ou lidar com grande variedade de tipos de dados, incorporar informações não estruturadas (textos, sons, imagens), ou numa adaptação automática a alterações nos dados.

Quando se pensa num modelo preditivo, a sua interpretabilidade assume uma menor importância, pois o essencial é que o modelo possa ser validado adequadamente e realize predições precisas. A base de um modelo preditivo eficaz é estabelecida com bom senso e um profundo conhecimento do contexto do problema, considerando dados fiáveis e relevantes, realizando validação e visualização desses dados e considerando um amplo conjunto de ferramentas de modelação para lidar com os diferentes cenários possíveis.

As técnicas de modelação utilizadas com informação estruturada podem ser classificadas como aprendizagem supervisionada (quando existe uma variável objetivo observada nos dados) ou não supervisionada (quando o objetivo é encontrar padrões ou relações entre eles). Na aprendizagem supervisionada destacam-se as NN, SVM, DT e os classificadores bayesianos; na não supervisionada, destacam-se as técnicas de *clustering* e as de redução de dimensionalidade, além das NN que também existem nesta classe [13, 14].

As DT podem ser aplicadas a problemas de regressão ou de classificação. Opta-se por neste artigo focar brevemente as árvores de classificação, pois, além do algoritmo de regressão logística de aprendizagem automática, foram os algoritmos baseados em árvores de classificação, (*Random Forest* (RF), C5.0 e XGBoost), que se mostraram mais eficientes na predição dos acidentes de viação. Uma árvore de decisão é uma estrutura hierárquica que representa um processo de tomada de decisão. No caso das árvores de classificação, a resposta é qualitativa. A árvore é composta por nós de decisão, ramos e folhas. Cada nó representa uma questão (ou teste) sobre um atributo específico, cada ramo indica um resultado (ou caminho possível), e cada folha representa uma categoria da variável resposta. Cada nó pode conduzir a novos nós ou a uma folha. O processo de construção de uma árvore envolve dividir repetidamente o conjunto de dados em subconjuntos com base nos atributos, de modo a maximizar a pureza do nó em cada ramificação. Usualmente é utilizado o Índice Gini ou a entropia para avaliação dessa pureza. O RF é um algoritmo, incluído nas designadas técnicas de *ensemble learning*, que envolvem combinar vários modelos para melhorar o desempenho, a precisão e a capacidade de generalização. Utiliza a abordagem de *Bagging Bootstrap Aggregating* para criar um conjunto de árvores de decisão individuais, sendo as predições de cada árvore combinadas ([15]). O algoritmo C5.0 é uma extensão avançada das árvores de decisão tradicionais, que melhora o desempenho do modelo [16]. Finalmente, o XGBoost (*Extreme Gradient Boosting*) combina várias árvores de decisão, melhorando a precisão das predições por um processo iterativo de construção de árvores [17].

3. ALGUNS RESULTADOS DE ÍNDOLE EXPLICATIVA

Nesta secção apresentam-se alguns resultados obtidos que resultaram da aplicação de cartas de controlo e do modelo de regressão logística descritos na secção anterior. Refira-se que o projeto MOPREVIS permitiu, apesar de várias condicionantes, nomeadamente ao nível da disponibilidade e qualidade dos dados sobre a sinistralidade rodoviária, obter vários resultados importantes, mostrando que a extensão deste projeto para outros distritos pode ser um importante contributo em termos de prevenção e segurança rodoviária. Nos resultados de índole mais explicativa podem destacar-se:

1. Definição dos principais determinantes para a ocor-

rência de acidentes, para a ocorrência de acidentes com feridos graves ou mortos (sinistralidade grave), para a natureza dos acidentes (atropelamento, colisão ou despiste), para a ocorrência de acidentes envolvendo motociclos e para a ocorrência de atropelamentos com vítimas;

2. Conceção de um sistema de informação espacial sobre os acidentes, não só determinando de uma forma consistente *hotspots* e *clusters* de acidentes, mas também concebendo atlas de suscetibilidade de ocorrência de acidentes;

3. Implementação de um novo indicador de gravidade, mais robusto e consistente que os existentes, que considera: a gravidade do acidente; um efeito amortecedor do número de vítimas; um ponderador espacial; e um ponderador temporal;

4. Determinação do perfil dos intervenientes nos acidentes de viação (condutores e vítimas).

A Área e os Dados do Estudo

O distrito de Setúbal está situado a Sul de Lisboa e abrange uma área de 5.064 km^2, divididos em 13 municípios que compreendem áreas urbanas e rurais. O distrito é acessível por importantes vias de acesso a Lisboa e possui diversos pontos turísticos que aumentam o fluxo de trânsito nos períodos de pico.

