Ana Rita Diniz Teles Nunes

Licenciada em Matemática Aplicada e Computação

MODELAÇÃO ESPACIAL DE ACIDENTES RODOVIÁRIOS NA CIDADE DE LISBOA

Dissertação para obtenção do Grau de Mestre em Matemática e Aplicações

Orientador: Doutora Isabel Cristina Maciel Natário, Faculdade de Ciências e Tecnologia, UNL, Portugal.

Co-orientador: Doutora Sílvia Shrubsall, Instituto Superior Técnico, UTL, Portugal.

Júri:

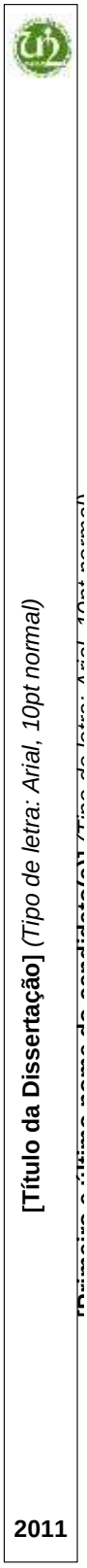
Presidente: Prof. Doutor Manuel Leote Tavares Inglês Esquível

Vogal: Prof. Doutora Gracinda Rita Diogo Guerreiro

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Dezembro 2011

LOMBADA


[Título da Dissertação] (Tipo de letra: Arial, 10pt normal)
[Primeiro e último nome do candidato(a)] (Tipo de letra: Arial, 10pt normal)
2011

Ana Rita Diniz Teles Nunes

Licenciada em Matemática Aplicada e Computação

MODELAÇÃO ESPACIAL DE ACIDENTES RODOVIÁRIOS NA CIDADE DE LISBOA

Dissertação para obtenção do Grau de Mestre em Matemática e Aplicações

Orientador: Doutora Isabel Cristina Maciel Natário, Faculdade de Ciências e Tecnologia, UNL, Portugal.

Co-orientador: Doutora Sílvia Shrubsall, Instituto Superior Técnico, UTL, Portugal.

Júri:

Presidente: Prof. Doutor Manuel Leote Tavares Inglês Esquível

Vogal: Prof. Doutora Gracinda Rita Diogo Guerreiro

FACULDADE DE CIÊNCIAS E TECNOLOGIA
UNIVERSIDADE NOVA DE LISBOA

Dezembro 2011

COPYRIGHT

Agradecimentos

Um estudo tão abrangente e detalhado como este só foi possível graças à mobilidade de informação e tratamento de dados de uma equipa multidisciplinar. Desta forma, começo por agradecer à Autoridade Nacional de Segurança Rodoviária pela cedência da profunda base de dados dos acidentes rodoviários na cidade de Lisboa. Do mesmo modo agradeço ao Laboratório Nacional de Engenharia Cívil pelo trabalho que fez na georreferenciação das localizações dos acidentes e pela partilha dessa informação. A vós, um muito obrigado!

Este trabalho teve ainda o apoio dos Fundos Nacionais através da FCT – Fundação para a Ciência e a Tecnologia, no âmbito do projecto SACRA (*Spatial Analysis of Child Road Accidents*), PTDC/TRA/66161/2006, à qual agradeço.

Quero também agradecer à Ana Raposo por todo o apoio prestado e por todo o trabalho que efectou anteriormente a mim e possibilitou a realização deste estudo. Ao Frederico Henriques por toda a ajuda concedida a vários níveis e que foi crucial para o desenvolvimento desta dissertação.

Às minhas orientadoras, Professora Isabel Natário e Professora Sílvia Shrubsall, um especial obrigado por todo o apoio e atenção prestada ao longo destes meses e por toda a compreensão prestada em todos os sentidos.

Por fim, um agradecimento muito especial aos meus pais, por todo o enorme apoio que sempre me prestaram durante todos estes meses e mais alguns e que, sem dúvida, contribuiram para o desenvolvimento deste trabalho.

Resumo

Os acidentes rodoviários em meio urbano contribuem para o decréscimo da qualidade de vida e para a inequidade social das cidades. Em Portugal, o número e gravidade dos acidentes rodoviários decresceu muito nos últimos vinte anos, mas a situação continua a ser preocupante.

Na cidade de Lisboa, apesar da evolução positiva que se tem verificado nos últimos anos, o número de acidentes rodoviários continua elevado, afastando-se inaceitavelmente da média europeia. Deste modo, com base no conjunto de todos os acidentes com vítimas ocorridos na cidade de Lisboa entre 2004 e 2007, foi realizada uma análise exploratória das características dos mesmos, com vista a encontrar os factores mais importantes na explicação da ocorrência e gravidade dos acidentes com vítimas. Recorrendo aos modelos lineares generalizados, concluiu-se que factores humanos, ambientais e circunstanciais têm influência na gravidade dos acidentes e algumas variáveis de exposição foram consideradas importantes na explicação da ocorrência dos mesmos por freguesia.

Estes dados foram ainda georreferenciados, com o intuito de uma exploração da natureza espacial dos mesmos, tentando perceber padrões geográficos existentes, identificando factores de risco associados. Considerando a localização da ocorrência de cada acidente como aleatória, enquadra-se este problema na teoria dos processos pontuais espaciais. Uma análise envolvendo testes de hipótese e descrições sumárias da localização dos acidentes permitiu concluir que estes não se encontram uniformemente distribuidos no espaço, rejeitando-se a designada *hipótese de aleatoriedade completa*.

Pretende-se, assim, contribuir para a identificação de medidas eficientes, tendo em conta as condições prevalencentes a nível local, bem como facilitar a realização de comparações com realidades internacionais no que respeita à segurança rodoviária.

Palavras-chave: Acidentes Rodoviários, Modelos Lineares Generalizados, Modelação espacial, Processos Pontuais Espaciais.

Abstract

Road accidents in urban areas contribute to the decline in quality of life and social inequity in cities. In Portugal, the number and severity of road accidents has decreased considerably in the last twenty years, but the situation remains worrying.

In Lisbon, despite the positive developments in recent years, the number of road accidents remains higher than the European average. Descriptive explanatory analysis were carried out to reported injury road accidents that occurred in Lisbon between 2004 and 2007, in order to find the most important factors that explain the number and severity of injury accidents. Results from subsequent application of generalized linear models, indicate that human factors, environmental and circumstantial factors influence the severity of accidents and some exposure variables were considered important in explaining their occurrence by parish.

Road accident data were also geo-referenced, with a view to exploiting their spatial nature and subsequently try to understand existing spatial patterns and identify associated risk factors. Considering as random the location of the occurrence of each accident, this problem can be studied using theory of spatial point processes. Analysis involving hypothesis testing and brief descriptions of the location of the accidents concluded that these are not randomly distributed in space, rejecting the so-called hypothesis of *complete spatial randomness*.

This study intends to contribute to the identification of effective measures, taking into account the local conditions, as well as allow comparison with international road safety contexts.

Keywords: Road accidents, Generalized Linear Models, Spatial Modelling, Spatial Point Process.

Índice do Texto

Índice de Figuras

Índices de tabelas

Índice de siglas

ANSR – Autoridade Nacional de Segurança Rodoviária.

CARE – *Community database on Accidents on the Roads in Europe*.

CSR – *Complete Spatial Randomness* (Aleatoriedade Espacial Completa).

DGV – Direcção Geral de Viação.

ENSR – Estratégia Nacional de Segurança Rodoviária.

GNR – Guarda Nacional Republicana.

INE – Instituto Nacional de Estatística.

IRTAD - International Road Traffic and Accident Database.

LNEC – Laboratório Nacional de Engenharia Cívil.

NA – Dado omisso.

NISPT – Núcleo de Infra-Estruturas, Sistemas e Políticas de Transporte.

UE – União Europeia.

USM – *Urban Safety Management*.

Glossário

Acidente com vítimas – Ocorrência na via pública envolvendo pelo menos um veículo registado pelas entidades oficiais da qual resultem vítimas e/ou danos materiais. Os acidentes com vítimas podem ser fatais/mortais, graves ou ligeiros dependendo da gravidade dos ferimentos das vítimas.

Acidente mortal – Acidente do qual resulte pelo menos uma vítima mortal.

Acidente grave – Acidente do qual resulte pelo menos um ferido grave, não tendo ocorrido qualquer morte.

Acidente ligeiro – Acidente do qual resulte pelo menos um ferido leve, e em que não se tenham registado vítimas mortais nem feridos graves.

Condutor – Pessoa que detém o comando de um veículo na via pública.

Ferido grave – vítima de acidente cujos danos corporais obriguem a um período de hospitalização superior a 24 horas.

Ferido ligeiro – Vítima de acidente que não seja considerado ferido grave.

Morto ou vítima mortal (30 dias) – Vítima com ferimentos decorrentes do acidente rodoviário cujo óbito ocorra no período de 30 dias após o mesmo.

Morto ou vítima mortal (local) – Vítima de acidente cujo óbito ocorra no local do evento ou no seu percurso até à unidade de saúde.

Passageiro – Pessoa afecta a um veículo na via pública e que não seja condutora.

Peão – Pessoa que transita na via pública a pé. Consideram-se ainda peões todas as pessoas que conduzam à mão, velocípedes.

Variável categórica – Variável estatística medida numa escala nominal.

Vítima rodoviária– Ser humano que, em consequência de acidente rodoviário, sofra danos corporais.

1 Introdução

1.1 Contexto

A sinistralidade rodoviária é um problema grave em Portugal a que está associada uma elevada taxa de mortalidade. Para fazer face a esta problemática, ao longo dos últimos anos têm sido aplicadas algumas medidas, como por exemplo alterações ao código da estrada e campanhas de sensibilização, e criados novos enquadramentos instituicionais, como é o caso de comissões de acompanhamento e da Autoridade Nacional de Segurança Rodoviária (ANSR). Tem-se verificado, também, um aumento considerável do nível de consciencialização em relação à segurança rodoviária que conduziram, por exemplo, à formulação da Estratégia Nacional de Segurança Rodoviária (ENSR), em vigor até ao ano 2015. Estas acções visam a redução do número e gravidade de acidentes, de vítimas mortais e feridos graves, na tentativa de aproximação aos valores da média europeia, que se situam numa posição mais favorável. Desta forma, um dos grandes desafios nacionais actuais é reduzir a elevada taxa de sinistralidade existente, passando também por "colocar Portugal entre os 10 países da U.E. com mais baixa sinistralidade rodoviária" (ANSR & ISCTE, p. 10).

Nos últimos vinte anos, Portugal registou uma evolução positiva significativa relativamente à redução da sinistralidade rodoviária (Tabela 1.1), tendo mesmo sido o país com a melhor evolução de toda a Europa entre 1999 e 2006 (54.5% versus 23.8% da média comunitária). Reconhece-se, apesar do mérito Português, ser mais fácil um decréscimo considerável na sinistralidade quando os valores ainda são muito elevados do que quando o número de acidentes e vítimas por milhão de habitantes já é reduzida, como no caso do Reino Unido, por exemplo.

Tabela 1.1 Evolução da sinistralidade rodoviária europeia desde 1975 até 2006.

	Evo 75/06	Evo 91/06	Evo 03/06	Evo 99/06	Evo 99/02
Alemanha	-71,8%	-56,3%	-22,5%	-34,7%	-12,6%
Áustria	-74,8%	-58,2%	-27,0%	-37,8%	-11,9%
Bélgica	-59,1%	-47,9%	-16,2%	-28,5%	-7,3%
Chipre		-36,0%	-17,6%	-32,1%	-19,4%
Dinamarca	-64,5%	-50,8%	-27,5%	-40,2%	-11,3%
Eslováquia		-16,4%	-19,2%	-19,2%	-5,8%
Eslovénia	-61,2%	-44,6%	5,8%	-24,3%	-20,1%
Espanha	-48,8%	-62,6%	-34,6%	-41,0%	-9,0%
Estónia		-51,4%	25,6%	-9,5%	-2,4%
Finlândia	-65,9%	-47,6%	-9,6%	-21,4%	-4,8%
França	-72,5%	-59,2%	-25,7%	-48,3%	-11,0%
Grécia	8,5%	-27,5%	2,7%	-23,1%	-23,6%
Hungria	-18,6%	-36,3%	-0,8%	2,4%	10,2%
Irlanda	-52,8%	-31,0%	2,4%	-21,6%	-13,5%
Itália	-50,5%	-35,7%	-13,2%	-22,0%	0,0%
Letónia		-49,0%	-22,4%	-29,8%	-12,3%
Lituânia		-29,7%	8,8%	5,2%	-5,2%
Luxemburgo	-77,5%	-63,9%	-33,9%	-42,6%	2,9%
Malta		-44,4%	-37,5%	127,3%	272,7%
P. Baixos	-74,8%	-49,4%	-31,7%	-37,7%	-11,6%
Polónia	-16,9%	-33,8%	-7,4%	-21,3%	-12,6%
Portugal	-73,7%	-71,8%	-38,5%	-54,5%	-20,0%
Reino Unido	-52,9%	-32,5%	-9,7%	-8,2%	-1,6%
Rep. Checa	-36,1%	-19,4%	-26,8%	-26,2%	-0,7%
Suécia	-65,8%	-43,7%	-16,9%	-25,8%	-4,5%
M. Europeia		-46,9%	-16,5%	-28,3%	-8,3%

Fonte: ENSR, citando fontes IRTAD (até 1990) e CARE (a partir de 1991)

No entanto, a posição de Portugal, no contexto da União Europeia, ainda não é satisfatória, apresentando uma razão de mortos por milhão de habitantes elevado e superior à média, que em 2010 se traduzia por 83 versus 76.

É de fazer notar que a evolução do número de acidentes mortais, isto é, que resultaram em mortes, tem sido muito mais significativa relativamente à do número de acidentes no total, com uma redução de 49% contra apenas 16% verificados no número de acidentes com vítimas no geral, segundo dados das estatísticas europeias, base de dados CARE referente ao período de 2001-2010. Mas é de fazer notar que, segundo a ENSR, "a sustentabilidade da diminuição do número de mortos só pode ser alcançada através da redução do total de acidentes com vítimas e das suas consequências" (ANSR&ISCTE, 2009, p.11).

Em suma, apesar de ser claro que Portugal tem evoluído positivamente na diminuição da sinistralidade rodoviária, o número de vítimas ainda é inaceitável. É, assim, importante continuar a procurar novas medidas, de forma a atingir valores mais baixos de sinistralidade, na tentativa de minorar o sofrimento pessoal dos utentes rodoviários, diminuir os custos sociais dos acidentes e aumentar a qualidade de vida no país.

1.2 Segurança rodoviária em meio urbano

O aumento considerável da taxa de motorização associado a uma rápida urbanização que se tem vindo a verificar nas últimas décadas, contribuiu para um aumento do número de acidentes rodoviários em meio urbano, que durante muito anos ocorreu nos países europeus. Esse crescimento das cidades obrigou ao aumento da capacidade das suas redes viárias, comprometendo, contudo, a segurança dos utilizadores mais vulneráveis.

Segundo dados estatísticos europeus, em 2008 mais de metade dos acidentes rodoviários ocorreram em zonas urbanas, com um número considerável de mortes que ascendeu aos 13502 no total de 19 países da União Europeia, correspondente a cerca de 38% do total do número de acidentes com vítimas. Contudo, muitas das vítimas destes acidentes não são os ocupantes dos veículos, mas os peões, motociclistas e ciclistas.

O problema da sinistralidade rodoviária a nível urbano só pode ser ultrapassado através do diagnóstico das características da mesma, atendendo ao seu número, distribuição e características dos acidentes, identificação das principais causas e, ainda, de informações complementares como, por exemplo, demográficas, de tráfego e económicas.

Uma abordagem que tem sido adoptada frequentemente no diagnóstico dos problemas de segurança, e posterior tratamento, envolve a criação de mapas de acidentes que permitem identificar os designados "pontos de acumulação de acidentes"[1] e a elaboração do respectivo diagrama de colisão. Ao invés de estarem concentrados em "pontos de acumulação", os acidentes podem estar dispersos numa área vasta, implicando análises mais complexas e a implementação de medidas de natureza diferente.

Assim, nos anos 90, surgiu a Gestão de Segurança Urbana (USM[2]), uma nova abordagem para redução e prevenção dos acidentes rodoviários nas áreas urbanas, incluindo um conjunto de estratégias, políticas e directrizes que pretendem apoiar as autoridades locais no desenvolvimento de

[1] Pontos de acumulação de acidentes são locais onde os acidentes se acumulam, existindo um elevado risco para a sua ocorrência. No entanto, é de fazer notar que não existe um consenso internacional para esta definição.
[2] *USM* é a abreviatura do do inglês para *Urban Safety Management*.

planos e respectivas medidas de segurança rodoviária em meio urbano. No entanto, a nível europeu, a aplicação desta abordagem ainda não tem sido sistemática, apesar do projecto europeu *Developing Urban Management and Safety* (EU, 2001), desenvolvido com o intuito de encorajar a sua implementação e definir iniciativas de segurança rodoviária a aplicar a nível urbano nos países da União Europeia.

É importante referir o papel crucial que as autoridades locais têm na redução da sinistralidade rodoviária, havendo necessidade de um esforço concertado entre elas. É, nomeadamente, essencial uma forte ligação entre câmaras municipais (responsáveis pela gestão das vias) e as autoridades policiais na recolha e divulgação de informação das características e localização dos acidentes. A análise destes dados apoia a definição de estratégias quanto a esta problemática e a identificação de soluções, que devem ser aplicadas de um modo multidisciplinar.

Em Portugal, o cenário da sinistralidade rodoviária em meio urbano é ainda preocupante. Os desenvolvimentos que o país tem revelado a nível da redução dos acidentes e vítimas mortais centram-se, essencialmente, nas estradas inter-urbanas. Entre 2004 e 2008, cerca de 69% do total dos acidentes rodoviários com vítimas aconteceram em zonas urbanas, requerendo, portanto, especial atenção. Neste sentido, e como complemento da Estratégia Nacional de Segurança Rodoviária (ENSR), a ANSR apoiou a elaboração do documento *Guia para a elaboração de Planos Municipais de Segurança Rodoviária* (ANSR, 2009). Este documento pretende apoiar as Câmaras Municipais na definição, desenvolvimento e aplicação de medidas adequadas à segurança rodoviária nos seus municípios. Contudo, a segurança rodoviária ainda não é uma actividade sistematicamente desenvolvida por essas entidades.

Apesar das acções visíveis na área da segurança rodoviária no panorama nacional, verifica-se que a preparação e execução de planos municipais de segurança rodoviária em Portugal ainda constituem um desafio. Uma fraqueza relevante é a morosidade no acesso aos dados de acidentes, em particular os que incluem a sua localiação exacta, que permitem o diagnóstico da sinistralidade urbana. No âmbito do projecto de investigação *Spatial Analysis of Child Road* Accidents (SACRA) foi possível estabelecer protocolos com várias entidades, nomeadamente a ANSR e o Laboratório Nacional de Engenharia Cívil (LNEC), que permitiram constituir uma base de dados de acidentes para a cidade de Lisboa, com dimensão e detalhe incomum em Portugal. Outros projectos do NISPT (Núcleo de Infra-Estruturas, Sistemas e Políticas de Transporte do Instituto Superior Técnico) contribuiram com dados de exposição, nomeadamente tráfego rodoviário, enriquecendo a referida base de dados.

Verificou-se em Lisboa, no ano de 2010, cerca de 7449 acidentes com vitimas, entre os quais 89 resultaram em mortes. Actualmente, assiste-se a uma maior consciencialização para esta problemática na cidade e tenciona-se desenvolver um conjunto de políticas e intervenções na área da Mobilidade e Segurança Rodoviária, visando a redução do tráfego, o aumento da segurança dos peões e utentes de

risco e a introdução de modos de mobilidade suave. Novas ferramentas que têm sido desenvolvidas nesta matéria incluem o projecto IRUMS- Infra-Estruturas Rodoviárias Urbanas Mais Seguras, desenvolvido pelo LNEC como um projecto académico, sem, no entanto, propor soluções práticas. Outras propostas de acções a desenvolver envolvem a criação de uma base de dados sobre a rede rodoviária e sinistralidade da cidade, o desenvolvimento de modelos de estimativas de frequência dos acidentes e definição de métodos para a identificação de zonas de acumulação dos mesmos, bem como um conjunto de campanhas de sensibilização e educação dos utentes. Estes dados, se analisados devidamente, poderão contribuir para a identificação de medidas de segurança rodoviária eficazes a aplicar em Lisboa, quer na abordagem dos pontos de acumulação, quer na de USM, contribuindo para a redução do número e gravidade dos acidentes na capital do país.

A melhoria da sinistralidade rodoviária beneficia da coordenação de esforços e actuação entre várias entidades responsáveis pela segurança rodoviária, nomeadamente a nível municipal, reconhecendo a importância do poder autárquico na resolução deste problema. Para isso, é fundamental o diagnóstico prévio da situação actual e a sua aplicação em Lisboa poderá constituir um incentivo à sua aplicação noutras cidades portuguesas.