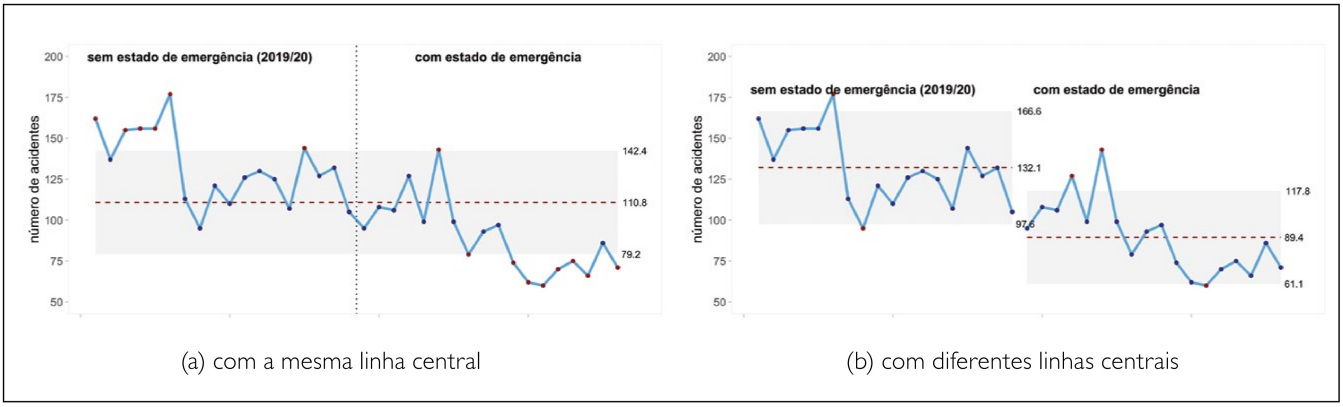
A principal fonte de dados foi o Boletim Estatístico do Acidente de Viação (BEAV), preenchido pelas autoridades policiais para cada acidente de viação, no período entre 2016 e 2022 (com variações do horizonte temporal consoante as abordagens e o momento em que ocorreram) atualizados pela ANSR para as vítimas a 30 dias (entre 2016 e 2019, pois a partir daí não foi recebida nenhuma atualização). Estes dados foram delimitados geograficamente aos acidentes registados pelo Comando Territorial da GNR de Setúbal ocorridos naquele distrito (a GNR tem jurisdição territorial em cerca de 96% do distrito). Os dados obtidos dos registos oficiais foram validados por cruzamento de várias variáveis e as coordenadas dos acidentes com vítimas foram validadas pela mesma entidade. Foram, ainda, incluídas diversas variáveis meteorológicas fornecidas pelo Instituto Português do Mar e da Atmosfera (IPMA) e outras relacionadas com as vias de circulação, informação fornecida pela Infraestruturas de Portugal (IP).

Note-se que o BEAV é apenas o retrato, mas não o filme, do momento do acidente, uma vez que não tem informações recolhidas *a posteriori* como, por exemplo, as vítimas a 30 dias ou os testes de álcool não realizados no momento. Sendo um registo manual, tem naturalmente imprecisões em alguns campos e em algumas coordenadas geográficas, bem como dados em falta, entre outros. O processo de validação foi muito moroso e a equipa de investigação sensibilizou as principais instituições envolvidas na problemática da sinistralidade rodoviária para a necessidade de melhorar a qualidade dos dados e proceder à sua validação. Também não existe informação oficial de variáveis importantes, como a intensidade de tráfego, a velocidade dos veículos, idade do parque automóvel e diversas características dos condutores.

A base final de análise foi composta por mais de 1000 variáveis de diferentes tipos: espaciais, temporais, ambientais, veículos envolvidos, intervenientes, via, tipologia e consequências do acidente. No período entre 2016 e 2019 e entre maio de 2021 e junho de 2022 (em que foram realizadas grande parte das análises), foram registados 36701 acidentes, 207 dos quais originaram vítimas mortais, 555 originaram feridos graves e 7199 originaram feridos leves. No total registaram-se 233 vítimas mortais, 686 feridos graves e 10040 feridos leves.

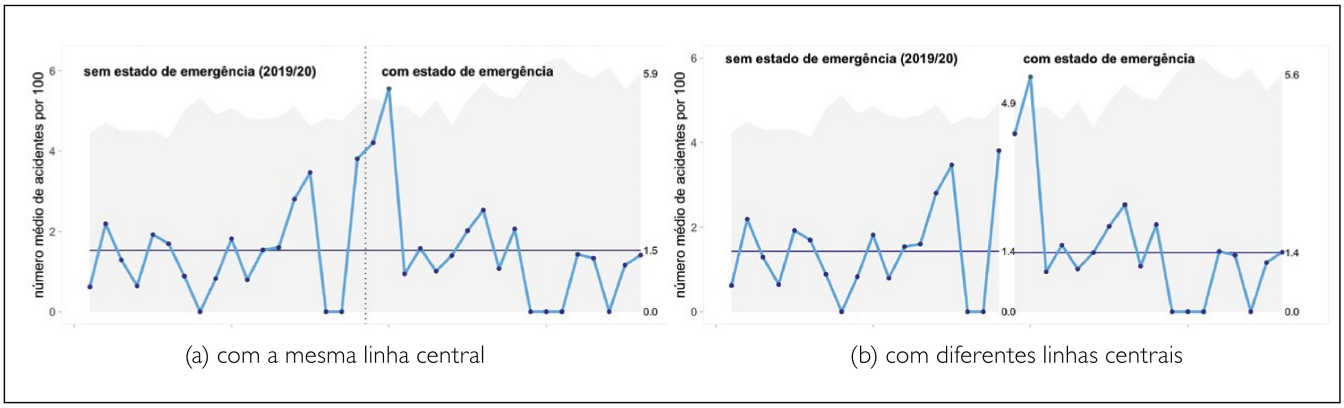
A intensidade de tráfego é explicativa para o número de acidentes, mas não para a sua gravidade?