1.3 Objectivos

No contexto da segurança rodoviária em meio urbano e, em particular, na sinistralidade registada na cidade de Lisboa, é essencial, após a recolha e sistematização dos dados relevantes sobre a matéria, um tratamento estatístico de exploração dos mesmos e a utilização de modelos adequados que identifiquem e incorporem os principais factores de risco associados ao problema.

Assim, esta dissertação visa, essencialmente, o desenvolvimento de conhecimentos que apoiem a identificação de políticas e medidas que contribuam na diminuição da sinistralidade rodoviária na cidade de Lisboa.

Esse objectivo passa por uma análise fina e detalhada dos acidentes rodoviários nesta cidade, recorrendo a uma metodologia capaz de identificar os factores de risco associados à ocorrência dos acidentes com vítimas.

Adicionalmente, a georreferenciação da localização dos acidentes permite uma exploração da natureza espacial da ocorrência dos mesmos, na tentativa de identificar padrões geográficos existentes e os respectivos factores de risco associados. Desta forma, espera-se uma caracterização pormenorizada do problema, contribuindo, também, para a identificação de áreas da cidade de Lisboa

com maior incidência de acidentes, com o intuito de encontrar e adoptar políticas eficazes que contribuam para a melhoria da segurança rodoviária das mesmas.

1.4 Metodologia

Os métodos estatísticos utilizados neste estudo são de natureza diversa e actual e foram aplicados em duas fases. A par destas metodologias é feita uma análise exploratória dos dados obtidos, identificando as principais causas a nível estrutural e urbano, humano e proveniente de factores circunstanciais, como factores atmosféricos, e das freguesias que potenciam a ocorrência dos acidentes com vítimas na cidade.

Depois de uma análise descritiva dos dados, numa primeira fase, aplicou-se uma metodologia baseada em modelos lineares generalizados, com o intuito de descrever os factores mais relevantes na ocorrência dos acidentes com vítimas em Lisboa. Estes modelos visam não só explicar essa influência no número de acidentes mas também na sua gravidade. Para tal, recorreu-se aos modelo binomial (logit) e de log-Poisson.

Para além disso, a georreferenciação da localização dos acidentes permitiu uma abordagem através de processos pontuais espaciais, permitindo a estimação da superfície de risco associada em função de factores extrínsecos ao acidente.

Estes estudos estatísticos foram implementados com recurso ao software estatístico *Rproject*.

1.5 Estrutura

Nesta dissertação, os acidentes analisados reportam-se ao período de 2004 a 2007, pelo que qualquer dado que esteja fora deste período será mencionado devidamente.

Distinguem-se duas fases na análise dos acidentes rodoviários com vítimas em Lisboa: uma primeira fase em que é feita uma análise exploratória dos dados, aplicando-se a teoria dos modelos lineares generalizados; e uma segunda fase em que se explora a natureza espacial dos acidentes, recorrendo-se à teoria dos processos pontuais espaciais.

Deste modo, a dissertação desenvolve-se, essencialmente, ao longo de três capítulos. No capítulo 2 é feita uma análise preliminar dos dados adquiridos, tentando perceber que factores influenciam

mais a ocorrência e gravidade dos acidentes rodoviários com vítimas em Lisboa, com base, essencialmente, em gráficos descritivos.

No capítulo 3, após selecção informal das variáveis que se mostraram mais influentes na ocorrência e gravidade dos acidentes, é feita uma análise de regressão, baseada nos modelos lineares generalizados (GLM). Após um enquadramento teórico desta temática, é efectuada a referente análise na secção 3.8, distinguindo-se dois modelos de regressão: o modelo binomial e o modelo de Poisson. Este primeiro é aplicado para modelar a gravidade dos acidentes, tendo em conta que se trata de uma variável resposta dicotómica, e o segundo modela o número de ocorrências de acidentes.

No capítulo 4 inicia-se a exploração da natureza espacial dos acidentes com vítimas em Lisboa, considerando a localização da ocorrência dos mesmos. Para tal, recorre-se à teoria dos processos pontuais espacias. Após uma introdução teórica dos mesmos nas secções 4.1 a 4.4 é realizada a análise espacial dos acidentes na secção 4.5. Nesta secção distinguem-se, ainda, dois momentos: um primeiro em que é feita uma análise para cada ano separadamente (secção 4.5.1), e um segundo em que a análise é feita no global, ou seja, com todos os acidentes considerados do período em estudo.

Por fim, no capítulo 5 apresentam-se as conclusões desta dissertação, bem como algumas sugestões para desenvolvimentos futuros nesta matéria.

2 Acidentes rodoviários com vítimas em Lisboa

Após a recolha dos dados é essencial uma análise exploratória dos mesmos para a realização de procedimentos de análises posteriores que se adequem a esses. Deste modo, foi feita uma análise preliminar dos dados recolhidos, obtendo-se uma visão mais global das suas variações.

Os dados recolhidos relativos aos acidentes rodoviários com vítimas em Lisboa respeitam ao período de 2004 a 2008. Estes dados foram obtidos pela Autoridade Nacional de Segurança Rodoviária (ANSR), que gere a base de dados dos acidentes e partilhou essa informação. A base de dados referida advém do preenchimento por parte das forças de segurança, Guarda Nacional Republicana (GNR) e Polícia de Segurança Pública (PSP), ao tomarem conhecimento de um acidente rodoviário (Anexo 1). Posteriormente, essa informação foi complementada com georreferenciação efectuada pelo Laboratório Nacional de Engenharia Civil (LNEC) e, após esse processo, foi uniformizada, tratada e completada pelo NISPT no âmbito do projecto SACRA. Dados complementares, nomeadamente de tráfego, características geométricas das vias e localização de equipamentos urbanos foram obtidos pelo NISPT no âmbito de outros projectos.

Os dados provenientes da ANSR dizem respeito aos anos de 2004 a 2008, embora os dados do LNEC apenas se refiram aos anos 2004 a 2007. Para facilitar a análise considerou-se, então, apenas os acidentes com vítimas referentes aos anos de 2004 a 2007. No entanto, cerca de metade dos acidentes ocorridos e representados na ANSR foram, também, excluídos aquando do cruzamento de dados provenientes das duas entidades, visto que não foi possível obter a georreferenciação de todos, tornando-se uma desvantagem considerável para este estudo.

Os dados encontram-se divididos em três categorias, havendo diferente informação relativa a peões, condutores e veículos envolvidos no acidente. No entanto, algumas características, devido à falta de exactidão da informação ou por serem consideradas irrelevantes para o estudo, tiveram de ser omitidas.

Esta análise distinguiu, ainda, os acidentes quanto à gravidade, sendo estes classificados como "Graves" se resultou em pelo menos um ferido grave ou um morto e "Ligeiros", caso contrário.

A amostra final é, assim, constituída por 9263 acidentes, sendo 994 graves e 8269 ligeiros, e inclui informação sobre factores demográficos, ambientais e características urbanas envolventes. Desta forma, foi feita uma análise exploratória com uma abordagem por temas.

2.1 Análise exploratória dos principais factores

Nesta secção será feita uma descrição dos dados e uma análise exploratória dos mesmos, recorrendo às características observadas dos acidentes rodoviários com vítimas em Lisboa. Essas variáveis foram agrupadas por temas de modo a facilitar a análise.

a) Natureza do acidente

Quanto à natureza do acidente, esta subdivide-se em dezassete tipos:

Despiste	1	Despiste simples, com transposição do separador lateral
	2	Com dispositivo de retenção
	3	Sem dispositivo de retenção
	4	Com transposição do dispositivo de retenção lateral
	5	Com capotamento
	6	Com colisão com veículo imobilizado ou obstáculo
	7	Com fuga

Colisão	8	Frontal
	9	Traseira com outro veículo em movimento
	10	Lateral com outro veículo em movimento
	11	Com veículo ou obstáculo na faixa de rodagem
	12	Choque em cadeia
	13	Com fuga
	14	Outras situações

Atropelamento	15	De peões
	16	De animais
	17	Com fuga

A Figura 2.1 traduz a distribuição do número de acidentes observado por natureza do mesmo, concluindo-se que o seu número é consideravelmente mais elevado para o atropelamento de peões, seguido de colisão lateral com outro veículo em movimento.

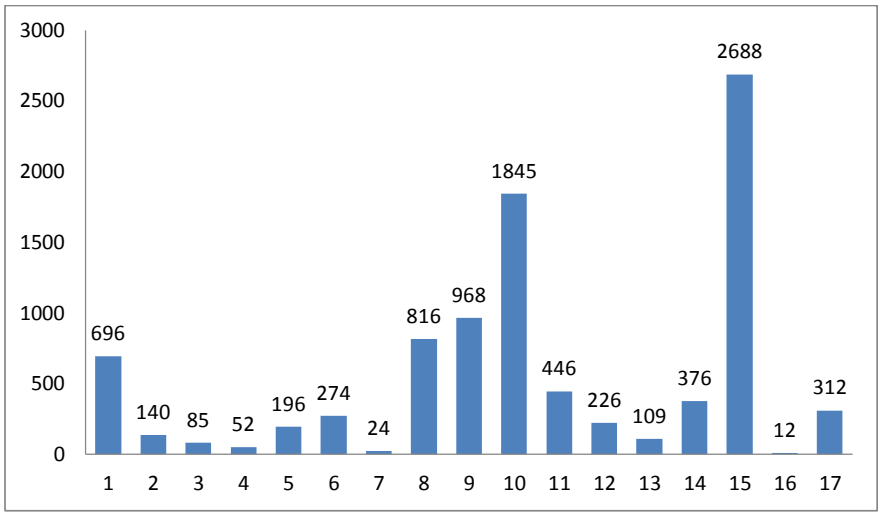


Figura 2.1 Número de acidentes com vítimas consoante a natureza do acidente.

Para uma compreensão facilitada, optou-se por considerar apenas três tipos de acidentes quanto à sua natureza, de acordo com a divisão apresentada atrás: Despiste, Colisão e Atropelamento.

Assim, o tipo de acidente mais frequente com vítimas é a colisão de veículos, como se pode observar pela Figura 2.2.

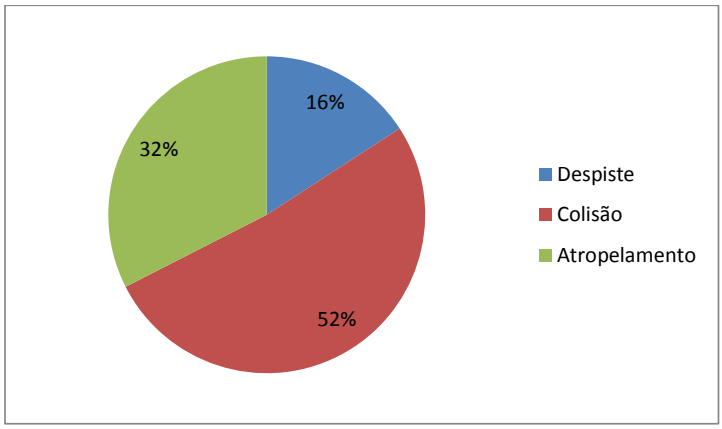


Figura 2.2 Distribuição do número de acidentes por tipo de acidente.

Quanto à gravidade tendo em conta o tipo de acidente (Figura 2.3), verifica-se que a proporção de atropelamentos e despistes é maior para acidentes graves relativamente à percentagem dos mesmos para acidentes ligeiros. Tem-se, ainda, que os atropelamentos são a maior causa de acidentes graves em Lisboa no período considerado. Por sua vez, as colisões geram mais acidentes ligeiros que graves.

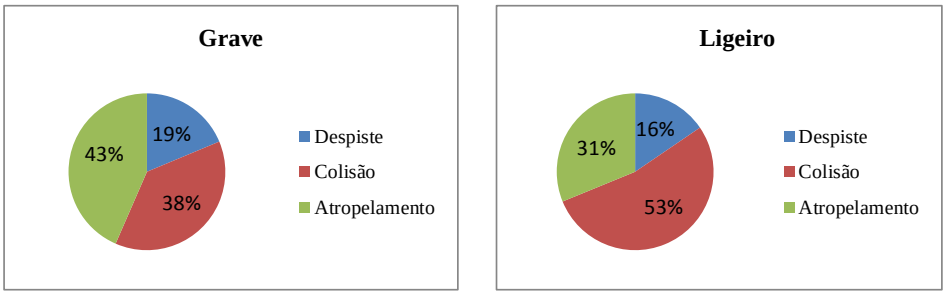


Figura 2.3 Distribuição do número de acidentes por tipo de acidentes, para acidentes Graves e Ligeiros.

(verificar)

b) Factores geográficos

Uma das variáveis em estudo é a freguesia onde ocorreu o acidente. Na data referente aos dados, a cidade de Lisboa contava com cinquenta e três freguesias (Anexo 2)Anexo 2: Mapa das freguesias da cidade de Lisboa, apesar de actualmente contar com mais uma, que não será contabilizada neste estudo.

A freguesia que apresenta mais acidentes rodoviários com vítimas na base de dados considerada é Santa Maria dos Olivais, com um total de 1092 acidentes, seguida de Alcântara e Campo Grande, ambas com um total de 572 acidentes (Figura 2.4). Ao considerar-se os acidentes por gravidade (Graves e Ligeiros), verifica-se que a situação mantém-se (Figura 2.5). Note-se, ainda, que o número de acidentes graves é relativamente menor que o número de acidentes ligeiros.

No entanto, tendo em conta que, por exemplo, Santa Maria dos Olivais é a freguesia com maior área e população, tal não significa, necessariamente, que esteja propensa a mais acidentes por unidade de área ou por habitante (Anexo 3). Para tal, considerou-se o número de acidentes por 10000 m^2 de área e por 1000 habitantes (Figura 2.6 e Figura 2.7).

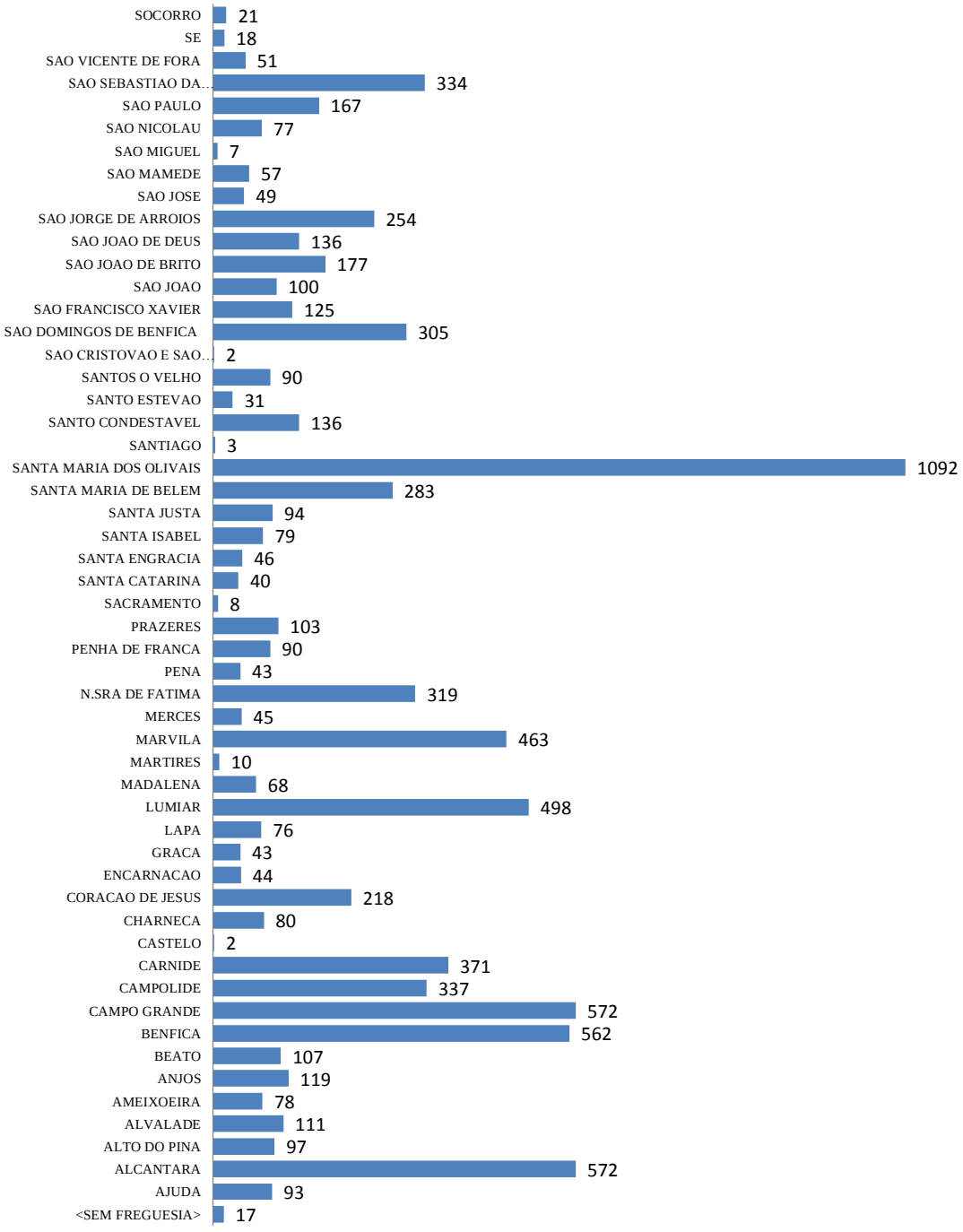


Figura 2.4 Número de acidentes rodoviários com vítimas na cidade de Lisboa entre 2004 e 2007 por freguesia.

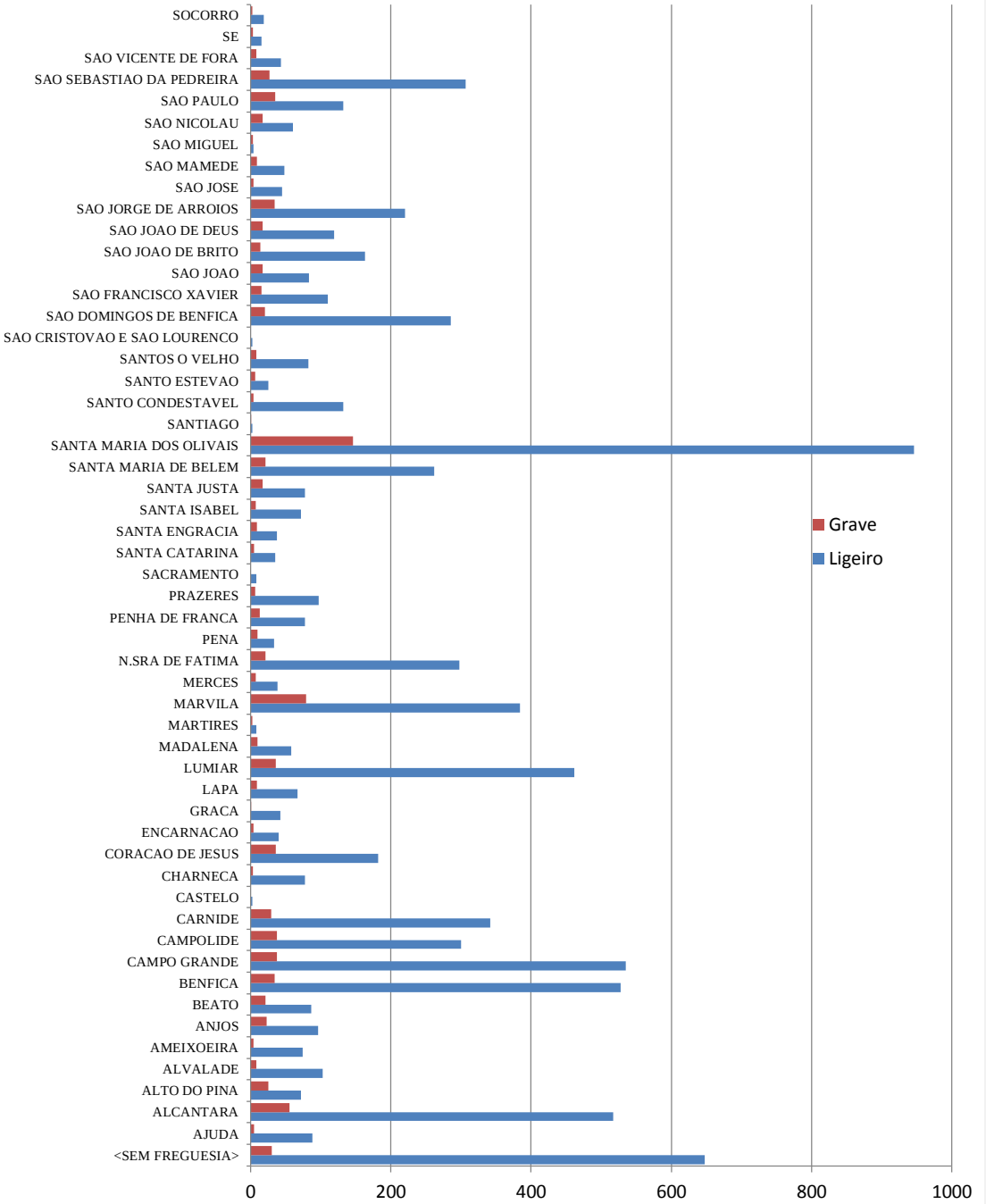


Figura 2.5 Número de acidentes rodoviários com vítimas na cidade de Lisboa entre 2004 e 2007 por freguesia e gravidade do acidente.