Pretendendo comparar o número de acidentes ocorridos no segundo período do estado de emergência (09/11/2020 a 16/03/2021) com os ocorridos no período homólogo dos anos anteriores, em vez de uma comparação transversal (que não tem em consideração a evolução com o tempo), podem usar-se cartas de controlo. Neste caso, definindo a semana como unidade de tempo, o número de acidentes nesta unidade de tempo pode ser monitorizado por uma *carta c*. Colocando os dois períodos (sem estado de emergência e com estado de emergência) na mesma carta de controlo, pode considerar-se que se usa uma abordagem *rational ordering*, por serem categorias de uma variável global (período). Contudo, neste caso, as observações, no seu conjunto, constituam uma sequência temporal. Na Figura 1(a) pode observar-se que os dois períodos têm distribuições de acidentes claramente diferentes, estando vários pontos acima do LSC e quase todos os pontos acima da LC no período sem estado de emergência, e vários pontos abaixo do LIC e abaixo da LC no período do estado de emergência. As cartas separadas por período (Figura 1(b)), embora ambas com alguns pontos fora dos



(a) com a mesma linha central (b) com diferentes linhas centrais

Figura 1. *Cartas c* para o número de acidentes por semana.



(a) com a mesma linha central (b) com diferentes linhas centrais

Figura 2. *Cartas u* para o número médio de acidentes de sinistralidade grave, em cada 100, por semana.

limites, permitem concluir que o número médio de acidentes foi menor no período em que vigorou o estado de emergência, onde, como se sabe, houve uma muito menor intensidade de tráfego.

Pensando, agora, no número de acidentes com sinistralidade grave (envolvendo feridos graves ou vítimas mortais), tomando como unidade de contagem o total de acidentes ocorridos numa semana, e atendendo a que este valor é variável, pode aplicar-se neste caso uma *carta u* e adotar a mesma abordagem que foi usada na *carta c*. Na figura 2(a) pode observar-se que os dois períodos têm distribuições de acidentes idênticas. Surge apenas um ponto acima do LSC no período do estado de emergência, existindo 8 pontos consecutivos abaixo da linha central que podem ser um indício de diminuição do valor da estatística, embora a probabilidade associada seja um pouco superior (0.0039) à de se obter um valor fora dos limites. Contudo, observando a Figura 2(b), onde se estimou a estatística eliminando o valor acima do LSC, conclui-se que não há base estatística para afirmar que se tenha registado uma redução no número médio de acidentes com sinistralidade grave no período em que decorreu o estado de emergência.

Quartas-feiras com menos acidentes com vítimas mortais?

Para estudar o efeito do dia da semana na ocorrência de acidentes com vítimas mortais, foi obtida uma *carta u*, uma vez que o número de acidentes por unidade de tempo é variável. Neste caso, para evitar um efeito excessivo da assimetria da distribuição (a probabilidade de ocorrer um acidente deste género é muito pequena) selecionou-se o trimestre como período temporal. Aplicando uma abordagem de *rational ordering* com os dados ordenados temporalmente dentro de cada dia da semana, representam-se todos os dias da semana na carta com a mesma linha central. Da sua análise, a quarta-feira destacou-se como causa assinalável relativamente aos restantes dias. Ao representar a carta com duas classes (quarta-feira *vs* restantes dias) com a mesma linha central, pode observar-se um padrão não aleatório nos acidentes com vítimas mortais ocorridos às quartas-feiras, surgindo 17 dos 20 pontos abaixo da

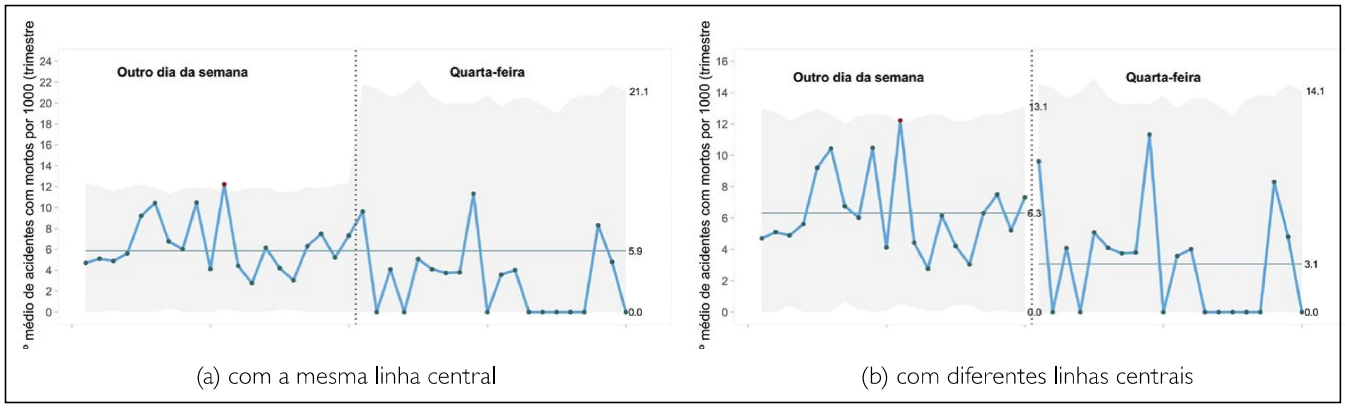


Figura 3. Cartas *u* para o número médio de acidentes com vítimas mortais, em cada 100, por trimestre

linha central (Figura 3(a)), a que está associada uma probabilidade de 1 em 1000. Construindo cartas separadas (Figura 3(b)), pode concluir-se que o número médio de acidentes com vítimas mortais às quartas-feitas é inferior ao ocorrido nos restantes dias da semana, estimando-se em aproximadamente metade (3 em cada 1000 *vs* 6 em cada 1000).

Determinantes para a Ocorrência de Acidentes Envolvendo Motociclos

O modelo de regressão logística permitiu encontrar, por exemplo, determinantes de acidentes envolvendo vítimas e também, de entre estes, determinantes para a ocorrência de feridos graves ou vítimas mortais ([5] e [6]). Também permitiu encontrar fatores que estão na origem de acidentes por atropelamento, que envolvem vítimas, constituindo um passo importante para delinear medidas para a sua prevenção. Uma caracterização de acidentes desta natureza, com base na informação disponível no BEAV, foi incluída em [7].

Neste artigo apresentam-se os resultados obtidos com o ajustamento de um modelo de regressão logística para os acidentes envolvendo motociclos, ciclomotores, triciclos e quadriciclos, aqui designados por motociclos por simplicidade de escrita. Estes acidentes merecem uma atenção particular, pois no período em análise não só aumentou a sua ocorrência como estão tipicamente associados a uma maior gravidade. Este modelo (Tabela 1) ajustou-se bem aos dados e registou uma ótima capacidade discriminativa (*Area Under the ROC Curve* $= 0.84$).

A interpretação dos seus coeficientes permite avaliar o efeito de cada fator sobre a probabilidade de ocorrência de acidentes envolvendo motociclos. Com base no modelo e nos seus coeficientes pode concluir-se que os principais fatores que contribuem para uma maior probabilidade de ocorrência de um acidente envolvendo motociclos são:

- ESPACIAIS - A chance/possibilidade aumenta:
  - Aproximadamente o triplo se ocorrer nos concelhos de Almada, Sesimbra e Setúbal, um pouco menos do triplo se ocorrer no Barreiro, Palmela ou Seixal ou próximo dos 60% se ocorrer em Alcochete ou na Moita, relativamente aos acidentes que ocorrem nos concelhos de Alcácer do Sal, Grândola, Montijo, Santiago do Cacém e Sines;
  - Para quase o dobro se ocorrer a menos de 100 m de uma escola;
  - Para cerca do dobro se ocorrer fora de uma zona de estacionamento;

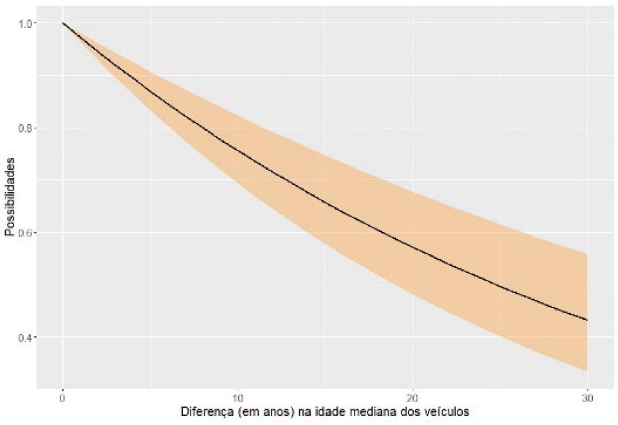
- TEMPORAIS - A chance/possibilidade aumenta:
  - Cerca de 20% se ocorrer ao domingo;
  - Nos horários compreendidos entre as 7h e as 9h e entre as 19h e as 21h;

- RELACIONADOS COM O CONDUTOR - A chance/possibilidade aumenta mais de 6 vezes em acidentes onde a maioria dos condutores envolvidos é do sexo masculino;

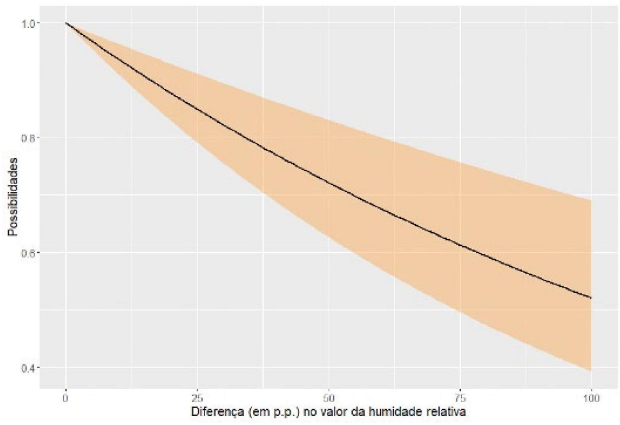
- RELACIONADOS COM A VIA - A chance/possibilidade aumenta:
  - Cerca de 70% em acidentes que ocorram numa via com separador central;
  - Cerca de 5 vezes se ocorrer numa Estrada Nacional

Tabela 1. Modelo de regressão Logística para a ocorrência de um acidente envolvendo motociclos. Variáveis e respetivas categorias, coeficientes do modelo, desvios padrão dos coeficientes e valores *p*.