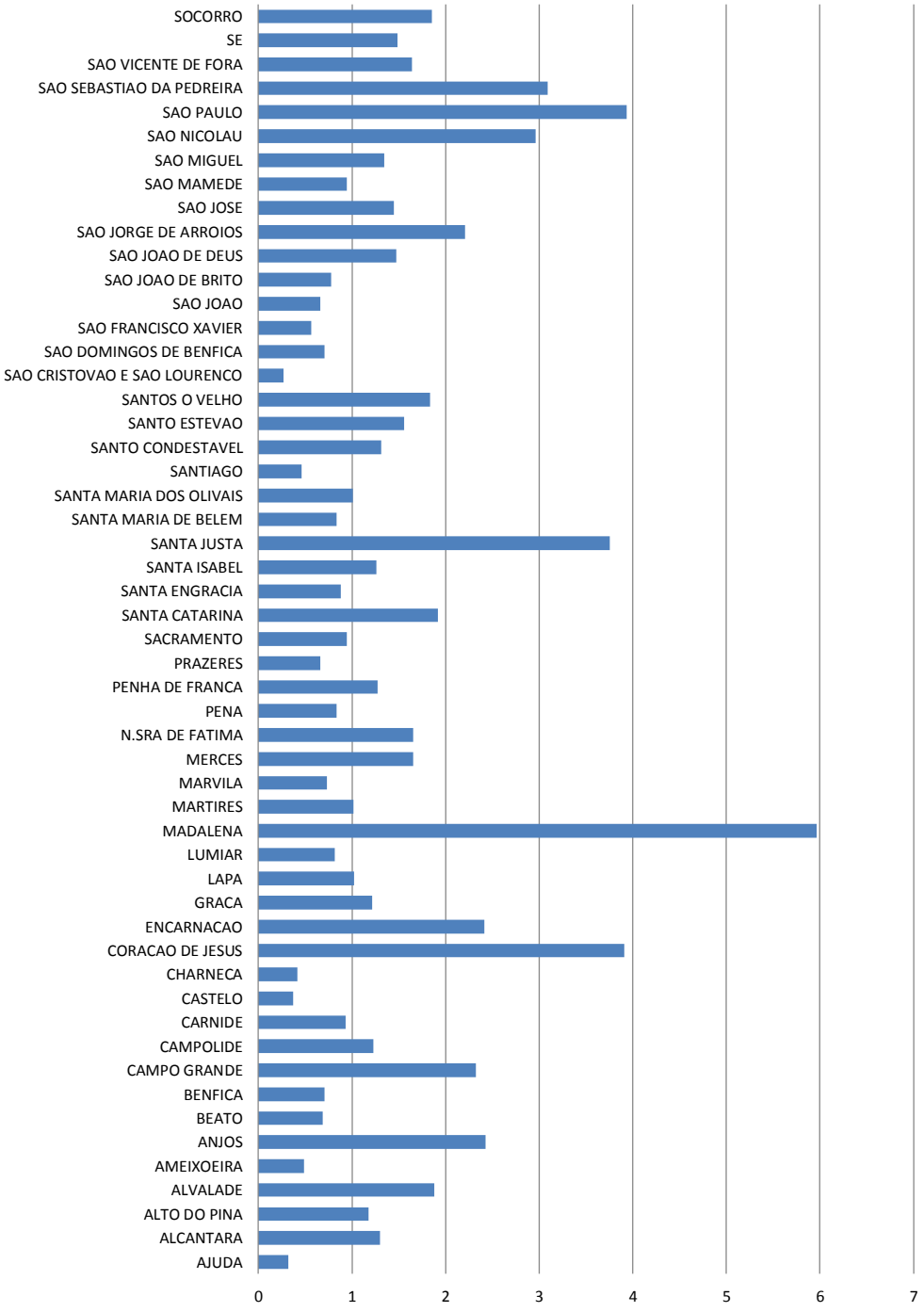


Figura 2.6 Número de acidentes rodoviários com vítimas na cidade de Lisboa entre 2004 e 2007 por freguesia, por 10000 $\boldsymbol{m^2}$.

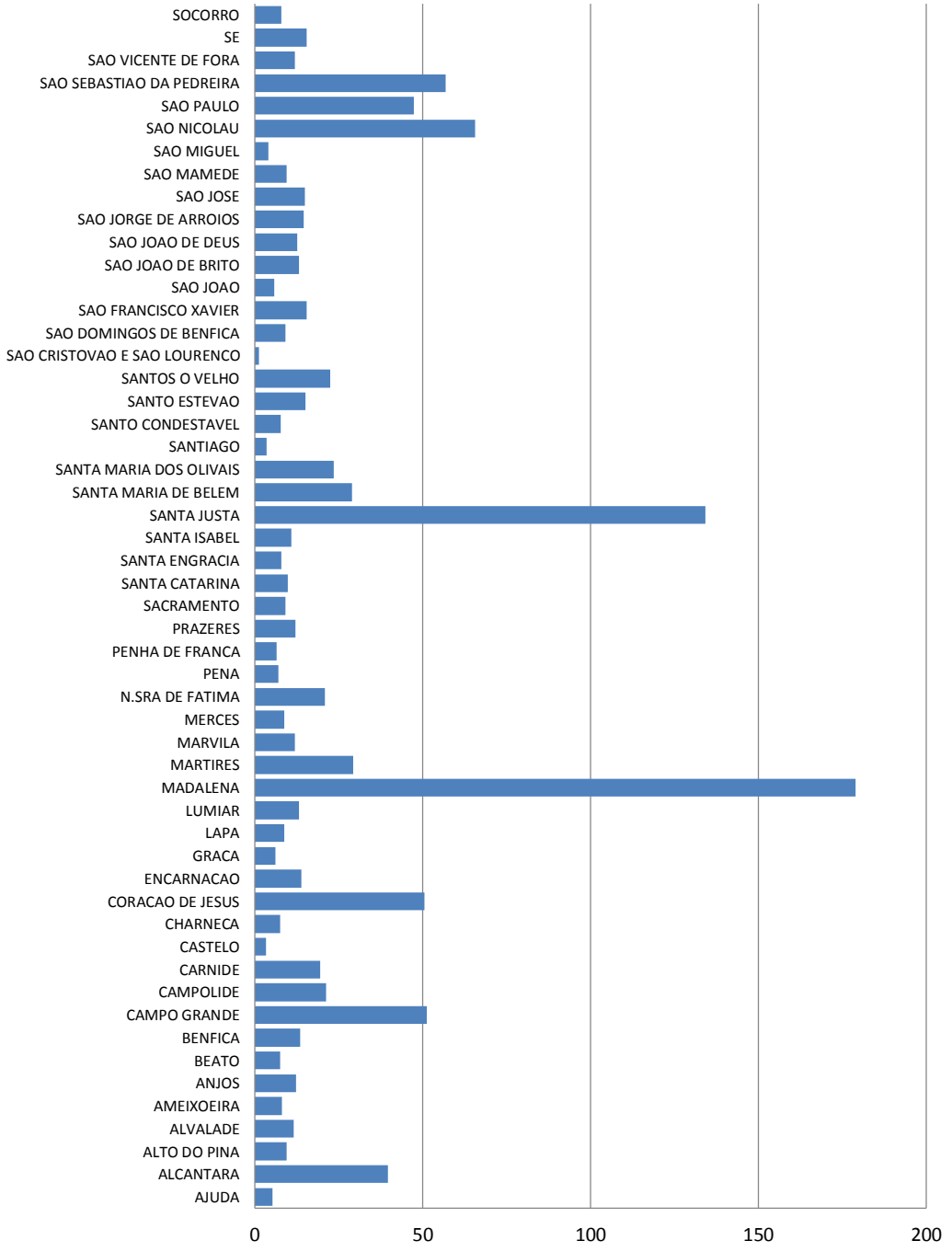


Figura 2.7 Número de acidentes rodoviários com vítimas na cidade de Lisboa entre 2004 e 2007 por freguesia, por 1000 habitantes.

Ao analisar-se o número de acidentes por freguesia por 10000 m^2 de área e por 1000 habitantes, destacam-se as freguesias de Coração de Jesus, Madalena, São Paulo, São Sebastião da Pedreira e Santa Justa com mais de três acidentes por 10000 m^2 de área e as freguesias de Madalena e Santa Justa com mais de 100 acidentes por 1000 habitantes (Figura 2.6 e Figura 2.7).

c) Factores circunstanciais

Nesta secção são analisados alguns factores circunstanciais que podem contribuir para o número de acidentes com vítimas em Lisboa, como factores atmosféricos, de luminosidade ou hora do dia.

- **Factores atmosféricos:**

Relativamente a estes factores, constatou-se que os acidentes ocorrem na sua maioria em estado de bom tempo, em 83% dos casos, seguido de condições de chuva em 13% .

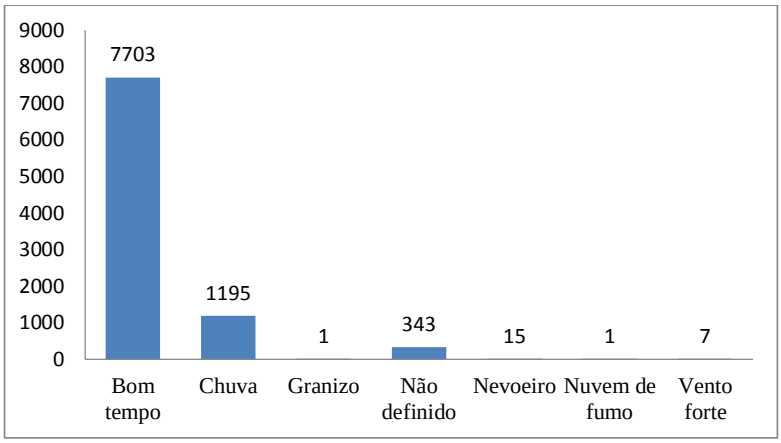


Figura 2.8 Número de acidentes consoante as condições atmosféricas.

Por serem poucas observações relativas às classes de *Nevoeiro*, *Nuvem de fumo* e *Vento forte*, decidiu-se agruparem-se estas classes numa só classe designada *Outros*. Foi ainda criada uma nova classe *Humidade*, que inclui situações de *Chuva* e *Granizo*, visto condicionarem o piso e afins. Assim, o número de acidentes em condições de bom tempo continua a prevalecer com 83%, seguido de condições de mau tempo em 13% dos casos.

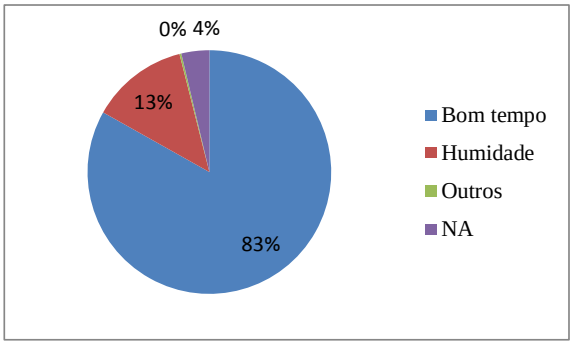


Figura 2.9 Distribuição do número de acidentes consoante condições atmosféricas adaptadas.

Relativamente à gravidade do acidente levando em conta as condições atmosféricas, não há diferenças significativas, como se verifica na Figura 2.10.

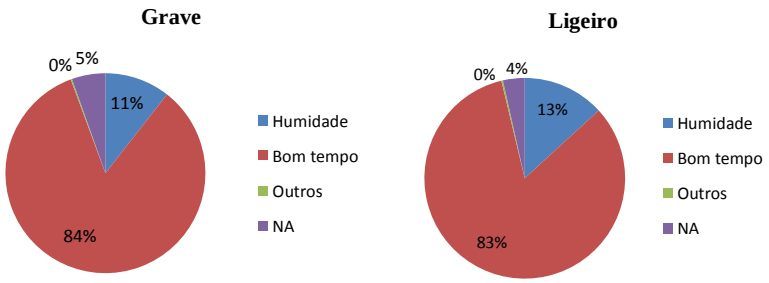


Figura 2.10 Distribuição do número de acidentes consoante condições atmosféricas, para acidentes graves e ligeiros.

- **Luminosidade e hora do dia**

Quanto à luminosidade, verificou-se que a maioria dos acidentes ocorre em pleno dia, portanto, com boa luminosidade, e poucos ocorrem com pouca luminosidade (Figura 2.11). Portanto, a falta de luminosidade não parece ser um factor relevante para a ocorrência de acidentes com vítimas.

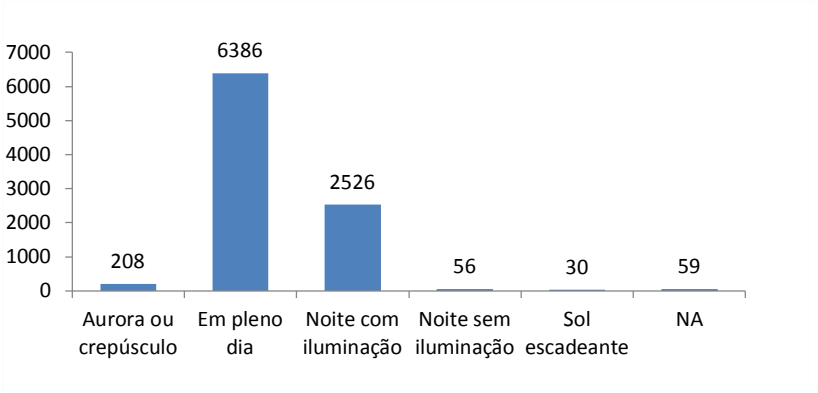


Figura 2.11 Número de acidentes consoante a luminosidade.

Tendo em conta o baixo número de acidentes com *sol escadeante* e de *noite sem iluminação*, agrupou-se os dados de modo a que a categoria *Pleno dia* incorporasse situações de *Sol escadeante* e formou-se uma só classe *Noite* que inclui *Noite com iluminação* ou *Noite sem iluminação.*

Ao analisar-se o número de acidentes em pleno dia, de noite ou em aurora ou crepúsculo tendo em conta as condições meteorológicas, conclui-se que de noite e em aurora ou crepúsculo a percentagem de acidentes em situação de humidade é maior do que em pleno dia, como pode ser verificado na Figura 2.12.

.

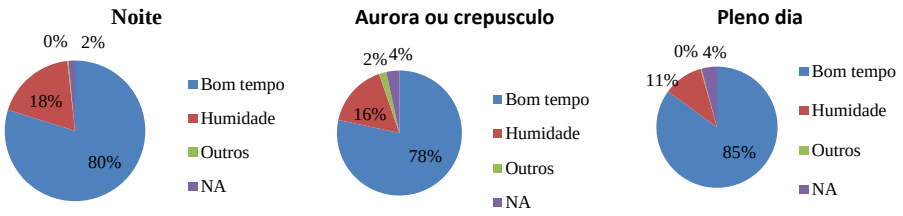


Figura 2.12 Distribuição dos acidentes consoante condições meteorológicas, por luminosidade.

Analisando a gravidade do acidente tendo em conta a luminosidade, verifica-se que de noite a proporção de acidentes graves é maior que a de acidentes ligeiros, ao contrário do que acontece em pleno dia ou em áurora ou crepúsculo (Figura 2.13).

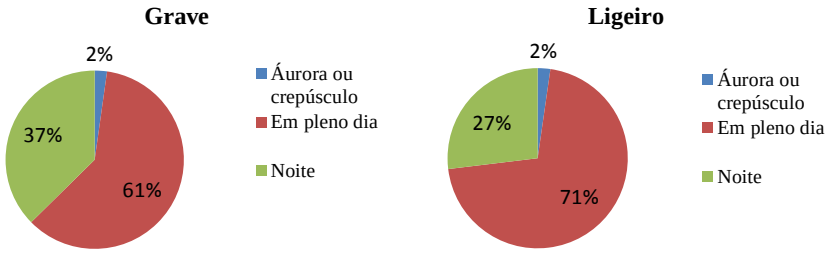


Figura 2.13 Proporção de acidentes graves e ligeiros consoante luminosidade.

Relativamente à hora do dia em que o acidente ocorre, foram consideradas cinco classes (0:00 – 6:59; 7:00 – 10:59; 11:00 – 15:59; 16:00 – 20:59; 21:00 – 23:59) , tendo em conta motivações como a hora de ponta para a sua divisão.

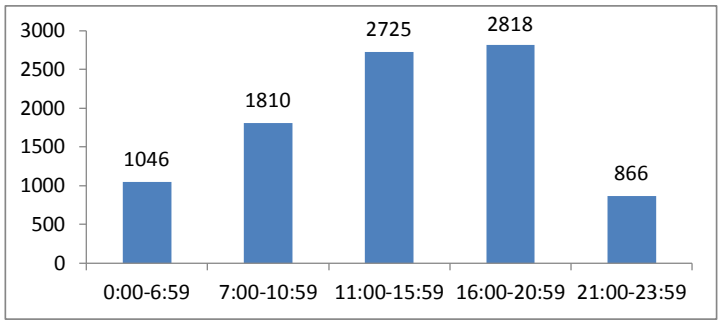


Figura 2.14 Número de acidentes por hora do dia.

Pela análise da Figura 2.14, verifica-se que a maior parte dos acidentes se situa ente as 11:00 e as 21:00.

- **Dias da semana**

Quanto ao número de acidentes em cada dia da semana, este mantém-se relativamente constante, apesar de uma pequena descida aos dias de fim-de-semana. Esta descida poderá ser explicada pela menor afluência de tráfego nesses dias, tal como nas análises das variáveis anteriores.

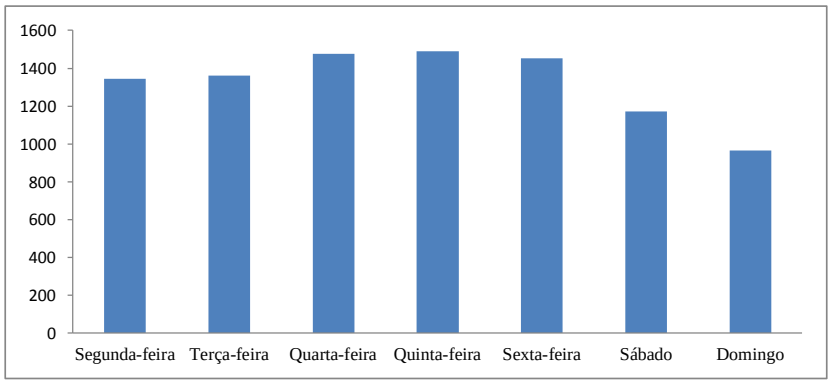


Figura 2.15 Número de acidentes por dia da semana.

O gráfico da Figura 2.16 sugere, ainda, que durante os dias de fim-de-semana (Sábado e Domingo) existe uma percentagem muito mais significativa de acidentes à noite comparando com os restantes dias.

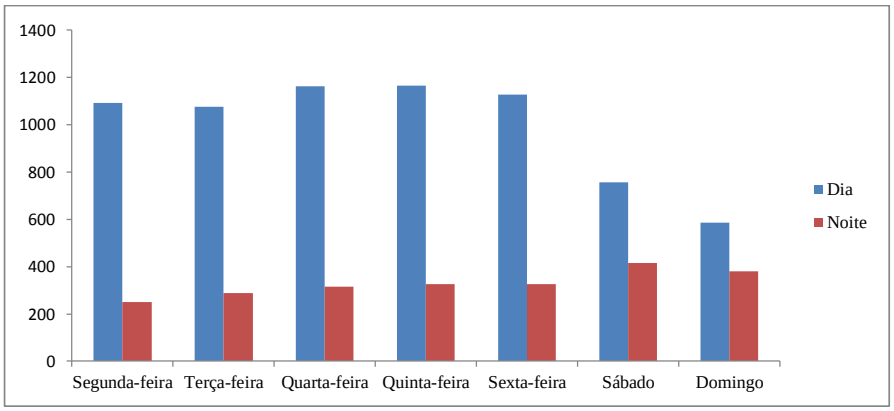


Figura 2.16 Número de acidentes por dia da semana, consoante seja Dia ou Noite.

d) Características da via

Esta secção diz respeito à análise de factores relacionados com as características urbanas existentes na situação do acidente, como as características do pavimento, da via, a existência de sinalização, de obras de arte como túneis ou pontes e o regime de circulação afecto.

Por serem factores muito específicos e por estarem, muitas vezes, sem preenchimento, foi feita uma análise mais breve, mas não menos informativa, dos mesmos. Deste modo, apresentam-se abaixo as tabelas correspondentes aos factores relacionados com a via na situação do acidente.