Variável	Coef.	D. Padrão	Valor *p*
Constante	–12.28	0.54	<0.001
Concelho (ref: Restantes Concelhos)			
Seixal / Barreiro / Palmela	0.97	0.09	<0.001
Setúbal / Sesimbra / Almada	1.19	0.09	<0.001
Moita / Alcochete	0.46	0.13	<0.001
Zona de Estacionamento (ref: Sim)			
Não	0.74	0.26	0.004
Proximidade de Escolas a 100m (ref: Sim)			
Não	–0.63	0.24	0.008
Tipo de via (ref: Autoestrada)			
EN	1.61	0.20	<0.001
Ponte / IC / IP	0.73	0.18	<0.001
Outra Via	1.27	0.20	<0.001
Existência de Separador Central (ref: Não)			
Sim	0.53	0.14	<0.001
Humidade Relativa	–0.007	0.002	<0.001
Idade Mediana dos Veículos Envolvidos	–0.03	0.01	<0.001
Percentagem de Condutores do Sexo Masculino (ref: [0,50]			
[50, 100]	1.82	0.13	<0.001
Hora do Acidente (ref: Restantes)			
6h-7h / 22h-0h	–0.31	0.13	0.015
1h-5h	–1.34	0.27	<0.001
7h-8h / 19h-21h	0.31	0.07	<0.001
Veículo Ligeiro (ref:Sim)			
Não	6.88	0.25	<0.001
Veículo Pesado (ref: Sim)			
Não	6.56	0.30	<0.001
Condições Meteorológicas (ref: Bom Tempo)			
Chuva / Outros Elementos Atmosféricos	–0.74	0.12	<0.001
Dia da Semana (ref: segunda a sábado)			
domingo	0.17	0.08	0.047



(a) função da idade mediana dos veículos envolvidos



(b) função das diferenças na humidade relativa

Figura 4. *Odds Ratio* (OR) de ocorrência de um acidente envolvendo motociclos.

(EN), o dobro se ocorrer na Ponte Vasco da Gama ou em Itinerário Complementar ou Principal (IC/IP) e cerca de 3.5 vezes se ocorrer noutras vias, relativamente aos acidentes deste tipo que ocorrem numa autoestrada;

▸ RELACIONADOS COM O VEÍCULO - A chance/possibilidade aumenta:

- Nos veículos mais recentes (aumenta com a diminuição da idade dos veículos envolvidos, Figura 4(a));
- No caso do acidente não envolver veículos ligeiros;
- No caso do acidente não envolver veículos pesados;

▸ METEOROLÓGICOS - A chance/possibilidade aumenta:

- Para aproximadamente o dobro se estiver bom tempo;
- Quando a humidade relativa diminui (Figura 4(b)).

4. PREDIÇÃO DE ACIDENTES DE VIAÇÃO

Com base na sinistralidade grave ocorrida, foram selecionadas 4 vias: Estrada Nacional 4 (EN4), Estrada Nacional 10 (EN10), Itinerário Complementar 1 (IC1) e Autoestrada 33 com a Ponte Vasco da Gama (A33+PVG). A metodologia, descrita integralmente em [9] para a EN10 (predição de ocorrência de acidentes), utiliza uma abordagem mista entre inteligência artificial, estatística e sistemas de informação geográfica (SIG), consistindo em cinco etapas fundamentais: (1) divisão da via em segmentos (a dimensão de 500 m foi a que conduziu a melhores resultados); (2) geração de amostras negativas para obter informações sobre os períodos e segmentos em que não ocorreram acidentes; (3) ajuste de modelos de ML com 3 abordagens diferentes para mitigar o forte desbalanceamento dos dados (a abordagem ROSE - *Random Over-Sampling Examples* foi a que produziu melhores resultados); (4) comparação do desempenho dos modelos para cada via; (5) implementação, numa aplicação digital, do modelo com melhor desempenho para cada via.

Após o processo de segmentação, baseado em técnicas de SIG, os segmentos da via possuem apenas as amostras positivas, ou seja, as informações referentes às ocorrências dos acidentes de trânsito. Portanto, não há dados sobre quando não ocorreram acidentes (que usualmente se designam por amostras negativas). Para criar tais amostras, são geradas todas as datas (ano, mês, dia e hora) no período temporal pretendido, que neste caso foi de 1 de janeiro de 2016 a 31 de dezembro de 2022, retirando o período compreendido entre março de 2020 e abril de 2021, que correspondeu ao período de maior influência da pandemia COVID-19, por serem obtidos numa situação atípica. A estas datas (excluindo os casos em que houve acidentes) foram adicionadas informações sobre outras variáveis que podem ser consideradas preditivas, isto é, que sejam conhecidas antes da ocorrência do acidente de viação: variáveis meteorológicas (precipitação, direção e velocidade do vento e temperatura), variáveis temporais (dia da semana,

movimento sazonal, pico de tráfego, turnos diurnos - ida para o trabalho, manhã ou tarde, saída do trabalho -, período escolar, período do nascer ou do pôr do sol e período de férias), variáveis relacionadas com as características da estrada (segmento, traçado da estrada, tipo de cruzamento rodoviário, número de vias, qualidade do pavimento, tipo de berma, existência de túnel/ponte no segmento e número de árvores no segmento), variáveis relacionadas com sinalização e informação sobre a velocidade no segmento (limite de velocidade e valores históricos da velocidade média).