Circulação	
Dois sentidos	6901
Sentido único	2265
NA	137

Condições de aderência	
Seco e limpo	7692
Outros	1524
NA	47

Marcas da via	
Com marcas separadoras de sentido de trânsito	2608
Com marcas separadoras de sentido e de vias de trânsito	4224
Sem marcas rodoviárias ou pouco visíveis	2248
NA	183

Características técnicas da via	
Estrada com separador Auto-estrada	14
Estrada com separador Outra via	3104
Estrada sem separador	3094
NA	3051

Intersecção de vias	
Nivelada	3817
Desnivelada	126
Fora da intersecção	2537
NA	2783

Obras de arte	
Viaduto -Ponte	205
Túnel	125
Passagem estreita	20
NA	8913

Obstáculos	
Correctamente sinalizados	267
Insuficientemente sinalizados	46
Não sinalizados	29
Inexistentes	8675
NA	246

Sinais	
Cedência de passagem	426
Stop	122
Passagem de peões	1053
Proibição ultrapassagem	9
Outros	2247
NA	5406

Sinais luminosos	
A funcionar normalmente	2855
Falha/Intermitente	104
Inexistentes	6164

Tipo de piso	
Betuminoso	8752
Betão cimento	58
Calçada	401
Terra batida	20
NA	32

Estado de conservação	
E m bom estado	6554
Em estado regular	2532
Em mau estado	116
NA	61

Traçado	
Recta	7266
Curva	1792
NA	205

Perfil da via	
Com inclinação ou lomba	3041
Em patamar	6068
NA	154

Traçado da via	
Berma pavimentada	6040
Berma não pavimentada	559
Sem berma ou impraticável	2014
NA	650

Situação do acidente	
Em plena via	8726
Outros	398
NA	139

Via de transito	
Central	1247
Direita	4556
Esquerda	2250
NA	1210

Sentidos	
Crescente quilometragem	1309
Decrescente quilometragem	1254
NA	6700

Destas variáveis seleccionaram-se as que se consideram mais relevantes para o estudo e que não apresentem um número elevado de valores omissos (*NA*), de modo a obter-se uma análise mais coerente. Assim, serão analisadas graficamente apenas as variáveis *condições de aderência*, *sinais luminosos* e *estado de conservação.*

Relativamente às condições de aderência, verifica-se que para piso seco e limpo a proporção de acidentes graves é maior que a de acidentes ligeiros, ao contrário do que se verifica para aderência de piso com humidade ou outros (Figura 2.17). É, também, em condições de aderência com piso seco e limpo que ocorre a maioria dos acidentes com vítimas, na sua totalidade.

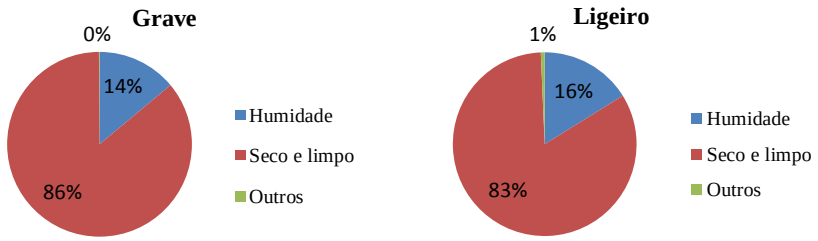


Figura 2.17 Proporção de acidentes graves e ligeiros dependendo das condições de aderência do piso.

Quanto ao funcionamento dos sinais luminosos, verfica-se que há uma maior percentagem de acidentes graves quando os sinais estão a funcionar normalmente do que acidentes ligeiros. No entanto, caso estes sejam inexistentes, há mais acidentes ligeiros que graves (Figura 2.18). Esta categoria é, também, a predominante no número total de acidentes.

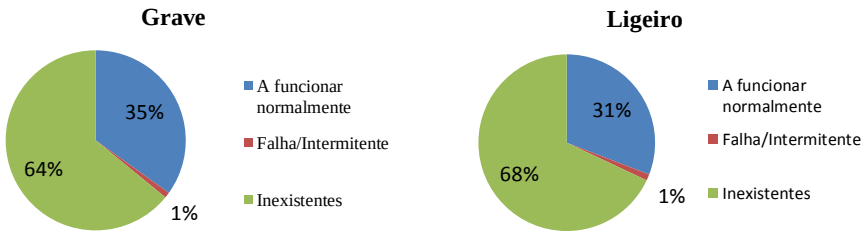


Figura 2.18 Funcionamento dos sinais luminosos consoante gravidade do acidente.

Quanto ao estado de conservação da via, os acidentes ocorrem, na sua maioria, em vias em bom estado, com um ligeiro aumento na gravidade para esta categoria. Em estado regular, ocorrem menos acidentes graves que ligeiros, apesar dessa diferença ser pouco significativa (Figura 2.19).

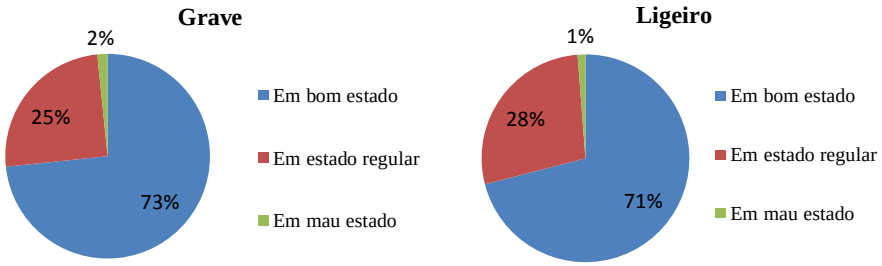


Figura 2.19 Estado de conservação da via em função da gravidade do acidente.

e) Factores humanos

Tendo em conta que a presente análise se refere a acidente com vítimas, analisou-se, deste modo, a classificação das mesmas e as características relativas a cada uma delas, consoante tratar-se do condutor, passageiro do veículo ou peões vítimas do acidente. Uma primeira análise mostra que o número de vítimas que são feridos ligeiros é consideravelmente maior do que feridos graves ou mortes (Figura 2.20).

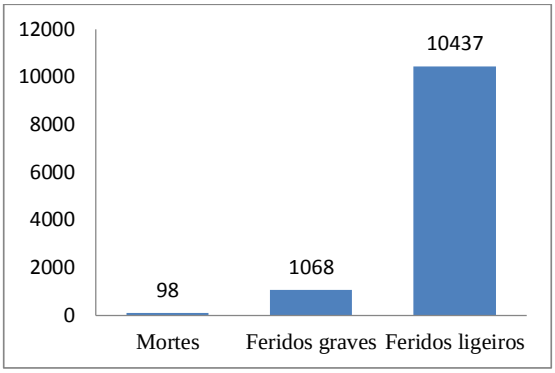


Figura 2.20 Número de acidentes por lesões causadas nas vítimas.

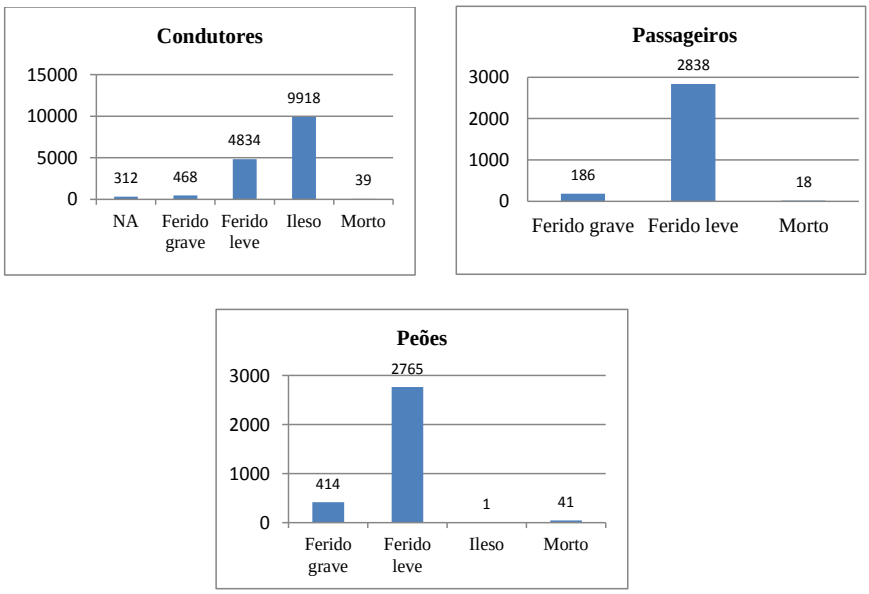


Figura 2.21 Número de acidentes por tipo de lesão, para condutores, passageiros e peões.

Reflectindo a Figura 2.21 na distinção de vítimas constatou-se que a maior parte dos condutores sai ileso do acidente ou como ferido ligeiro. Quanto aos passageiros, a maior parte das vítimas são feridos leves, o mesmo acontecendo com os peões. Relativamente ao sexo e idade das vítimas (Figura 2.22 e

Figura 2.23), verifica-se que de entre os condutores há predominância do sexo masculino e das idades compreendidas entre os 30 e os 40 anos e, também, entre os 20 e 25 anos de idade. Os passageiros vítimas são, também, maioritariamente homens com idades entre os 20 e os 30 anos de idade. Quanto aos peões apenas se tem informação da idade, constatando-se ser relativamente uniforme, embora sensivelmente menor para peões com idades inferiores a 10 anos e superiores a 90 anos de idade, devendo-se, possivelmente, ao facto de haver menos afluência destes nas ruas.

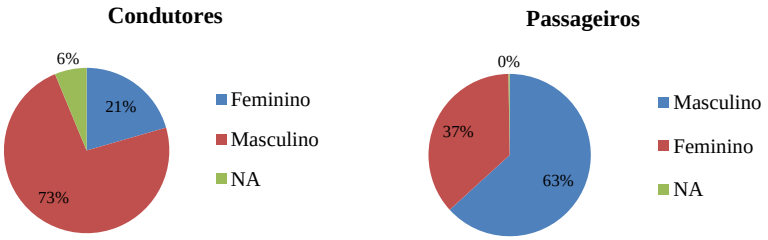


Figura 2.22 Distribuição do número de acidentes consoante o sexo, para condutores e passageiros.

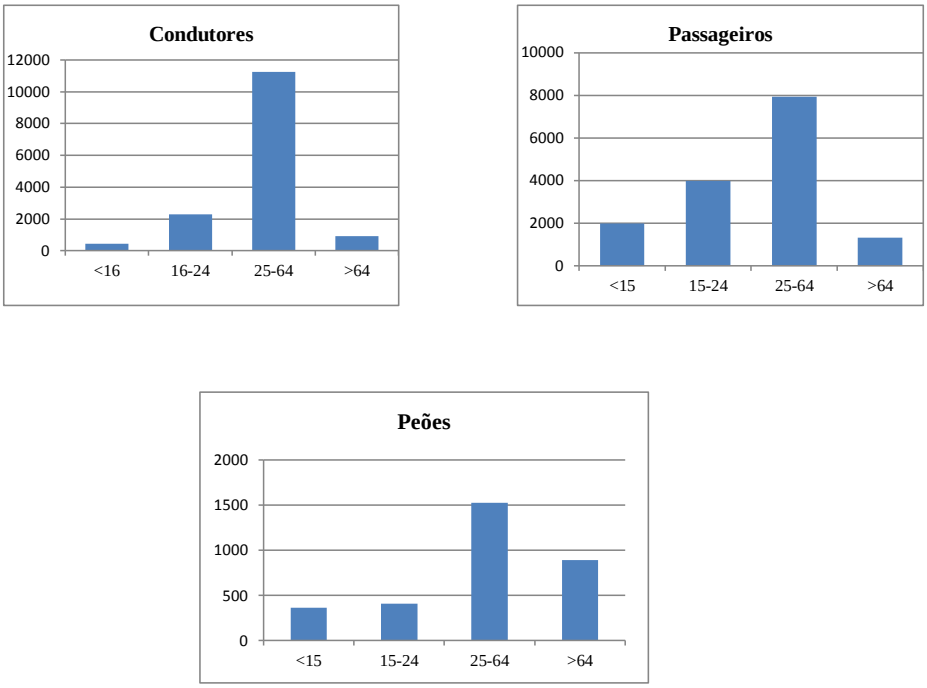


Figura 2.23 Distribuição das vítimas por idade, em número.

Considerando agora a densidade populacional na cidade de Lisboa distribuída por idades, obtem-se o seguinte gráfico da percentagem da população que são peões vítimas de acidentes rodoviários, consoante idade (Figura 2.24).

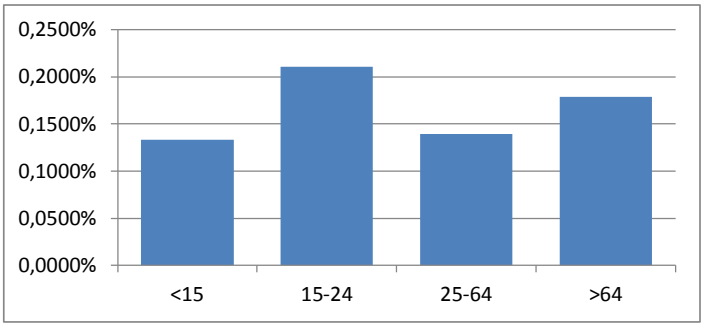


Figura 2.24 Distribuição de peões vítimas na população de Lisboa, por idade.

Note-se que anteriormente verificava-se que a principal faixa etária de peões vítimas de acidentes rodoviários em Lisboa era a dos 70 aos 80 anos. No entanto, com esta *padronização* dos dados constata-se que é dos 15 aos 24 anos que existem mais peões vítimas. Esta padronização é feita assumindo que todas as faixas etárias na população estão igualmente representadas na população dos peões.

Em relação à gravidade do acidente tendo em conta a idade do peão (Figura 2.25), verifica-se um aumento da percentagem de peões com lesões graves para a faixa dos menores de 15 anos e para a dos maiores de 64 anos. Isto deve-se, provavelmente, ao facto de serem mais vulneráveis. As restantes faixas etárias não apresentam grandes diferenças.

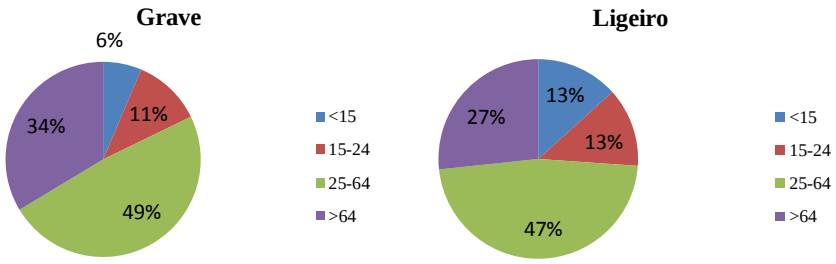


Figura 2.25 Distribuição da idade dos peões por gravidade do acidente.

Quanto às restantes variáveis subjacentes às vítimas dos acidentes, verifica-se que a grande maioria dos condutores e dos passageiros se encontravam seguros com cinto de segurança ou capacete (Figura 2.26).

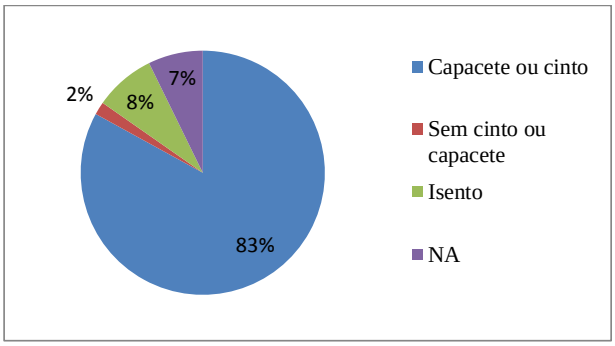


Figura 2.26 Distribuição da utilização dos acessórios.

A maior parte dos condutores tem licença de carta adequada ao veículo inerente (Figura 2.27), circulavam em marcha normal e foram submetidos ao teste de alcoolémia. Os peões, na sua maioria, atravessavam em passagem sinalizada ou a 50m da mesma e encontravam-se isolados.

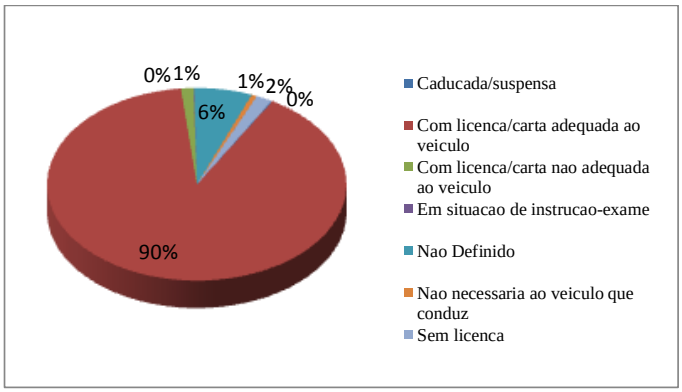


Figura 2.27 Distribuição da situação da licença dos condutores.

Quanto aos anos de carta dos condutores, a maior parte tem carta há mais de dois anos. A percentagem de condutores com carta há menos de dois anos é maior nos acidentes graves que nos ligeiros.

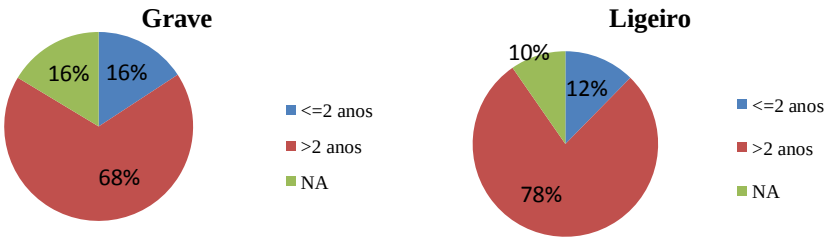


Figura 2.28 Distribuição dos anos de carta dos condutores, por gravidade do acidente.

Pela análise da Figura 2.23 e relativamente aos condutores, verifica-se que há condutores com menos de 18 anos de idade, referindo-se aos condutores de velocípedes, motociclos ou ciclomotores. Fazendo uma análise mais detalhada desse facto, obteve-se o gráfico da distribuição de idades para condutores desse tipo de viaturas (Figura 2.29).

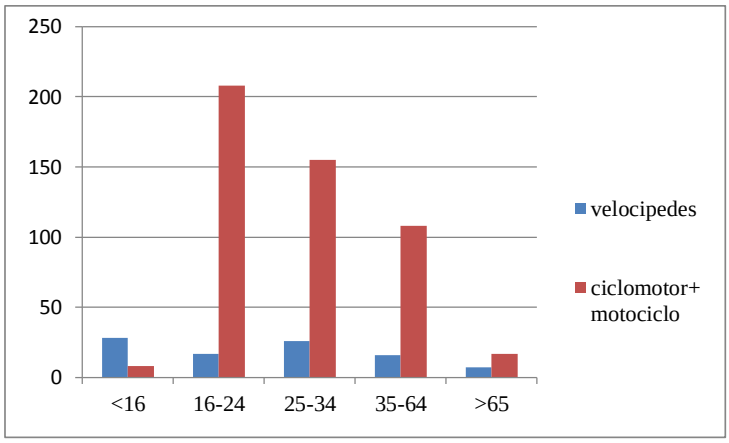


Figura 2.29 Distribuição das idades dos condutores de veículos de duas rodas.

Conclui-se, assim, que é a faixa dos 16 aos 24 anos que apresenta mais vítimas condutores de ciclomotores e motociclos, e na faixa dos menores de 16 anos relativamente a velocípedes.

Quanto à gravidade dos acidentes relativamente a estes veículos em detrimento dos restantes, não há alterações significativas, apenas mais 2% dos acidentes graves são com estes tipos de veículos do que com outros (Figura 2.30).

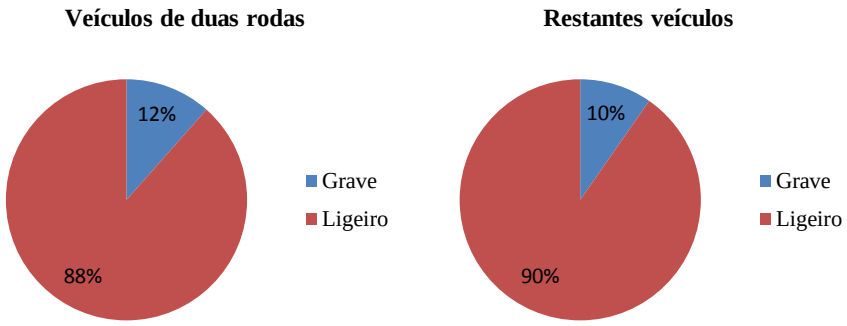


Figura 2.30 Distribuição da gravidade dos acidentes por tipo de veículo (veículos de duas rodas e restantes).

Quanto à idade do veículo, verifica-se que a proporção de acidentes entre veículos com menos de cinco anos e entre cinco e vinte anos, não apresenta grande diferença (Figura 2.31).

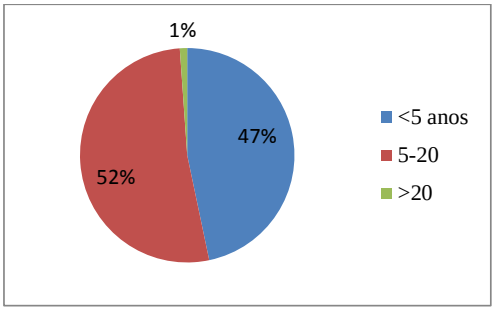


Figura 2.31 Distribuição dos acidentes por idade do veículo.

2.2 Principais conclusões

A análise exploratória dos dados efectuada nesta secção permitiu conhecer as diversas variáveis consideradas na ocorrência de acidentes rodoviários com vítimas em Lisboa e correspondentes implicações na gravidade e número de acidentes.

Verifica-se que há factores que têm mais impacto na gravidade que outros, como é o caso, por exemplo, da natureza dos acidentes ou da luminosidade na hora do acidente. Os factores atmosféricos, no entanto, mostram-se pouco relevantes, visto que a maioria dos acidentes ocorrer em condições de bom tempo poderá ser irrealista, pois na maior parte dos dias as condições meteorológicas são boas. Verifica-se, ainda, que com humidade e de noite, a gravidade dos acidentes é maior comparativamente a outras condições. Quanto às condições de aderência do piso e ao funcionamento dos sinais luminosos, há uma maior proporção de acidentes graves quando o piso está seco e limpo e quando os sinais luminosos funcionam normalmente. O estado de conservação da via aquando da ocorrência dos acidentes é, normalmente, boa ou regular.

Quando há peões envolvidos, a idade dos mesmos também se mostrou relevante, na medida em que há uma maior proporção de crianças e idosos com ferimentos graves ou morte, do que nas outras faixas etárias.

Verifica-se, também, que há mais homens envolvidos em acidentes do que mulheres e esses encontram-se, prioritariamente, na faixa etária dos 20 aos 30 anos. Os anos de carta do condutor têm também relevância, na medida em que a obtenção de carta de condução há menos de dois anos resulta numa maior percentagem de acidentes graves relativamente a acidentes ligeiros.

Na maioria dos casos, os condutores dos veículos encontravam-se protegidos com cinto de segurança e tinham licença ou carta adequada ao veículo. A idade do veículo não parece significativa, visto que os acidentes ocorrem basicamente com a mesma frequência para veículos com mais ou com menos de 5 anos.

Esta breve análise exploratória foi o primeiro passo para uma análise mais focada nos factores associados aos acidentes rodoviários, num estudo dos efeitos dos mesmos através de um modelo de regressão apropriado e, também, para uma análise espacial da localização das ocorrências dos acidentes.

3 Análise de regressão

A análise de regressão é uma das técnicas mais utilizadas na Estatística, com aplicações em diversas áreas. O principal objectivo desta análise é desenvolver um modelo que seja capaz de prever valores de uma variável aleatória dependente Y em função de uma ou mais variáveis independentes x, através de uma relação que pode ou não ser linear.

Usualmente, a variável dependente Y é designada de variável resposta, enquanto as variáveis x se designam de variáveis independentes, variáveis explicativas ou covariáveis.

Os primeiros estudos deste método, levados a cabo por Gauss e Legendre (Nelder & McCullagh, 1989) basearam-se num modelo linear, onde a relação entre as variáveis dependente e as independentes é linear e a variável resposta é de natureza contínua.

Considere-se uma amostra aleatória de n unidades, para cada observação tem-se um par de valores das variáveis aleatórias Y e $\boldsymbol{x}$ denotado por $(y_i, \boldsymbol{x}_i), i = 1, \dots, n$, independentes. No modelo de regressão linear clássico tem-se a seguinte relação:

$$Y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_k x_{ik} + \epsilon_i \,, \qquad i = 1, \dots, n \tag{3.1}$$

onde $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)'$ é o vector de coeficientes de regressão, que se assume serem fixos e desconhecidos e ϵ_i é uma variável aleatória designada de erro aleatório.

Para estes modelos considera-se a variável resposta Y de natureza contínua e as variáveis explicativas x de natureza qualitativa dicotómica ou ordinal, discreta ou contínua. Supõe-se, ainda, que os erros aleatórios são variáveis aleatórias independentes e $E[\epsilon_i] = 0$ e $Var[\varepsilon_i] = \sigma^2$. Desta forma, a equação acima implica que

$$\mu_i = \boldsymbol{E}[Y_i|\boldsymbol{x}_i] = \beta_0 + \beta_1 . x_{i1} + \cdots + \beta_k . x_{ik} \,, \qquad i = 1, \dots, n. \tag{3.2}$$

Escrevendo a equação acima na forma matricial tem-se

$$\boldsymbol{Y} = \begin{pmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{pmatrix} \qquad \boldsymbol{X} = \begin{pmatrix} 1 & x_{11} & \cdots & x_{1k} \\ 1 & x_{21} & \ddots & \vdots \\ \vdots & \vdots & & \\ 1 & x_{n1} & \cdots & x_{nk} \end{pmatrix} \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \quad \boldsymbol{\varepsilon} = \begin{pmatrix} \varepsilon_0 \\ \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}$$

e

$$\boldsymbol{Y} = \boldsymbol{\beta X} + \boldsymbol{\varepsilon}.$$

Usualmente assume-se que os erros $\boldsymbol{\varepsilon}$ seguem uma distribuição normal. Este pressuposto permite obter uma distribuição amostral para os estimadores dos coeficientes de regressão e assim construir testes de hipótese e intervalos de confiança para eles.

Apesar das inúmeras aplicações deste modelo de regressão linear clássico, existem situações em que esses modelos não representam uma explicação plausível da realidade, nomeadamente quando a variável resposta em estudo não é de natureza contínua ou normal. Tal como referem Nelder e MacCullagh (1989), situações envolvendo ensaios de diluição, proporções, dados de contagens ou análise de sobrevivência são explicadas mais adequadamente por modelos não lineares e/ou não normais, como os modelos complementar log-log, probit, logit ou os modelos log-lineares. Estes modelos são casos particulares dos modelos lineares generalizados, abreviadamente referidos por GLM.

3.1 Modelos Lineares Generalizados

Os modelos lineares generalizados são uma extensão dos modelos lineares clássicos, incluindo inúmeros outros modelos bastante úteis na análise estatística. Estes têm, ainda, uma estrutura de linearidade na média e consistem basicamente em três componentes:

- Uma distribuição de probabilidade da variável resposta na família exponencial;

- Um predictor linear $\eta = \sum_{i=1}^{k} z_i \beta_i$;

- Uma função de ligação g tal que $\eta = g(\mu)$.

No caso do modelo linear clássico, a função de ligação é a identidade e a distribuição de probabilidade associada é a normal.

3.1.1 A Família Exponencial

Como foi referido anteriormente, nos modelos lineares generalizados a variável resposta Y tem distribuição pertencente à família exponencial.

Definição 3.1 Seja Y uma variável aleatória com distribuição pertencente à família exponencial. Então, a sua função massa de probabilidade pode escrever-se na forma

$$f(y|\theta,\phi) = \exp\left\{\frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi)\right\}, \tag{3.3}$$

onde θ *e* ϕ são parâmetros escalares e $a(.), b(.)$ e $c(.,.)$ são funções reais conhecidas.

O parâmetro θ é denominado parâmetro canónico e caracteriza a distribuição e ϕ é estritamente positivo e designado de parâmetro de escala. Admite-se, ainda, que a função $b(.)$ é diferenciável e o suporte da distribuição não está dependente dos parâmetros e, nesta situação, a família exponencial pertence à classe de famílias regulares.

Considere-se a função log-verosimilhança dada por $\ell(\theta;\phi,y) = \ln\big(f(y|\theta,\phi)\big)$. A *função score* é definida por

$$S(\theta) = \frac{\partial \ell(\theta;\phi,Y)}{\partial \theta}. \tag{3.4}$$

Para a família exponencial tem-se

$$\ell(\theta;\phi,y) = \frac{y\theta - b(\theta)}{a(\phi)} + c(y,\phi), \tag{3.5}$$

portanto,

$$S(\theta) = \frac{Y - b'(\theta)}{a(\phi)} \quad e \quad \frac{\partial S(\theta)}{\partial \theta} = -\frac{b''(\theta)}{a(\phi)},$$

onde $b'(\theta)$ e $b''(\theta)$ são a primeira e segunda derivadas de $b(\theta)$, respectivamente.

Sob as condições de regularidade, sabe-se que

$$E[S(\theta)] = 0 \quad e \quad E[S^2(\theta)] = E\left[\left(\frac{\partial \ell(\theta;\phi,Y)}{\partial \theta}\right)^2\right] = -E\left[\frac{\partial^2 \ell(\theta;\phi,Y)}{\partial \theta^2}\right],$$

donde o valor esperado e a variância de Y vêm dados por

$$\mu = E[Y] = b'(\theta), \tag{3.6}$$

$$Var[Y] = a^2(\phi)Var[S(\theta)] = a^2(\phi)\frac{b''^{(\theta)}}{a(\phi)} = a(\phi)b''(\theta). \quad (3.7)$$

À função $b''(\theta)$ dá-se o nome de *função de variância* e depende apenas do parâmetro de localização, sendo, geralmente, representada por $V(\mu)$. A função $a(\phi)$, que depende apenas do parâmetro de escala, é, na maior parte dos modelos, da forma

$$a(\phi) = \frac{\phi}{\omega},$$

com ω uma constante conhecida. Desta forma, a expressão da variância simplifica-se tomando a forma

$$Var[Y] = b''(\theta)\frac{\phi}{\omega}.$$

Se o parâmetro ϕ é conhecido, então representa a família exponencial uniparamétrica indexada por θ.

Deste modo, o modelo pode ser reescrito como

$$f(y|\theta, \phi, \omega) = \exp\left\{\frac{\omega}{\phi}(y\theta - b(\theta)) + c(y, \phi, \omega)\right\}. \quad (3.8)$$

Exemplo:

Considere-se a função distribuição de uma $Binomial(n, \pi)$ tal que

$$f(y) = \binom{n}{y}\pi^y(1-\pi)^{n-y}, \quad y \in \{0,1,\dots,n\}. \quad (3.9)$$

Equivalemente, a equação (3.9) pode ser reescrita como

$$f(y) = \binom{n}{y} + \left[\exp(y)\log\left(\frac{\pi}{1-\pi}\right) + n\log(1-\pi)\right]. \quad (3.10)$$

Logo, a distribuição binomial pertence à família exponencial com parâmetro canónico

$$\theta = \log\left(\frac{\pi}{1-\pi}\right).$$

3.2 As componentes dos GLM

Como já foi referido anteriormente, os modelos lineares generalizados são uma extensão do modelo linear clássico e a sua estrutura é formada por duas partes: uma componente aleatória e uma componente sistemática e função de ligação.

- *Componente aleatória*

Esta componente estipula que a variável aleatória resposta Y segue uma distribuição pertencente à família exponencial, independente sobre observações e tal que

$$E[Y_i|\boldsymbol{x}_i] = \mu_i = b'(\theta_i), \qquad i = 1, \dots, n \tag{3.11}$$

em que $\boldsymbol{x}_i$ é o vector de covariáveis.

- *Componente sistemática e função de ligação*

A componente sistemática dos GLM, também designada de preditor linear, é uma função linear dos parâmetros desconhecidos $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \dots, \beta_k)'$ representada por $\eta_i = \boldsymbol{z}_i^T\boldsymbol{\beta}$, onde $\boldsymbol{z}_i$ é um vector de especificação de dimensão $k+1$, função do vector de covariáveis $\boldsymbol{x}_i$, $\boldsymbol{z}_i = (1, x_{i1}, x_{i2}, \dots, x_{ik})'$.

Outra característica desta componente é a relação do valor esperado μ com o preditor linear η através de

$$\mu_i = h(\eta_i) = h\left(\boldsymbol{z}_i^T \boldsymbol{\beta}\right), \tag{3.12}$$

onde h é uma função monótona e diferenciável e tal que $g = h^{-1}$ é a designada *função de ligação.*

Quando o preditor linear coincide com o parâmetro canónico, isto é, $\theta_i = \eta_i$, então diz-se que a função de ligação é a *função de ligação canónica.*

No modelo de regressão linear clássico, a função de ligação é a identidade ($\eta = \mu$). A escolha da função de ligação depende do problema a ser estudado e da variável resposta. Por exemplo, para o modelo binomial, as três principais funções de ligação usadas são:

1. logit (ou logística): $g(\mu) = \ln\left(\frac{\mu}{1-\mu}\right)$;

2. probit: $g(\mu) = \Phi^{-1}(\mu)$;

3. complementar log-log : $g(\mu) = \ln[-\ln(1-\mu)]$.

Já para o modelo Poisson, a função de ligação usual é a função logarítmica dada por $g(\mu) = \ln(\mu)$.

3.3 Estimação dos parâmetros

Depois de escolhido o modelo adequado ao problema em estudo, é necessário realizar inferências sobre o mesmo. Os modelos lineares generalizados baseiam essa inferência na metodologia de máxima verosimilhança.

As estimativas dos parâmetros dos GLM concentram-se, essencialmente, na estimação do parâmetro $\boldsymbol{\beta}$, estimado pelo método da máxima verosimilhança (Wedderburn & Nelder, 1972). Já o parâmetro ϕ, não sendo conhecido, é estimado pelo método dos momentos.

As estimativas do parâmetro $\boldsymbol{\beta}$ podem, portanto, ser obtidas pela maximização da verosimilhança, ou log-verosimilhança, supondo fixos os dados observados. No entanto, as equações de máxima verosimilhança não têm, em geral, uma solução analítica, sendo necessário resolvê-las através de métodos numéricos. Para os GLM, Nelder e Wedderburn (1972) construíram um algoritmo para resolver tais equações, o que proporcionou o sucesso destes modelos, tendo em conta que se usa o

mesmo algoritmo para os diversos modelos. Este é designado de *método iterativo de mínimos quadrados ponderados* e baseia-se no método de *scores de Fisher* (secção 3.3.1).

3.3.1 Método dos scores de Fisher

Tal como formulado anteriormente, considere-se y_i a observação da variável resposta para a i-ésima unidade experimental, $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \dots, x_{ik})'$ o correspondente vector das observações das variáveis explicativas (ou covariáveis). Tem-se, ainda, o vector de especificação $\boldsymbol{z}_i$ de dimensão *k+1* , já definido anteriormente.

Considere-se um modelo linear generalizado definido por

$$f(y_i|\theta_i, \phi, \omega_i) = \exp\left\{\frac{\omega_i}{\phi}\left(y_i\theta_i - b(\theta_i)\right) + c(y_i, \phi, \omega_i)\right\} \tag{3.13}$$

com função de ligação $g(\mu_i) = \boldsymbol{z}_i^T\boldsymbol{\beta}$, com y_i variáveis aleatórias independentes.

A função de verosimilhança, dependendo do parâmetro que se quer estimar, $\boldsymbol{\beta}$, é dada por

$$L(\boldsymbol{\beta}) = \prod_{i=1}^{n} f(y_i|\theta_i, \phi, \omega_i) = \prod_{\mathrm{i}=1}^{\mathrm{n}} \exp\left\{\frac{\omega_i}{\phi}\left(y_i\theta_i - b(\theta_i)\right) + c(y_i, \phi, \omega_i)\right\}$$

$$= \exp\left\{\frac{1}{\phi}\sum_{i=1}^{n}\omega_i\left(y_i\theta_i - b(\theta_i)\right) + \sum_{i=1}^{n} c(y_i, \phi, \omega_i)\right\}. \tag{3.14}$$

Desta forma, a função log-verosimilhança vem dada por

$$\ell(\boldsymbol{\beta}) = \ln L(\boldsymbol{\beta}) = \sum_{i=1}^{n}\left(\frac{\omega_i\left(y_i\theta_i - b(\theta_i)\right)}{\phi} + c(y_i, \phi, \omega_i)\right) = \sum_{i=1}^{n}\ell_i(\boldsymbol{\beta}). \tag{3.15}$$

Assim, os estimadores de máxima verosimilhança para $\boldsymbol{\beta}$ são as soluções do sistema de equações de verosimilhança

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{i=1}^{n} \frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j} = 0, \quad j = 1, \dots, k+1$$

Pela regra da cadeia tem-se

$$\frac{\partial \ell_i(\boldsymbol{\beta})}{\partial \beta_j} = \frac{\partial \ell_i(\theta_i)}{\partial \theta_i} \frac{\partial \theta_i(\mu_i)}{\partial \mu_i} \frac{\partial \mu_i(\eta_i)}{\partial \eta_i} \frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_j}$$

e

$$\frac{\partial \ell_i(\theta_i)}{\partial \theta_i} = \frac{\omega_i(y_i - b'(\theta_i)}{\phi} = \frac{\omega_i(y_i - \mu_i)}{\phi}$$
$$\frac{\partial \mu_i}{\partial \theta_i} = b''(\theta_i) = \frac{\omega_i Var(Y_i)}{\phi},$$
$$\frac{\partial \eta_i(\boldsymbol{\beta})}{\partial \beta_j} = z_{ij}.$$

Deste modo, tem-se que as equações de verosimilhança para $\boldsymbol{\beta}$ são

$$\sum_{i=1}^{n} \frac{(y_i - \mu_i) z_{ij}}{Var[Y_i]} \frac{\partial \mu_i}{\partial \eta_i} = 0, \quad j = 1, \dots, k+1 \tag{3.16}$$

Para a implementação do método de scores de *Fisher*, uma generalização do método de Newton-Raphson, considera-se o processo iterativo

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \widehat{\boldsymbol{\beta}}^{(k)} + [\mathfrak{T}(\widehat{\boldsymbol{\beta}}^{(k)})]^{-1} S\big(\widehat{\boldsymbol{\beta}}^{(k)}\big), \tag{3.17}$$

onde $\mathfrak{T}(.)^{-1}$ é a inversa da matriz de informação de *Fisher* e $S(.)$ o vector de *scores*, ambos calculados para $\boldsymbol{\beta} = \widehat{\boldsymbol{\beta}}^{(k)}$.

O vector dos *scores* é um vector de dimensão $k+1$ dado por

$$S(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} = \left\{ \sum_{i=1}^{n} \frac{(y_i - \mu_i) z_{ij}}{Var[Y_i]} \frac{\partial \mu_i}{\partial \eta_i} \right\}, \qquad j = 1, \dots, k+1. \tag{3.18}$$

A matriz de informação de *Fisher* $\mathfrak{T}(\boldsymbol{\beta}) = E[-\frac{\partial S(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}]$ obtém-se pelas segundas derivadas de $\ell_i(\boldsymbol{\beta})$ e tem-se, para famílias regulares,

$$-E\left[\frac{\partial^2 \ell_i(\boldsymbol{\beta})}{\partial\beta_j\partial\beta_k}\right] = E\left[\left(\frac{(Y_i-\mu_i)z_{ij}}{Var[Y_i]}\frac{\partial\mu_i}{\partial\eta_i}\right)\left(\frac{(Y_i-\mu_i)z_{ik}}{Var[Y_i]}\frac{\partial\mu_i}{\partial\eta_i}\right)\right] = \frac{z_{ij}z_{ik}}{Var[Y_i]}\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2 \tag{3.19}$$

Assim, tem-se sob a forma matricial que

$$\mathfrak{I}(\boldsymbol{\beta}) = \boldsymbol{Z^T W Z}, \tag{3.20}$$

com

$$\boldsymbol{Z} = (\boldsymbol{z_1}, \boldsymbol{z_2}, \dots, \boldsymbol{z_n})^T = \begin{pmatrix} z_{11} & z_{12} & \cdots & z_{1p} \\ z_{21} & z_{22} & \ddots & \vdots \\ \vdots & \vdots & & \\ z_{n1} & x_{n2} & \cdots & z_{np} \end{pmatrix}, \text{com } p = k+1 \tag{3.21}$$

e $\boldsymbol{W}$ uma matriz diagonal de pesos definidos por

$$w_i = \frac{\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2}{Var[Y_i]} = \frac{\omega_i\left(\frac{\partial\mu_i}{\partial\eta_i}\right)^2}{\phi V(\mu)}, \tag{3.22}$$

visto que $V(\mu) = b''(\theta)$ e $Var[Y_i] = \frac{\phi}{\omega}V(\mu)$.

A equação (3.17) pode ser escrita por

$$\left[\mathfrak{I}\left(\widehat{\boldsymbol{\beta}}^{(k)}\right)\right]\hat{\beta}^{(k+1)} = \left[\widehat{\mathfrak{I}\left(\widehat{\boldsymbol{\beta}}^{(k)}\right)}\right]\beta^{(k)} + s\left(\widehat{\boldsymbol{\beta}}^{(k)}\right) = Z^T W^{(k)}\boldsymbol{u}^{(k)}, \tag{3.23}$$

com $\boldsymbol{u}^{(k)}$ um vector tal que

$$u_i^{(k)} = \sum_{j=1}^{p} z_{ij}\beta_j^{(k)} + \left(y_i - \mu_i^{(k)}\right)\frac{\partial\eta_i^{(k)}}{\partial\mu_i^{(k)}} = \eta_i^{(k)} + \left(y_i - \mu_i^{(k)}\right)\frac{\partial\eta_i^{(k)}}{\partial\mu_i^{(k)}}. \tag{3.24}$$

Logo, a estimativa de $\boldsymbol{\beta}$ na (k+1)-ésima iterada é dada por

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \left(\boldsymbol{Z}^T\boldsymbol{W}^{(k)}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{W}^{(k)}\boldsymbol{u}^{(k)}, \tag{3.25}$$

onde $\boldsymbol{W}^{(k)}$ é a representação da matriz $\boldsymbol{W}$ calculada em $\widehat{\boldsymbol{\mu}}^{(k)}$.