Quando existe uma grande diferença no número de casos em cada classe da variável resposta, costuma referir-se que se está na presença de dados desbalanceados. Nesta situação, num problema de classificação binária, a classe que corresponde à não ocorrência de acidentes tem muito mais casos do que a classe de ocorrência de acidentes (basta pensar no número de horas em cada segmento de via em que ocorreram e não ocorreram acidentes). Tal pode levar a diversos problemas nos modelos, tais como o enviesamento da classe maioritária (uma vez que tem mais ocorrências dessa classe), um desempenho irrealista (modelo pode ter uma precisão elevada predizendo bem apenas a classe maioritária) ou uma menor sensibilidade à classe minoritária (maior dificuldade em aprender padrões dessa classe), conduzindo a decisões erradas. Para lidar com este problema, existem várias estratégias que podem ser aplicadas: 1) realizar sobreamostragem (*oversampling*) da classe minoritária ou subamostragem (*undersampling*) da classe maioritária para equilibrar as classes; 2) atribuir pesos diferentes às classes durante o treino para aumentar a importância da classe minoritária; 3) criar exemplos sintéticos da classe minoritária para aumentar a sua representação. Neste caso concreto, a abordagem que deu melhores resultados foi a ROSE [18], que consistiu em obter uma amostra aleatória dos casos negativos (classe maioritária), sendo quatro vezes o número de casos positivos (classe minoritária), e em seguida sobreamostrar os casos negativos replicando a classe minoritária. Desta forma, obtêm-se acidentes de viação sintéticos a partir dos existentes. Consequentemente, após a amostragem das amostras negativas para obter uma relação de 4:1, a sobreamostragem permite obter uma relação de aproximadamente 1:1 entre os casos negativos e positivos.

A aplicação dos modelos ML faz-se dividindo o conjunto de dados em dois subconjuntos. Um primeiro, designado por conjunto de dados de treino (neste caso entre 2016 e 2021, excluindo o período de influência da pandemia), usado para treinar o modelo, estimando os seus parâmetros para se ajustar aos padrões nos dados. Um segundo, designado por conjunto de dados de teste (neste caso o ano de 2022), usado para avaliar o desempenho do modelo com novos dados, sendo fundamental para medir a capacidade preditiva do modelo. No caso da EN10, o modelo preditivo atualmente implementado na aplicação digital é para acidentes com vítimas (portanto menos ocorrências), pelo que foi considerado um maior período de dados de teste (maio de 2021 a dezembro de 2022). Existem diferentes medidas de desempenho dos modelos, devendo ter-se um cuidado particular com as medidas a utilizar no caso de dados desbalanceados [19].

Os modelos atualmente implementados em cada uma das 4 vias e algumas medidas de desempenho são apresentados na Tabela 2. A Sensibilidade (SEN) é a probabilidade do modelo predizer corretamente um acidente de trânsito, a especificidade (ESP) é a probabilidade do modelo predizer corretamente um não acidente. O valor preditivo positivo (VPP) é a proporção de acidentes corretamente preditos e o valor preditivo negativo (VPN) é a proporção de não acidentes corretamente preditos. Estes valores variam consoante o ponto de corte (valor que separa as predições do modelo entre acidentes e não aci-

Tabela 2: Modelos ajustados (via de implementação) e valor de algumas medidas de desempenho.

Medida	C5.0 (EN10)	XGBoost (EN4)	LR (A33+PVG)	XGBoost (IC1)
SEN	0.888	0.810	0.828	0.739
ESP	0.611	0.607	0.502	0.545
VPP	0.621	0.618	0.617	0.616
VPN	0.884	0.803	0.750	0.679
AUC	0.881	0.755	0.729	0.707

dentes), o qual neste caso foi obtido para maximizar a sensibilidade, mantendo a especificidade próximo dos 50%. É ainda apresentada a área abaixo da curva ROC (AUC).

Como se pode observar, os resultados obtidos são bastante animadores. Recorde-se que os modelos predizem a ocorrência de um acidente rodoviário num dado segmento de estrada num determinado período horário de um dado dia. Este tipo de modelos nunca foi implementado em Portugal e a nível internacional não há conhecimento da aplicação de modelos preditivos para a ocorrência de acidentes em vias com caraterísticas tão heterogéneas como, por exemplo, a EN10, e faltando informação sobre algumas variáveis de grande interesse (como informação específica sobre a intensidade de tráfego). Estes modelos foram implementados numa aplicação digital que apoia o processo de tomada de decisão da GNR de Setúbal. Com esta informação, a GNR de Setúbal envia patrulhas aos segmentos com maior risco de ocorrência de acidentes e a sua presença reduz a probabilidade de ocorrência de um acidente naquele local. Na Figura 5 é apresentado um *output* da aplicação digital relativo à A33+PVG num dado período temporal. Os segmentos vermelhos têm probabilidade superior a um valor pré-estabelecido pela GNR de Setúbal, os segmentos amarelos têm probabilidade superior ao ponto de corte e inferior ao valor pré-estabelecido e os segmentos verdes têm probabilidade inferior ao ponto de corte. Cada troço contém informação sobre o segmento e o respetivo histórico de acidentes.

A aplicação digital desenvolvida tem outras valências, permitindo a visualização do: i) Passado - *dashboard* com consulta do histórico das variáveis principais que caracterizam o acidente; ii) Presente - atlas de suscetibilidade de ocorrência de acidentes e mapas com o indicador de gravidade por troço de estrada; iii) Futuro - predição de ocorrência de acidentes em tempo real para troços das 4 vias e predição de ocorrência de *hotspots* (locais de ocorrência de muitos acidentes). Dado que existe uma reconhecida limitação de recursos para a vigilância rodoviária no distrito de Setúbal, a GNR de Setúbal pode utilizar esta ferramenta preditiva, gerindo melhor a eficiência dos recursos ao seu dispor em prol do aumento da eficácia da segurança rodoviária.

5. CONSIDERAÇÕES FINAIS

Muitas das decisões que afetam a vida das pessoas nos mais variados aspetos são tomadas com base em ideias pré-concebidas, sem critérios rigorosos, mas por afirmações aparentemente irrefutáveis, tornado-se uma prática

Figura 5. Mapa da aplicação digital de apoio à decisão: predição da ocorrência de um acidente de viação numa determinada hora do dia por segmentos de 500 m.

que alguns denominam por "achómetro". Muitas vezes, parece haver certezas numa determinada área de atuação, baseadas na experiência que alguns têm nessa área e nas perceções que lhes são passadas por outros com mais experiência. Contudo, o "achómetro" frequentemente não está calibrado e a sua fiabilidade é muito pequena. A mudança de atitude é fundamental nestes casos, procurando a evidência científica para apoio à tomada de decisão.

Este artigo procurou, com base num projeto, dar exemplos do que se pode conseguir ligando o meio académico e a ciência aplicada transdisciplinar para a resolução de um problema social concreto. Foram apresentados exemplos de alguns resultados alcançados no projeto MOPREVIS (atualizados para este artigo), o qual teve fortes alicerces matemáticos e conjugou a matemática com outras áreas do saber e com o próprio saber da experiência adquirido ao longo do tempo. Este caminho precisa de ser seguido apenas em prol de um objetivo comum: SALVAR VIDAS!

É o conhecimento de base científica que possibilita tomar decisões coerentes e consistentes, definindo uma estratégia proativa e vetores de atuação, de modo a alcançar uma segurança rodoviária mais eficaz. E TODOS ganham com isso!

Com este projeto, um primeiro passo foi dado. E se esse passo levar a que menos uma pessoa tenha morrido nas estradas do distrito de Setúbal, então o projeto já valeu a pena!

REFERÊNCIAS

[1] WHO. "Preventing Injuries and Violence: An Overview". *Technical Report*. World Health Organization: Geneva, Switzerland, 2020.

[2] European Commission. "European Road Safety Observatory". Em *Annual Statistical Report on Road Safety in the EU, 2022*. European Commission, Directorate General for Transport: Brussels, Belgium, 2023.

[3] Silva, C. M., Bravo, J. M., Gonçalves, J., *Impacto Económico e Social da Sinistralidade Rodoviária em Portugal*. CEGE - Centro de Estudos de Gestão do ISEG e Autoridade Nacional de Segurança Rodoviária (ANSR), Lisboa, 2021.

[4] Infante P., Nogueira V., Rebelo Manuel P., Góis P., Afonso A., Santos D., Jacinto G., Saias J., Rego L., Silva M., Carocha Gonçalves N., Rebisco P., Quaresma P., Nogueira P., Pisco Costa R., Clemente R. *A Sinistralidade Rodoviária no Distrito de Setúbal*. Évora: Imprensa da Universidade de Évora, 2023, ISBN 978-972-778-300-7, doi: *https://doi.org/10.24902/uevora.33*.

[5] Infante, P., Afonso, A., Jacinto, G., Rego, L., Nogueira, P., Silva, M., Nogueira, V., Saias, J., Quaresma, P., Santos, D., Gois, P., Manuel, P.R., Some Determinants for Road Accidents Severity in the District of Setúbal. In: Bispo, R., Henriques-Rodrigues, L., Alpizar-Jara, R., de Carvalho, M. (eds), *Recent Developments in Statistics and Data Science*. Springer Proceedings in Mathematics and Statistics, 398, 203-214, 2022, ISBN 978-3-031-12765-6, *https://doi.org/10.1007/978-3-031-12766-3_14*.

[6] Infante, P., Jacinto, G., Afonso, A., Rego, L., Nogueira, V., Quaresma, P., Saias, J., Santos, D., Nogueira, P., Silva, M., Costa, R.P., Gois, P., Manuel, P.R., "Comparison of Statistical and Machine-Learning Models on Road Traffic Accident Severity Classification". *Computers*, 11(5), 80, 2022, doi: *https://doi.org/10.3390/computers11050080*.

[7] Infante P., Jacinto G., Afonso A., Rego L., Nogueira P., Silva M., Nogueira V., Saias J., Quaresma P., Santos D., Góis P., Manuel P.R., "Factors That Influence the Type of Road Traffic Accidents: A Case Study in a District of Portugal". *Sustainability*, 15(3), 2352, 2023, doi: *https://doi.org/10.3390/su15032352*.

[8] Nogueira, P., Silva, M., Infante, P., Nogueira, V., Manuel, P., Afonso, A., Jacinto, G., Rego, L., Quaresma, P., Saias, J., Góis, P.. Manuel, P. R., "Learning from Accidents: Spatial Intelligence Applied to Road Accidents with Insights from a Case Study in Setúbal District", Portugal. *ISPRS International Journal of Geo-Information*, 12(3), 93, 2023, doi: *https://doi.org/10.3390/ijgi12030093*.