3.4 Testes de hipóteses

Os testes de hipóteses sobre o vector de parâmetros β formulam-se, essencialmente, por

$$H_0: \boldsymbol{C\beta} = \boldsymbol{\xi} \quad versus \;\; H_1: \boldsymbol{C\beta} \neq \boldsymbol{\xi}$$

onde $\boldsymbol{C}$ é uma matriz $q \times p$ com $q \leq p$ e $\boldsymbol{\xi}$ é um vector de dimensão q previamente definido.

Os testes mais usuais para testar essas hipóteses são os testes de Wald, de Wilks (ou razão de verosimilhanças) e o de Rao (ou score). De seguida serão apresentados apenas o teste de Wald e o teste da razão de verosimilhanças, por serem os mais usuais.

3.4.1 Teste de Wald

Seja $\widehat{\boldsymbol{\beta}}$ o estimador de máxima verosimilhança de $\boldsymbol{\beta}$, que se verifica seguir, aproximadamente, uma distribuição normal multivariada

$$\widehat{\boldsymbol{\beta}} \sim N_p\left(\boldsymbol{\beta}, \mathfrak{I}^{-1}(\widehat{\boldsymbol{\beta}})\right). \tag{3.26}$$

Pelas propriedades desta distribuição e como $\boldsymbol{C}\widehat{\boldsymbol{\beta}}$ é uma transformação linear de $\widehat{\boldsymbol{\beta}}$,

$$\boldsymbol{C}\widehat{\boldsymbol{\beta}} \sim N_q\left(\boldsymbol{C\beta}, \boldsymbol{C}\,\mathfrak{I}^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{C}^T\right). \tag{3.27}$$

Sob a hipótese nula tem-se, assim, a *Estatística de Wald* definida por

$$\mathcal{W} = \left(\boldsymbol{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{\xi}\right)^T \left[\boldsymbol{C}\,\mathfrak{I}^{-1}(\widehat{\boldsymbol{\beta}})\boldsymbol{C}^T\right]^{-1}\left(\boldsymbol{C}\widehat{\boldsymbol{\beta}} - \boldsymbol{\xi}\right), \tag{3.28}$$

com distribuição assintótica $\mathcal{X}^2$ com q graus de liberdade.

3.4.2 Teste da razão de verosimilhanças

Este teste assenta na estatística de Wilks ou estatística de razão de verosimilhanças, definida por

$$\Lambda = -2\ln\frac{\max_{H_0} L(\boldsymbol{\beta})}{\max_{H_0 \cup H_1} L(\boldsymbol{\beta})} = -2\{\ell(\tilde{\boldsymbol{\beta}}) - \ell(\hat{\boldsymbol{\beta}})\} \quad (3.29)$$

com $\tilde{\boldsymbol{\beta}}$ o estimador de máxima verosimilhança restrito, i.e., o valor de $\boldsymbol{\beta}$ que maximiza a verosimilhança sujeito às restrições impostas pela hipótese $\boldsymbol{C\beta} = \boldsymbol{\xi}$. Esta estatística, sob a hipótese nula, tem, também, uma distribuição assintótica $\mathcal{X}^2$ com q graus de liberdade.

3.5 Qualidade do ajustamento – Desvio e resíduos

Na escolha do modelo mais adequado pretende-se escolher aquele que melhor se ajuste aos dados e melhor interprete o problema com o menor número de covariáveis envolvidas.

Uma das medidas que verifica a qualidade do ajustamento do modelo aos dados é a designada função desvio. Antes da sua definição é conveniente explicitar dois dos vários tipos de modelos mais comumente referidos - o modelo saturado e o modelo nulo.

O modelo completo ou saturado considera o modelo linear generalizado com n parâmetros $\mu_1, \mu_2, \ldots, \mu_n$, um para cada observação . Neste caso, o modelo ajusta-se exactamente aos dados, pois as estimativas de máxima verosimilhança dos μ_i são as próprias observações. Assim, a matriz do modelo é a matriz identidade de dimensão $n \times n$.

O modelo nulo é o modelo mais simples, em que se considera apenas um parâmetro que representa a média μ, comum a todas as observações y_i. Neste caso, a matriz do modelo é um vector coluna unitário.

Pretende-se, na prática, encontrar um modelo cujo número de parâmetros se situe entre o número de parâmetros de cada um dos modelos acima citados.

Partindo de um modelo em investigação M, com m parâmetros, introduz-se uma medida da distância entre os valores ajustados $\hat{\boldsymbol{\mu}}$ e os correspondentes valores observados $\boldsymbol{Y}$, baseada na estatística de Wilks. Assim, fazendo a comparação entre o modelo em investigação e o modelo completo (ou saturado), obtém-se

$$D^*(y;\mu) = -2\left(\ell_M(\widehat{\boldsymbol{\beta}}_M) - \ell_S(\widehat{\boldsymbol{\beta}}_S)\right), \tag{3.30}$$

com $\ell_M(\widehat{\boldsymbol{\beta}}_M)$ *e* $\ell_S(\widehat{\boldsymbol{\beta}}_S)$ o máximo da função log-verosimilhança para o modelo em investigação e para o modelo saturado, respectivamente.

Como já referido, a função log-verosimilhança de um modelo linear generalizado é dada por

$$\ln L(\boldsymbol{\beta}) = \ell(\boldsymbol{\beta}) = \sum_{i=1}^{n} \frac{\omega_i\left(y_i q(\mu_i) - b(q(\mu_i))\right)}{\phi} + c(y_i, \phi, \omega_i), \tag{3.31}$$

em que se substituiu θ_i por $q(\mu_i)$ de modo a salientar a relação entre θ_i e μ_i.

Tendo em conta que, para o modelo saturado, se tem $\hat{\mu}_i = y_i$, então

$$\ell_S(\widehat{\boldsymbol{\beta}}_S) = \sum_{i=1}^{n} \frac{\omega_i\left(y_i q(y_i) - b(q(y_i))\right)}{\phi} + c(y_i, \phi, \omega_i). \tag{3.32}$$

Para o modelo em investigação tem-se

$$\ell_M(\widehat{\boldsymbol{\beta}}_M) = \sum_{i=1}^{n} \frac{\omega_i\left(y_i q(\hat{\mu}_i) - b(q(\hat{\mu}_i))\right)}{\phi} + c(y_i, \phi, \omega_i), \tag{3.33}$$

onde se designou por $\hat{\mu}_i$ a estimativa de máxima verosimilhança de μ_i.

Assim, obtém-se o designado desvio reduzido $D^*(\boldsymbol{y};\boldsymbol{\mu})$

$$D^*(\boldsymbol{y};\boldsymbol{\mu}) = -2\sum_i \frac{\omega_i}{\phi}\{[y_i q(\hat{\mu}_i) - b(q(\hat{\mu}_i))] - [y_i q(y_i) - b(q(y_i))]\} = \frac{D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})}{\phi}. \tag{3.34}$$

A $D(\boldsymbol{y};\widehat{\boldsymbol{\mu}})$ dá-se o nome de desvio para o modelo em análise.

3.5.1 Análise dos Desvios

Nos modelos lineares clássicos é feita uma análise de variância para verificar o ajustamento do modelo aos dados. A análise do desvio é a generalização dessa análise para os modelos lineares generalizados, tendo como objectivo testar modelos encaixados.

Um modelo M_1 é um submodelo (encaixado) do modelo M, com vector parâmetro $\boldsymbol{\beta}$ de dimensão p, se é um modelo com vector de parâmetros $\boldsymbol{\beta_1}$ subvector de $\boldsymbol{\beta}$.

A comparação de modelos encaixados é feita pela função desvio e a diferença entre os seus desvios coincide com a estatística de razão de verosimilhanças. Neste caso tem-se que a hipótese nula representa o modelo menor, em que alguns dos parâmetros do modelo maior são nulos, contra a hipótese representando o modelo maior. Com esta análise da diferença dos desvios pretende-se explicar a variação dos dados com base nos termos que estão num modelo mas não no outro, considerando-se, portanto, significativos os termos que estão no modelo maior que não estão no modelo menor, em caso de rejeição da hipótese nula.

3.5.2 Informação de Akaike

Outro critério relevante usado na selecção dos modelos é o critério da informação de Akaike (AIC), baseado na função de log-verosimilhança, com o objectivo de medir a informação perdida na construção do novo modelo a partir de um maior.

Para o modelo em estudo, a estatística de Akaike correspondente é

$$AIC = -2\ell\left(\tilde{\boldsymbol{\beta}}_1, 0, \tilde{\phi}\right) + 2r, \tag{3.35}$$

com $r = \dim(\boldsymbol{\beta}_1)$.

Dado um conjunto de modelos candidatos para explicar os dados, aquele que tem menor AIC é preferível a todos os outros. Este critério não se restringe a compensar a qualidade de ajustamento modelos mas introduz, ainda, uma penalização que é uma função crescente do número de parâmetros estimados.

3.5.3 Análise dos Resíduos

Uma das principais etapas na selecção do modelo mais adequado é a análise dos resíduos. Relativamente aos GLM, os resíduos avaliam a qualidade de ajustamento de um modelo com respeito à escolha da distribuição, função de ligação e dos termos no preditor linear, medindo a discrepância entre os valores observados y_i e os valores ajustados $\hat{\mu}_i$. No entanto, os resíduos são, também, úteis na busca de pontos aberrantes, ou seja, observações que o modelo não explica adequadamente.

3.5.4 Matriz de projecção generalizada

Tal como visto anteriormente, o método iterativo dos mínimos quadrados conduz a

$$\widehat{\boldsymbol{\beta}}^{(k+1)} = \left(\boldsymbol{Z}^T\boldsymbol{W}^{(k)}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{W}^{(k)}\boldsymbol{u}^{(k)}, \tag{3.36}$$

sendo equivalente à solução obtida dos mínimos quadrados para o modelo

$$\boldsymbol{u_0} = \boldsymbol{Z}_0\boldsymbol{\beta} + \tilde{\boldsymbol{\epsilon}}, \tag{3.37}$$

onde $\boldsymbol{u}_0 = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{u}$, $Z_0 = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Z}$. Assim, a matriz de projecção generalizada é dada por

$$\boldsymbol{H} = \boldsymbol{Z}_0(\boldsymbol{Z}_0^T\boldsymbol{Z}_0)^{-1}\boldsymbol{Z}_0^T = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Z}\left(\boldsymbol{Z}^T\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Z}\right)^{-1}\boldsymbol{Z}^T\boldsymbol{W}^{\frac{1}{2}} = \boldsymbol{W}^{\frac{1}{2}}\boldsymbol{Z}\mathfrak{I}^{-1}(\boldsymbol{\beta})\boldsymbol{Z}^T\boldsymbol{W}^{\frac{1}{2}}. \tag{3.38}$$

Os elementos da diagonal desta matriz (h_{ii}) são importantes na análise de pontos influentes, como será discutido adiante.

3.5.5 Resíduos

Existem várias definições de resíduos, entre elas o resíduo de Pearson, o resíduo de Anscombe e o desvio residual, que apresentar-se-ão de seguida.

Seja R_i o resíduo referente à i-ésima observação. Considera-se que estes sejam padronizados e reduzidos, i.e., com variância constante unitária e que sejam aproximadamente normalmente. Note-se, ainda, que assimptoticamente se tem

$$Var(\hat{\mu}_i) = Var(Y_i)h_{ii}$$
$$Var(Y_i - \hat{\mu}_i) = Var(Y_i)(1 - h_{ii}),$$

com h_{ii} o $i-$ é$simo$ elemento da diagonal da matriz de projecção $\boldsymbol{H}$, atrás definida.

O resíduo de Pearson é dado por

$$R_i^P = \frac{y_i - \hat{\mu}_i}{\sqrt{\widehat{Var[Y_i]}}} = \frac{(y_i - \hat{\mu}_i)\omega_i}{\sqrt{\hat{\phi}V[\hat{\mu}_i]}}. \quad (3.39)$$

Este resíduo, apresenta, no entanto, a desvantagem de ter uma distribuição bastante assimétrica para modelos não normais.

O resíduo de Pearson padronizado, tendo em conta que $Var(Y_i - \widehat{\mu_i}) = Var(Y_i)(1 - h_{ii})$, é dado por

$$R^{*P}_i = \frac{(y_i - \hat{\mu}_i)\omega_i}{\sqrt{\phi V[\hat{\mu}_i](1 - h_{ii})}}. \quad (3.40)$$

Seja D o desvio definido em (3.34), considere-se a i-ésima observação da função desvio, d_i.

O desvio residual faz uso destas componentes d_i e é definido por

$$R_i^D = \delta_i\sqrt{d_i},$$

com $\delta_i = sinal(y_i - \hat{\mu}_i)$. Estes resíduos podem ser vistos como variáveis aleatórias com distribuição aproximadamente normal e, consequentemente, $R_i^{D\,2} = d_i$ tem distribuição aproximadamente χ_1^2.

O desvio residual padronizado é dado por

$$R^{*D}_i = \frac{R_i^D}{\sqrt{\hat{\phi}(1 - h_{ii})}}.$$

3.6 Formulação do modelo

No processo de modelação dos dados através de um GLM, existem três etapas fundamentais:

- Formulação do modelo: fase em que há a escolha da distribuição da variável resposta, das covariáveis e da função de ligação.

- Ajustamento do modelo: fase da estimação dos parâmetros do modelo e erros padrão; obtenção de intervalos de confiança e testes e realização de testes de qualidade do ajustamento

- Selecção e validação do modelo: pretende-se encontrar submodelos com um número moderado de parâmetros que ainda sejam adequados aos dados; detectar discrepâncias entre os dados e os valores preditos; análise de resíduos; etc.

De seguida, serão apresentados dois modelos que interessam na análise dos dados em estudo: o modelo binomial e o modelo Poisson.

3.6.1 Modelo Binomial e função de ligação logit

O modelo binomial é especialmente útil para modelar dados binários ou na forma de proporções.

Suponhamos que as variáveis resposta $Y_i, i = 1, \dots, n$ têm distribuição Binomial $B(1, \pi_i)$, i.e.,

$$f(y_i) = {\pi_i}^{y_i}(1 - \pi_i)^{1-y_i}, \;\; y_i = 0,1, \tag{3.41}$$

e cada unidade experimental tem associado um vector de especificação $\boldsymbol{z}_i$.

Tal como verificado anteriormente, a distribuição binomial $B(n_i, \pi_i)$ pertence à família exponencial e $\theta_i = \ln\left(\frac{\pi_i}{1-\pi_i}\right)$. Logo, fazendo $\theta_i = \eta_i = \boldsymbol{z}_i^T\boldsymbol{\beta}$, tem-se que a função de ligação canónica é a função logit. Resolvendo em ordem a π_i tem-se

$$\pi_i = \frac{\exp\left(\boldsymbol{z}_i^T\boldsymbol{\beta}\right)}{1 + \exp(\boldsymbol{z}_i^T\boldsymbol{\beta})}. \tag{3.42}$$

Note-se que a função $F: \mathbb{R} \to [0,1]$ dada por

$$F(x) = \frac{\exp(x)}{1 + \exp(x)}, \tag{3.43}$$

é a função de distribuição logística.

Desta forma, o GLM modelo binomial e a função de ligação canónica *logit* é o *modelo de regressão logística*.

Para este modelo, o vector dos *Score*s e a informação de Fisher são, respectivamente,

$$\begin{aligned} S(\boldsymbol{\beta}) &= \boldsymbol{Z}^T(\boldsymbol{y}-\boldsymbol{\mu}) \ \ e \\ \mathfrak{T}(\boldsymbol{\beta}) &= \boldsymbol{Z}^T\boldsymbol{W}\boldsymbol{Z}, \end{aligned} \tag{3.44}$$

com

$$\boldsymbol{W} = diag\{\pi_i(1-\pi_i)\}.$$

Sob o modelo em investigação, a função log-verosimilhança avaliada nas estimativas de máxima verosimilhança, é dada por

$$\ell_M(\hat{\pi}) = \sum_i (y_i \ln(\hat{\pi}_i) + (n_i - y_i)\ln(1-\hat{\pi}_i)) \tag{3.45}$$

com $\hat{\pi}_i = \frac{\hat{\mu}_i}{n_i}$. Já para o modelo saturado é substituir a estimativa $\hat{\pi}_i$ anterior por $\tilde{\pi}_i = y_i$.

Logo, o desvio para este modelo é expresso por

$$D^*(\boldsymbol{y};\boldsymbol{\mu}) = -2\big(\ell_M(\tilde{\pi}) - \ell_S(\hat{\pi})\big) = 2\sum_i \left\{ y_i \ln(y_i) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{\mu}_i}\right)\right\}. \tag{3.46}$$

Relativamente ao desvio residual, é dado por

$$R_i^D = \delta_i \left[2\left(\ln\left(\frac{y_i}{\hat{\mu}_i}\right) + (1-y_i)\ln\left(\frac{1-y_i}{1-\hat{\mu}_i}\right)\right)\right]^{\frac{1}{2}}, \tag{3.47}$$

com $\delta_i = sinal(y_i - \hat{\mu}_i)$.

3.6.2 Modelo Poisson e função de ligação logarítmica

O modelo Poisson é usualmente utilizado para modelar dados sob a forma de contagens, como por exemplo o número de acidentes por área, tal como se pretende neste estudo. Esta distribuição pertence à família exponencial, pelo que faz parte do conjunto de GLM.

Seja Y_i a variável resposta para a distribuição i, distribuidas independentemente com distribuição Poisson de valor médio μ_i. Considere-se, ainda, $\ln(\mu_i) = \boldsymbol{z}_i^T\boldsymbol{\beta}$, ficando

$$f(y_i|\boldsymbol{x}_i,\mu_i) = \exp(-\mu_i)\frac{{\mu_i}^{y_i}}{y_i!} = \exp\left(-e^{\boldsymbol{z}_i^T\boldsymbol{\beta}} + y_i\boldsymbol{z}_i^T\boldsymbol{\beta} - \ln y_i!\right), \quad y_i = 0,1,\ldots \tag{3.48}$$

Neste caso, tem-se um GLM designado *modelo de regressão de Poisson*. Nestes modelos é usual o uso da função de ligação *logarítmica*.

A função log-verosimilhança é, então, dada por

$$\ell(\boldsymbol{\beta}) \propto \sum_{i=1}^{n} y_i \ln(\mu_i) - \mu_i \tag{3.49}$$

e

$$\ell_M(\hat{\beta}_M) \propto \sum_{i=1}^{n} y_i \ln(\hat{\mu}_i) - \hat{\mu}_i$$
$$\ell_S(\hat{\beta}_S) \propto \sum_{i=1}^{n} y_i \ln(y_i) - y_i. \tag{3.50}$$

Portanto, o desvio para este modelo é dado por

$$D^*(\boldsymbol{y};\hat{\boldsymbol{\mu}}) = 2\left[\sum_{i=1}^{n} y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - \sum_{i=1}^{n}(y_i - \hat{\mu}_i)\right]. \tag{3.51}$$

O desvio residual é dado por

$$R_i^D = \delta_i 2^{1/2} \left[y_i \ln\left(\frac{y_i}{\hat{\mu}_i}\right) - y_i + \hat{\mu}_i \right]^{\frac{1}{2}}. \quad (3.52)$$

3.7 Observações discordantes

A análise dos resíduos permite investigar se existem desvios sistemáticos do modelo. Interessa, também, averiguar se existem desvios isolados do modelo, ou seja, observações mal ajustadas, que se distinguem das outras por não seguirem o mesmo padrão. A este tipo de observações estão associadas três medidas importantes- repercussão, influência e consistência.

3.7.1 Medida de repercussão ("leverage")

Esta medida avalia a influência de uma dada observação, ao medir o efeito que a mesma tem nos valores preditos. No caso dos GLM, a repercussão da *i-ésima* observação no valor predito da *j-ésima* variável é dada pelo *ij-ésimo* elemento da matriz de projecção generalizada (3.38).

O i-ésimo elemento da diagonal desta matriz h_{ii} dá uma medida do efeito de repercussão da i-ésima observação na determinação do valor predito $\hat{\mu}_i$. Como

$$tra(H) = \sum_{i=1}^{n} h_{ii} = p,$$

então cada valor de h_{ii} deve estar, em média, próximo de p/n. Desta forma, diz-se que um ponto tem repercussão elevada se

$$h_{ii} > \frac{2p}{n}.$$

3.7.2 Medida de influência

Esta medida avalia o efeito que uma dada observação tem nas estimativas dos parâmetros do modelo, ao ser modificada ou excluída do mesmo. Portanto, observações influentes podem levar a conclusões indevidas do modelo.

Considere-se $\widehat{\boldsymbol{\beta}}_{(i)}$ o estimador de máxima verosimilhança de $\boldsymbol{\beta}$ obtido da amostra sem a observação $(y_i, \boldsymbol{z}_i)$ e $\widehat{\boldsymbol{\beta}}$ o estimador de máxima verosimilhança obtido da amostra integral. Se houver uma diferença substancial entre estes dois valores, então a observação $(y_i, \boldsymbol{z}_i)$ é considerada influente.

Para avaliar esta característica, define-se a medida de influência de Cook, dada por

$$D_i = \frac{\left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\right)^T (\boldsymbol{Z}^T \boldsymbol{W} \boldsymbol{Z}) \left(\widehat{\boldsymbol{\beta}} - \widehat{\boldsymbol{\beta}}_{(i)}\right)}{p\hat{\phi}}. \tag{3.53}$$

A estimativa de $\widehat{\boldsymbol{\beta}}_{(i)}$ necessita de recurso a métodos iterativos, sendo um processo pesado do ponto de vista computacional quando feito para todas as observações. É, por isso, comum, o recurso a uma aproximação deste valor, ao fazer-se apenas o primeiro passo do processo iterativo. Resulta, então, que

$$\widehat{\boldsymbol{\beta}}_{(i),1} = \mathfrak{T}_{(i)}^{-1}(\widehat{\boldsymbol{\beta}}) \boldsymbol{Z}_{(i)}^T \boldsymbol{W}_{(i)}(\widehat{\boldsymbol{\beta}}) \boldsymbol{u}(\widehat{\boldsymbol{\beta}}). \tag{3.54}$$

3.7.3 Medida de consistência

Observações inconsistentes são aquelas que não seguem a tendência que as restantes observações sugerem, apresentando um resíduo elevado.

Assim, pode adaptar-se o modelo sem essa observação e calcula-se o resíduo que a observação eliminada produz em relação ao correspondente valor predito, isto é, $\hat{\mu}_{(i)} = h(\boldsymbol{z}_i^T \widehat{\boldsymbol{\beta}}_{(i)})$. Estes resíduos são designados de *resíduos de eliminação*.

3.8 Acidentes rodoviários com vítimas em Lisboa

A análise de regressão tem sido uma das técnicas usadas na modelação da ocorrência e gravidade de acidentes rodoviários. Vários autores demonstraram que o uso de um modelo de regressão linear não é o mais apropriado para estudar esta problemática, sugerindo modelos alternativos como o modelo de Poisson ou o modelo de regressão binomial negativa. Mercier *et al* (1997), Kim *et al* (1996), entre outros, desenvolveram modelos de regressão logística, para o estudo da influência de factores como a idade e o sexo na gravidade dos acidentes em zonas rurais e na segurança rodoviária infantil, respectivamente.

Mais recentemente, estudos desenvolvidos na Arábia Saudita revelaram que um modelo logístico é um bom ponto de partida na estimativa da influência de certos factores na gravidade dos acidentes (Al-Ghamdi, 2002).

No caso em estudo pretende-se analisar quais as variáveis que parecem ser explicativas, quer da gravidade, quer do número de acidentes rodoviários com vítimas na cidade de Lisboa por área.

Relativamente à análise da gravidade dos acidentes, pretende-se encontrar um modelo que inclua as variáveis que se mostrem mais relevantes para esta situação. A gravidade dos acidentes, por ser uma variável resposta binária é modelada pelo modelo binomial.

Relativamente ao número de acidentes por área, a modelação foi feita usando o modelo Poisson com função de ligação logarítmica.

Pela análise preliminar feita na secção anterior, verificou-se que as variáveis que parecem estar mais associadas à gravidade do acidente são as descritas no quadro seguinte, bem como a sua codificação.

Variável	Codificação	Nome
Condições de aderência	1=Seco e limpo 0=Outros	ADERENCIA
Estado de conservação	1=Em bom estado 2=Em estado regular 3=Em mau estado	CONS
Factores atmosféricos	1=Bom tempo 0=Outros	METEO

Hora (classes em horas)	1= [0:00, 6:59] 2= [7:00, 10:59] 3= [11:00, 15:59] 4= [16:00, 20:59] 5= [21:00, 23:59]	HORA_C
Luminosidade	1=Em pleno dia 2=Aurora ou Crepúsculo 3=Noite	LUMIN
Natureza	1=Despiste 2=Colisão 3=Atropelamento	NATUR
Sinais luminosos	1=A funcionar normalmente 2=Falha/Intermitente 3=Inexistentes	SINAIS

Na categorização da idade dos condutores, passageiros e peões, no sexo dos condutores, bem como nas variáveis acessórios dos condutores, anos de carta e anos de veículo, dado que existem mais vítimas que acidentes, foi necessária uma categorização que envolvesse o conjunto das vítimas/condutores do mesmo acidente. Desta forma foi feita a seguinte escolha:

Variável	Codificação	Nome
Idade dos condutores	1=Existe pelo menos um jovem envolvido mas nenhum idoso. 2=Existe pelo menos um idoso envolvido mas nenhum jovem. 3=Existe pelo menos um jovem e um idoso. 4=Todos os envolvidos são adultos.	IDD_COND
Peões envolvidos	1=Existem peões envolvidos. 2=Não existem peões envolvidos.	PEAO
Sexo dos condutores	1=Todos os condutores são do sexo feminino 2=Todos os condutores são do sexo masculino 3=Pelo menos um condutor do sexo feminino e um condutor do sexo masculino.	SEXOC
Anos de carta	1=Existe um condutor envolvido com menos de 2 anos de carta. 2=Todos os condutores têm mais de 2 anos de carta.	CARTA
Acessórios dos condutores	1= Todos os condutores envolvidos estavam com cinto de segurança ou capacete. 2=Pelo menos um dos condutores envolvidos não tinha cinto de segurança ou capacete ou estava isento.	ACESS

3.8.1 Análise da gravidade dos acidentes – Regressão Logística

Como referido anteriormente, na análise da gravidade dos acidentes usou-se o modelo de regressão logística para dados binários.

Numa primeira fase, realizou-se um teste de independência do Qui-quadrado para cada uma das variáveis independentes categóricas e a variável resposta *Gravidade dos acidentes*, de forma a verificar se possuem influência significativa em relação a esta. Na Tabela 3.1 é feito um resumo dos dados obtidos para esse teste.

Tabela 3.1 Valores da estatística de teste χ^2 e respectivos valores-p para testes de independência entre a variável resposta e as variáveis independentes consideradas.

Variável	χ^2	**Valor-p**
CONS	4.3359	0.1144
ADERENCIA	4.8062	0.02836
LUMIN	48.2022	3.412e-11
METEO	4.4277	0.03536
NATUR	88.1524	< 2.2e-16
HORA	52.1019	1.313e-10
SINAIS	6.8246	0.03297
IDD_COND	6.8525	0.07675
SEXOC	31.7726	1.261e-07
CARTA	0.0214	0.8837
ACESS	7.2776	0.006982
PEAO	1557.270	2.2e-16

A análise da Tabela 3.1 revela que não há evidências suficientes para afirmar que as variáveis *Estado de conservação* e *anos de carta* dos condutores têm uma influência significativa na gravidade dos acidentes, pelo que não serão incorporadas no modelo.

Ajustou-se, então, um primeiro modelo com todas as restantes variáveis explicativas através da metodologia anteriormente descrita, obtendo-se as estimativas dos parâmetros e testando a significância de cada nível das covariáveis, através do teste de Wald.

Na análise da variável *idade dos condutores*, a categoria 3 apresenta-se como não significativa, apresentando um valor-p bastante elevado para o teste de Wald, de cerca de 0.98. Assim, optou-se por recodificar esta variável, agrupando a categoria 3 à categoria 4.

A selecção do melhor modelo logístico baseia-se no método de selecção *stepwise*, onde são escolhidas as covariáveis baseando-se nos valores-p que a sua inclusão ou exclusão no modelo produzem. Estes valores-p são relativos ao teste da razão de verosimilhança de Wilks e a covariável correspondente é tão mais importante quanto menor o seu valor-p.

Assim, o método *stepwise*, partindo de um modelo com as covariáveis acima citadas, excepto *Estado de conservação* e *anos de carta* dos condutores, resultou num modelo logístico com as variáveis descritas na Tabela 3.2. Nesta tabela são ainda apresentados os valores-p resultante da análise de desvios aquando da inserção de cada uma das variáveis.

Tabela 3.2 Variáveis significativas após aplicação do método stepwise, e respectivos valores das funções desvio, considerando apenas cada uma das variáveis, e valores-p correspondentes.

Variável	x_i	**Função desvio**	**Valor-p**
METEO	x_1	9.407	0.002161
LUMIN	x_2	40.638	1.498e-09
NATUR	x_3	78.498	< 2.2e-16
HORA	x_4	14.458	0.005967
SINAIS	x_5	8.168	0.016836
IDD_COND	x_6	6.107	0.047196
SEXOC	x_7	19.158	6.918e-05
ACESS	x_8	2.168	0.140886
PEAO	x_9	9.259	0.002344

O método excluiu a variável *condições de aderência* , com base no critério de Akaike, pois um modelo incluindo esta variável produz um valor de AIC superior ao que um modelo sem esta variável envolvida produz. Fazendo uma análise dos desvios ao modelo logístico gerado, a Tabela 3.2 sugere que a variável *acessórios* não tem influência tão significativa no contexto geral (valor-p=0.1). Através da análise de desvios (ou teste de razão de verosimilhança) comparando um modelo que inclui esta variável com um que a exclui, não é rejeitada a hipótese da mesma não ser estatisticamente significativa quando incorporada no modelo, apresentando um valor-p de 0.1527. Assim, a variável *acessórios* é retirada do modelo global.

Note-se então que, apesar de as variáveis *condições de aderência* e *acessórios* apresentarem relevância estatística quando analisadas individualmente, na análise global a sua inclusão no modelo deixou de apresentar evidências estatísticas para afirmar que têm influência no mesmo. No entanto, como as condições de aderência do piso poderão estar relacionadas com os factores atmosféricos, foi feita uma análise com a incorporação desta interacção no modelo.

A análise dos desvios sugere, no entanto, que a hipótese da interacção entre essas variáveis ser estatisticamente relevante é rejeitada, com um valor-p bastante elevado para cada uma dessas. Portanto, a variável *factores atmosféricos* é, definitivamente, descartada do modelo.

Considerou-se, ainda, a hipótese de outras interacções serem relevantes na avaliação global do modelo logístico:

- Luminosidade e Hora (classes)
- Idade dos condutores e a Natureza do acidente
- Existência de peões e a Natureza do acidente

Destas interacções, apenas a *idade dos condutores* parece relacionar-se com a *natureza do acidente*. Na comparação do modelo sem esta interacção com o modelo que a introduz, resultou da análise dos desvios um valor-p de 0.06611. Apesar de não parecer significante ao nível de 5%, como o valor-p é ainda menor que 0.1, optou-se por realizar uma análise mais profunda da inclusão desta interacção. Deste modo, esta interacção foi incluída no modelo logístico final.

Na Tabela 3.3 encontram-se as estimativas dos coeficientes de regressão para cada variável, bem como os valores da estatística de Wald e os valores-p correspondentes ao teste de significância para cada categoria de cada covariável. O coeficiente nulo $\hat{\beta}_0$ é -246362.

Tabela 3.3 Coeficientes estimados do modelo considerado e respectivos valores da estatística de Wald e valores-p.

Covariável	$\widehat{\beta}_i$	Estatística de Wald	Valor-p
Factores atmosféricos:			
Outros	-0.34389	-2.727	0.006394
Luminosidade:			
Em pleno dia	-0.14811	0.545	0.585915
Noite	0.42522	1.566	0.117403
Natureza:			
Despiste	0.16125	0.413	0.679788
Colisão	-0.05236	-0.15	0.880994
Hora :			
Categoria 2	-0.34134	-1.826	0.067808
Categoria 3	-0.19913	-1.121	0.262466
Categoria 4	-0.29760	-2.09	0.036618
Categoria 5	-0.53281	-3.339	0.000842
Sinais luminosos:			
Falha/Intermitente	-0.14255	-0.393	0.694179
Inexistentes	-0.22923	-2.786	0.005339
Idade dos condutores:			
Categoria 2	-0.99413	-3.413	0.000642
Categoria 3	-0.37814	-2.406	0.016116
Sexo dos condutores:			
Categoria 2	0.55831	4.236	2.28e-05
Categoria 3	0.47484	2.759	0.005796
Peões envolvidos:			
sim	1.03282	3.324	0.000888
Idade dos condutores x Natureza:			
Categoria 2 x Colisão	0.85262	2.327	0.019977
Categoria 2 x Despiste	1.14297	1.935	0.052942
Categoria 3 x colisão	0.20335	1.003	0.316016
Categoria 3 x despiste	0.58102	2.107	0.035127

3.8.1.1 Validação do modelo

A adequabilidade do modelo é verificada através da análise dos desvios. Partindo da hipótese nula de o modelo se ajustar bem aos dados contra a hipótese de o modelo não se ajustar satisfatoriamente aos dados, a estatística de teste, como foi referido no capítulo anterior, tem distribuição aproximada Qui-quadrado.

O modelo corrente produz para a função desvio o valor de 4796.858 correspondente a um valor-p praticamente unitário, concluindo-se que há forte adequabilidade do modelo.

A análise dos resíduos foi feita com base nos desvios residuais, concluindo-se que os pontos apresentam um desvio residual padronizado de valor compreendido entre [-2,3] (Figura 3.1).

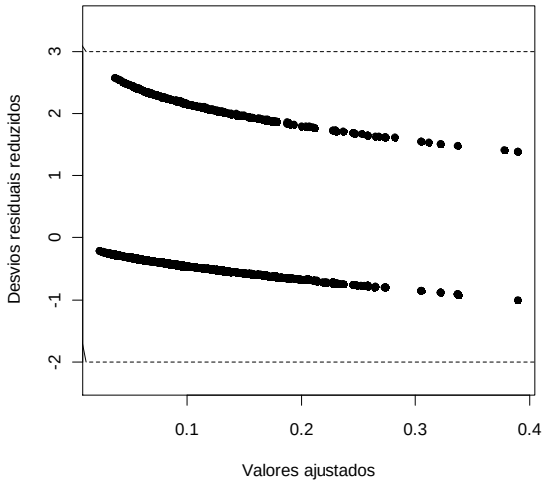


Figura 3.1 Desvios residuais reduzidos.

Na análise das observações com repercussão elevada, existem um número significativo de pontos tais que $\frac{nh_{ii}}{p} > 2$. Portanto, optou-se por considerar que são realmente discordantes as observações tais que $\frac{nh_{ii}}{p} > 10$.

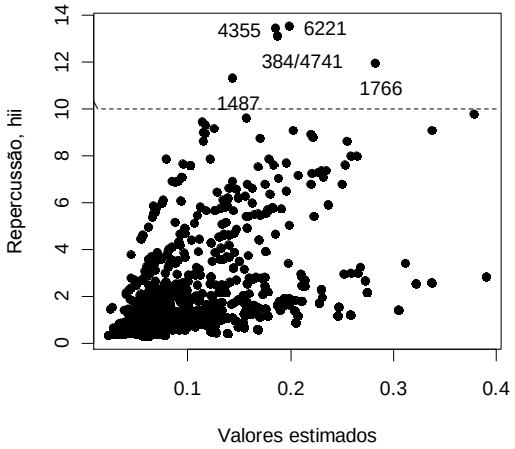


Figura 3.2 Observações com repercussão elevada (>10).

No Anexo 4 representa-se uma tabela com a descrição das características destas observações.

Dessa tabela pode concluir-se que os acidentes com repercussão elevada ocorrem todos de noite, maioritariamente despistes, excepto um que é atropelamento, quase sempre com um idoso envolvido e um condutor do sexo masculino.

Quanto à análise de pontos influentes, recorreu-se à distância de Cook, considerando-se observações influentes aquelas cuja distância de Cook é superior a 0.1. A Figura 3.3 mostra que não existem, portanto, pontos influentes.

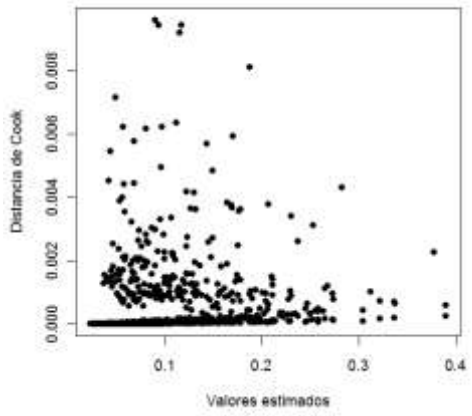


Figura 3.3 Distância de Cook para os valores estimados do modelo.

As observações discordantes são, então, apenas aquelas que contêm repercussão elevada (Anexo 4). Após retirar estas observações do modelo, observou-se que não houve grandes alterações nas estimativas dos coeficientes, nem na sua significância para o modelo.

3.8.1.2 Interpretação do modelo da gravidade dos acidentes

O modelo resultante inclui as variáveis descritas no Tabela 3.3, ou seja

- *Factores atmosféricos*: x_1
- *Luminosidade:* x_2
- *Natureza:* x_3
- *Hora:* x_4
- *Sinais luminosos:* x_5
- *Idade dos condutores:* x_6
- *Sexo dos condutores:* x_7
- *Peões envolvidos:* x_9
- *Idade dos condutores x Natureza:* x_{10}

Resulta, assim, o seguinte preditor linear:

$$\begin{aligned}\eta = g(\boldsymbol{x}) = \beta_0 + \beta_1 x_1 + \sum_{j=1}^{5} \beta_{2j}\, x_{2j} + \sum_{j=1}^{3} \beta_{3j}\, x_{3j} + \sum_{j=1}^{3} \beta_{4j}\, x_{4j} \\ + \sum_{j=1}^{3} \beta_{5j}\, x_{5j} + \sum_{j=1}^{3} \beta_{6j}\, x_{6j} + \sum_{j=1}^{3} \beta_{7j}\, x_{7j} + \beta_9 x_9 \\ + \sum_{j=1}^{3}\sum_{k=1}^{3} \beta_{10jk} x_{3j}\, x_{6k}\end{aligned} \tag{3.55}$$

Na interpretação do modelo interessa perceber qual o impacto das variáveis que se mostraram significativas relativamente à variável dependente. Ou seja, pretende-se estudar a associação das várias variáveis na gravidade dos acidentes. Desta forma, formaram-se as razões de chances[3] para cada um dos níveis das covariáveis.

Recordando o modelo logístico, tem-se que a variável resposta *gravidade* ($\boldsymbol{y}$) segue uma distribuição Bernoulli com probabilidade de sucesso π, isto é, probabilidade de o acidente ser grave.

$$\boldsymbol{y} \sim B(\pi)$$
$$E[\boldsymbol{y}] = \pi$$

A função de ligação *logit* que associa esta variável às variáveis independentes é dada por

$$\begin{aligned} logit(\pi) &= \log\left(\frac{\pi}{1-\pi}\right) = \log(chance) \\ &= \beta_0 + \beta_1 x_1 + \sum_{j=1}^{5} \beta_{2j}\, x_{2j} + \sum_{j=1}^{3} \beta_{3j}\, x_{3j} + \sum_{j=1}^{3} \beta_{4j}\, x_{4j} \\ &+ \sum_{j=1}^{3} \beta_{5j}\, x_{5j} + \sum_{j=1}^{3} \beta_{6j}\, x_{6j} + \sum_{j=1}^{3} \beta_{7j}\, x_{7j} + \beta_9 x_9 \\ &+ \sum_{j=1}^{3}\sum_{k=1,k}^{3} \beta_{10jk} x_{3j}\, x_{6k} \end{aligned} \tag{3.56}$$

[3] Traduz-se o termo *"odds ratio"* por razão de chances.

A razão de chances entre os níveis de uma covariável pode ser interpretada como o aumento estimado na probabilidade de sucesso aquando do aumento de uma unidade no valor predito dessa mesma variável, no caso de variáveis contínuas. Se a variável for categórica, a comparação é feita baseada nos níveis da mesma.

- *Factores atmosféricos*

Para a variável *factores atmosféricos*, a razão de chances relativamente à categoria de *Bom tempo* é dada por

$$\frac{chance|Bom\ tempo}{chance|outros} = \frac{\exp\left(\beta_0 + \sum_{j=1}^{5}\beta_{2j} + \sum_{j=1}^{3}\beta_{3j} + \sum_{j=1}^{3}\beta_{4j} + \sum_{j=1}^{3}\beta_{5j} + \sum_{j=1}^{3}\beta_{6j} + \sum_{j=1}^{3}\beta_{7j} + \beta_9 + \sum_{j=1}^{3}\sum_{k=1,k}^{3}\beta_{10jk}\right)}{\exp\left(\beta_0 + \beta_1 + \sum_{j=1}^{5}\beta_{2j} + \sum_{j=1}^{3}\beta_{3j} + \sum_{j=1}^{3}\beta_{4j} + \sum_{j=1}^{3}\beta_{5j} + \sum_{j=1}^{3}\beta_{6j} + \sum_{j=1}^{3}\beta_{7j} + \beta_9 + \sum_{j=1}^{3}\sum_{k=1,k}^{3}\beta_{10jk}\right)}$$
$$= \exp(-\beta_1) = 1.41.$$

Este valor indica que a chance de ocorrer um acidente grave em que as condições atmosféricas sejam de Bom tempo é cerca de 1.410 vezes a chance de ocorrer em *outras* condições.