[9] Infante P, Jacinto G, Santos D, Nogueira P, Afonso A, Quaresma P, Silva M, Nogueira V, Rego L, Saias J, Góis, P.. Manuel, P. R., "Prediction of Road Traffic Accidents on a Road in Portugal: A Multidisciplinary Approach Using Artificial Intelligence, Statistics, and Geographic Information Systems". *Information*, 14(4):238, 2023, *doi: https://doi.org/10.3390/info14040238*.

[10] Carey, R. G., *Improving HealthCare With Control Charts: Basic and Advanced SPS Methods and CAse Studies*. ASQ Quality Press, Milwaukee, Wisconsin, 2003.

[11]Dobson, A., Barnett, A., *An Introduction to Generalized Linear Models*. 4th Edition, Chapman & Hall, 2018.

[12] Hosmer Jr., D., Lemeshow, S., Sturdivant, R., *Applied Logistic Regression*. 3rd Edition, John Wiley & Sons, Inc., US, 2013.

[13] James, G., Witten, D., Hastie, T., Tibshirani, R., *An Introduction to Statistical Learning with Applications in R*. 2nd Edition, Springer, US, 2021.

[14] Witten, I., Frank, E., Hall, M., Christopher J., *Data Mining: Practical Machine Learning Tools and Techniques*. 4th Edition, Morgan Kaufmann, US, 2017.

[15] Breiman, L., "Random Forests". *Machine Learning*, 45(1), 5-32, 2001.

[16] Pang, Su-lin,Gong, Ji-zhang, "C5.0 Classification Algorithm and Application on Individual Credit Evaluation of Banks". *Systems Engineering-Theory & Practice*, 29(12), 94-104, 2009, doi *https://doi.org/10.1016/S1874-8651(10)60092-0*.

[17] Chen, T., Guestrin, C., "XGBoost: A Scalable Tree Boosting System". In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, 785-7944, 2016, New York, US: ACM, doi: *https://doi.org/10.1145/2939672.2939785*.

[18] Menardi, G.; Torelli, N., "Training and Assessing Classification Rules with Imbalanced Data". *Data Min Knowl Disc*, 28, 92-102, 2014, doi: *https://doi.org/10.1007/s10618-012-0295-5*.

[19] He, H.; Garcia, "Learning from Imbalanced Data". *EEE Transactions on Knowledge and Data Engineering*, 21 (9), 1263–1284, 2009, doi:*10.1109/TKDE.2008.239*.

Agradecimentos:
Os resultados obtidos no Projeto MOPREVIS só foram possíveis devido a uma equipa transdisciplinar que desde o início realizou um trabalho excecional. Pela Universidade de Évora, um profundo agradecimento aos Professores Vítor Nogueira, Anabela Afonso, Gonçalo Jacinto, José Saias, Paulo Quaresma, Pedro Nogueira e Rosalina Pisco Costa, e aos Bolseiros Mestres Daniel Santos, Leonor Rego, Marcelo Silva e Patrícia Góis. Pela GNR de Setúbal, um profundo agradecimento ao Tenente-Coronel Nuno Carocha Gonçalves, ao Cabo-Chefe Paulo Rebisco e ao Cabo Rui Clemente. E um profundo agradecimento ao timoneiro que desencadeou este projeto, Coronel Paulo Rebelo Manuel, dando sempre contributos fundamentais para o seu sucesso, quer inicialmente pela GNR de Setúbal (enquanto Comandante do Comando Territorial), quer depois integrado na equipa da Universidade de Évora. Finalmente, fica um agradecimento especial ao Instituto Português do Mar e da Atmosfera, à Infraestruturas de Portugal, à Waze Portugal e à Autoridade Nacional de Segurança Rodoviária, pelo apoio dado ao longo do projeto.

SOBRE O AUTOR

Paulo Infante é Professor Associado na Universidade de Évora, doutorado em Matemática pela mesma Universidade. Tem desempenhado diversos cargos, como o de Pró-Reitor para as áreas da Inovação, Transferência, Empreendedorismo e Cooperação, sendo atualmente Coordenador da Linha de Investigação em Modelação Matemática em Ciências da Vida e Aplicações do Centro de Investigação em Matemática e Aplicações (CIMA) e membro das Comissões de Curso da licenciatura em Matemática Aplicada à Economia e à Gestão e da licenciatura e mestrado em Inteligência Artificial e Ciência de Dados. Responsável por unidades curriculares dos diferentes ciclos de estudo, tendo orientado estudantes de licenciatura, mestrado e doutoramento. Coordenou vários projetos entre a Universidade e a comunidade, com diversas publicações em metodologia estatística e em modelação estatística e análise de dados, em diferentes áreas. Foi investigador responsável do projeto Modelação e Predição de Acidentes de Trânsito no Distrito de Setúbal (MOPREVIS), financiado pela FCT. Além da modelação estatística, os seus interesses atuais de investigação estão nas áreas de ciência de dados, controlo de qualidade e análise de sobrevivência.

Secção coordenada pela PT-MATHS-IN, Rede Portuguesa de Matemática para a Indústria e Inovação *pt-maths-in@spm.pt*