- *Luminosidade*

Da mesma forma, as razões de chances para cada categoria da variável Luminosidade são apresentadas no quadro abaixo:

Luminosidade	**Razão de chances**
Em pleno dia	0.564
Aurora ou Crepúsculo	0.758
Noite	1.77

Quanto à luminosidade, a chance de em pleno dia ocorrer um acidente grave é 0.564 vezes a chance de ocorrer durante a noite. Já em estado de aurora ou crepúsculo, a chance de ocorrer um acidente grave é 0.758 a chance de ocorrer nas outras situações. Há, portanto, maior chance de ocorrer um acidente grave de noite, cerca de mais 1.77 vezes a chance de ocorrer em pleno dia ou em aurora ou crepúsculo.

- *Sexo dos condutores*

Sexo	Razão de chances
Categoria 1	0.3558871
Categoria 2	1.087054
Categoria 3	0.9199173

Pela tabela anterior, conclui-se que a chance de um acidente ser grave é maior na categoria 2 quando comparado com qualquer uma das outras categorias. Ou seja, a chance do acidente ser grave se todos os condutores envolvidos forem do sexo masculino é 1.087054 a chance de ocorrer em qualquer das outras situações possíveis.

- *Natureza do acidente*

Natureza	Idade categoria 1	Idade categoria 2	Idade categoria 3
Despiste	1.238	1.655	1.806
Colisão	0.8076605	0.604	0.554
Atropelamento	1	0.122	0.409

Tendo em conta que a natureza do acidente interage com a idade do condutor, então a análise foi baseada nas categorias da variável idade. Desta forma, considerando a categoria 1 da variável idade dos condutores, ou seja, supõe-se que pelo menos um condutor é jovem e nenhum idoso, a chance de um acidente ser grave tendo em conta que se tratou de um despiste é 1.238 vezes maior do que se se tratar de uma colisão ou atropelamento. O mesmo raciocínio é aplicado às restantes categorias.

- *Hora do acidente*

Hora	Razão de chances
Categoria 1	3.939
Categoria 2	1.990
Categoria 3	2.645
Categoria 4	2.172
Categoria 5	1.357

A tabela acima mostra que a chance de ocorrer um acidente grave na categoria 1, ou seja, das 0h00 às 6h59, se destaca por ser 3.939 a chance de ocorrer em qualquer outra das categorias. A categoria 3, ou seja, das 11h00 às 15h59, também apresenta um valor mais elevado.

- *Sinais luminosos*

Sinais luminosos	**Razão de chances**
A funcionar normalmente	1.450
Falha/Intermitente	1.091
Inexistentes	0.917

A chance de um acidente ser grave numa situação em que os sinais luminosos estão a funcionar normalmente é 1.450 vezes a chance de ser grave se os sinais luminosos não existirem ou estiverem em falha ou intermitentes. Se não existirem, há uma menor chance de ocorrer um acidente grave relativamente às restantes categorias.

- *Peão*

Peão	**Razão de chances**
Existência de peões	2.809
Não existência de peões	0.356

A existência de peões aumenta a chance de ocorrer um acidente grave ao invés de ligeiro em quase três vezes do que se não existirem peões envolvidos.

De seguida são apresentadas probabilidades de ocorrência de um acidente grave tendo em conta as características dos acidentes que se mostraram explicativas:

$$\pi = P(acidente\ grave) = 1 - \frac{1}{1 + e^{\eta(x)}}, \tag{3.57}$$

com $\eta(\boldsymbol{x})$ dado pela equação (3.54).

Algumas estimativas dessa probabilidade, associando diferentes características inerentes ao acidente, estão representados no quadro do Anexo 5.

Uma breve análise a esse quadro sugere que a probabilidade de um acidente ser grave aumenta, por exemplo, quando de trata de um atropelamento, ou quando ocorre de madrugada e o estado do tempo não é bom, considerando que os condutores são do sexo masculino e pelo menos um é jovem e nenhum idoso. Essa mesma probabilidade diminui quando se tratam de condutores do sexo feminino, ou quando está envolvido um idoso e nenhum jovem.

3.8.2 Análise do número de acidentes por área– Regressão Poisson

Na análise do número de acidentes por área, foram considerados os acidentes em cada uma das 53 freguesias de Lisboa. A análise exploratória feita neste âmbito na secção 2.1 revelou que a distribuição do número de acidentes por freguesia não é homogénea. Numa primeira fase os acidentenes foram considerados em cada um dos quatro anos em estudo e, posteriormente, consideraram-se todos os anos numa só análise, de modo a esbater os efeitos aleatórios que uma análise anual possa conter.

As variáveis a considerar neste estudo são dadas pela Tabela 3.4.

Tabela 3.4 Variáveis a considerar na análise do número de acidentes por freguesia.

Variável	**Nome**	$\boldsymbol{x_i}$
Proporção da população que usa automóvel	$\text{Prop}_{\text{AUTOM}}$	$\boldsymbol{x_1}$
Proporção da população que trabalha na freguesia	Prop_{IN}	$\boldsymbol{x_2}$
Proporção da população sem nível de ensino	$\text{Prop}_{\text{SEMENS}}$	$\boldsymbol{x_3}$
Número de hospitais na freguesia	Hosp	$\boldsymbol{x_4}$
Número de centros de saúde na freguesia	C_Saúde	$\boldsymbol{x_5}$
Número de escolas na freguesia	Escolas	$\boldsymbol{x_6}$
Encargos mensais por habitação	Enc_mensais	$\boldsymbol{x_7}$
Proporção de população jovem (0-24 anos) por freguesia	$\text{Prop}_{\text{JOVEM}}$	$\boldsymbol{x_8}$
Proporção de população idosa (>64 anos) por freguesia	$\text{Prop}_{\text{IDOSO}}$	$\boldsymbol{x_9}$

Relativamente à variável da proporção da população que utiliza automóvel é necessário referir que o trânsito na freguesia em causa não se limita, claramente, apenas a esta população, mas também ao tráfego de passagem nessas freguesias de utilizadores provenientes de outros locais.

Numa primeira fase, foi analisada a multicolinearidade das variáveis explicativas, ou seja, o grau de correlação entre elas. Uma variável independente pode ser importante na explicação do modelo não só pela sua correlação com a variável dependente mas também pela sua correlação com as restantes variáveis independentes.

A análise da multicolinearidade pode ser feita através do VIF (Variance Inflation Factor), que fornece uma medida do quanto é inflacionada a variância da estimativa dos coeficientes devido à colinearidade. O VIF é dado por

$$VIF_p = \frac{1}{1 - R_p^2}, \tag{3.58}$$

com R_p^2 um coeficiente de determinação múltipla da regressão da covariável $\boldsymbol{x}_p$ em todas as outras covariáveis.

Se $R_p^2 \approx 0$ ($VIF_p \approx 1$) então tem-se independência das covariáveis e quando VIF_p é maior que 10 considera-se que as covariáveis são linearmente dependentes.

3.8.2.1 Número de acidentes rodoviários por ano

Na Tabela 3.5 é, então, analisada a multicolinearidade das várias variáveis independentes consideradas, tendo em conta o VIF.

Tabela 3.5 Valores do VIF para as diferentes variáveis, por ano.

Variável	**VIF (2004)**	**VIF (2005)**	**VIF (2006)**	**VIF (2007)**
$\text{Prop}_{\text{AUTOM}}$	4.223	4.044	4.219	4.226
Prop_{IN}	1.591	1.582	1.621	1.552
$\text{Prop}_{\text{SEMENS}}$	7.50	7.195	7.367	7.440
Hosp	1.389	1.407	1.414	1.344
C_Saúde	1.436	1.441	1.458	1.480
Escolas	4.436	4.365	4.387	4.072

Enc_mensais	3.268	3.184	3.193	3.083
$\text{Prop}_{\text{JOVEM}}$	11.088	10.526	11.025	10.877
$\text{Prop}_{\text{IDOSO}}$	9.538	9.094	9.289	9.232

Os resultados acima sugerem que existe alguma multicolinearidade nas variáveis Proporção de jovens e Proporção de idosos residentes na freguesia. A fim de ultrapassar esta questão, considerou-se apenas a Proporção de idosos e concluiu-se que essa multicolinearidade foi ultrapassada.

Tabela 3.6 Valores do VIF para a proporção de idosos, após a retirada da proporção de jovens.

Variável	**VIF (2004)**	**VIF (2005)**	**VIF (2006)**	**VIF (2007)**
$\text{Prop}_{\text{IDOSO}}$	3.347	3.111	3.228	3.148

Ao analisar-se o número de acidentes por freguesia, espera-se que este seja tanto maior quanto mais tráfego existir na mesma. Assim, de forma a uniformizar este número e ser possível uma análise coerente, é necessária a inclusão da variável tráfego sem que lhe seja atribuído nenhum coeficiente. Este dado é designado de *offset*. Trata-se, pois, de considerar λ como a taxa de acidentes por unidade de tráfego e tem-se

$$\lambda = \frac{\mu}{l},$$

com μ o número médio de acidentes numa dada freguesia e l o tráfego existente nessa mesma freguesia. Assim,

$$\log(\lambda_i) = \beta_0 + \sum_{j=1}^{7} \beta_j x_{ij}$$

$$\log\left(\frac{\mu_i}{l_i}\right) = \beta_0 + \sum_{j=1}^{7} \beta_j x_{ij}$$

$$\log(\mu_i) = \log(l_i) + \beta_0 + \sum_{j=1}^{7} \beta_j x_{ij}.$$

A informação que se tem do tráfego diz respeito ao considerado no período de ponta da manhã e ao número máximo de carros que circulam num determinada secção da via, por hora. Desta forma, considera-se um *offset*, referente a esse tráfego máximo.

O modelo inicialmente proposto é então, para cada ano e para a freguesia i, sendo y_i o número de acidentes na $i-ésima$ freguesia e $y_i \sim Poisson(\lambda_i)$,

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_5 x_{i5} + \beta_6 x_{i6} + \beta_7 x_{i7} + offset(log(l_i)).$$

Relativamente ao ano 2004, o método *stepwise* sugere que as variáveis *Proporção de idosos* e *Número de centros de saúde* não são importantes na explicação do modelo. Assim, o modelo final proposto para esse ano é

$$\log(\lambda_i) = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \beta_4 x_{i4} + \beta_6 x_{i6} + \beta_7 x_{i7} + offset(\log(l_i)). \quad (3.60)$$

No entanto, relativamente ao ano de 2005, o método sugere que a *Proporção da população sem ensino* não é relevante no modelo. Para o ano de 2006 as variáveis *Proporção da população sem ensino* e os *encargos mensais* auferidos por habitação não são relevantes; para 2007, a variável *Número de centros de saúde* parece não ter significância.

As estimativas dos parâmetros de regressão do modelo *log-linear* são apresentados na Tabela 3.7, tendo em conta cada ano separadamente.

Tabela 3.7 Coeficientes de regressão estimados do modelo de Poisson resultante, por ano.

	Ano 2004	**Ano 2005**	**Ano 2006**	**Ano 2007**
β_0	-3.986	-3.665	-1.170	-2.108
$\text{Prop}_{\text{AUTOM}}$	-0.007	-0.008	-0.010	-0.0089
Prop_{IN}	4.979	4.637	2.388	4.045
$\text{Prop}_{\text{SEMENS}}$	9.215	8.426	----------	5.048
Hosp	0.042	0.105	0.092	0.145
C_Saúde	----------	----------	-0.041	----------
Escolas	0.176	0.168	0.179	0.136
Enc_mensais	0.003	0.002	----------	0.0014
$\text{Prop}_{\text{îDOSO}}$	----------	----------	-2.403	-3.175

Validação do modelo

Na análise dos desvios residuais, o gráfico abaixo sugere que existem algumas observações que se destacam por se encontrarem com um resíduo mais elevado relativamente às restantes, observações que se consideram influentes.

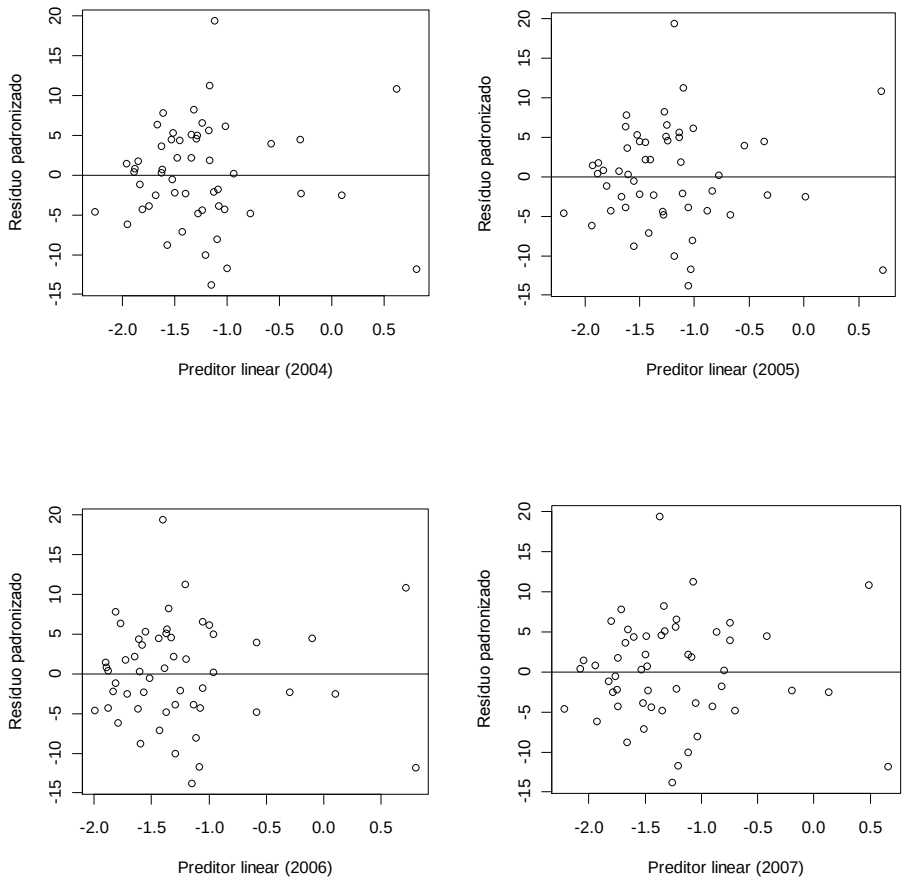


Figura 3.4 Resíduos padronizados resultantes do modelo de Poisson, por ano.

Através da análise das observações influentes, é possível saber quais as freguesias que têm mais influência na gravidade dos acidentes. Para tal, calculou-se a distância de Cook para as várias freguesias, concluindo-se que as freguesias Marvila e Santa Maria dos Olivais, apresentam uma distância de Cook relativamente mais elevada do que as restantes, em todos os anos. A freguesia de

Benfica destaca-se no ano de 2004 e a freguesia de São João de Brito destaca-se nos anos de 2005 e 2007. Um ponto foi considerado influente se a distância de Cook associada se destacar acentuadamente dos restantes, considerando-se o valor 2 como valor limite.

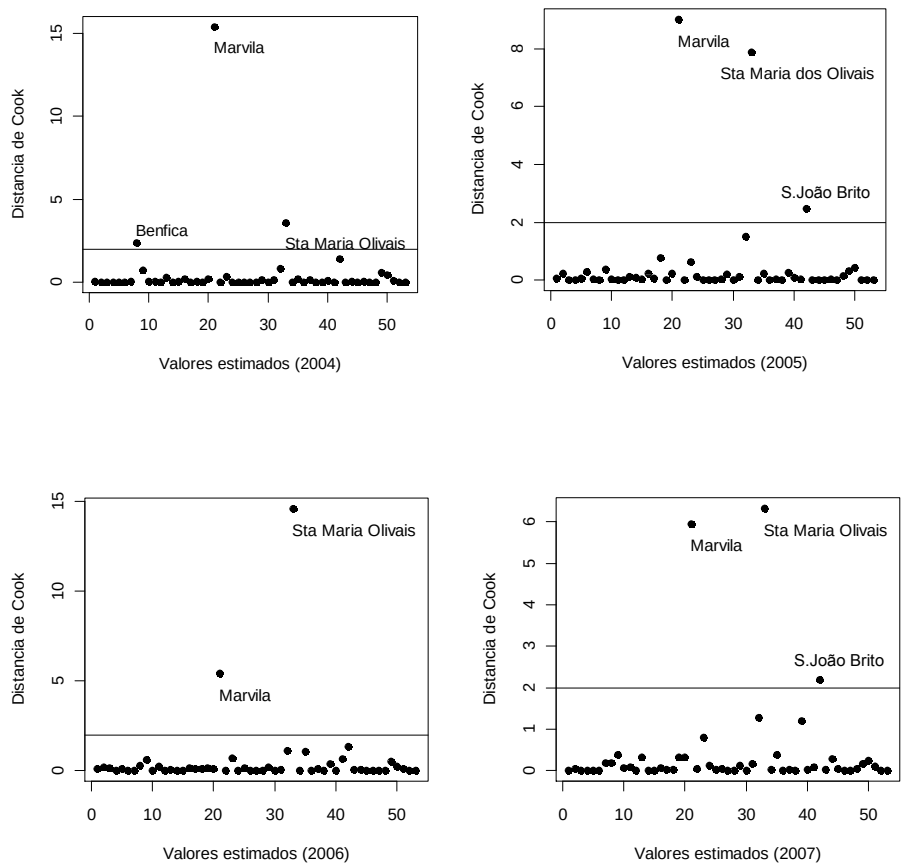


Figura 3.5 Distâncias de Cook para os valores estimados do modelo de Poisson, por ano.

Quanto às freguesias que apresentam repercussão elevada, continuam a destacar-se Marvila e Santa Maria dos Olivais, em todos os anos em estudo. Nos anos de 2004 e 2005 surge também a freguesia do Lumiar, com repercussão elevada. Consideraram-se pontos de repercussão elevada os que realmente se destacam dos restantes, e escolheram-se os diferentes valores limite com base nas figuras abaixo.

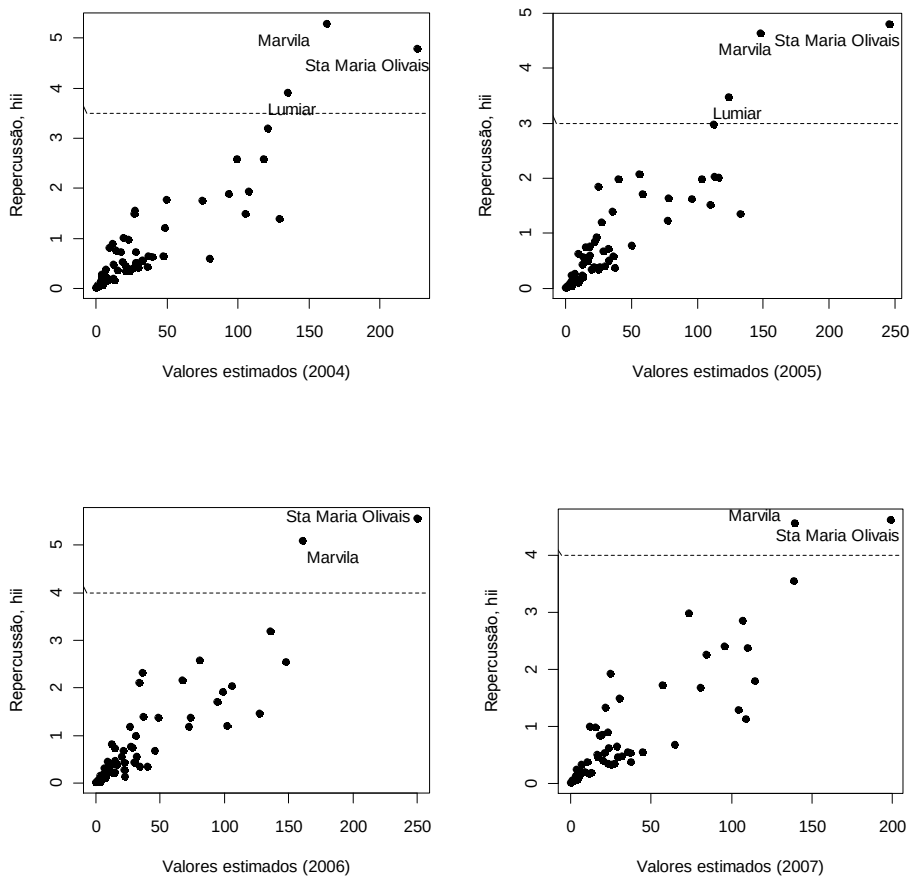


Figura 3.6 Observações com repercussão elevada, por ano.

3.8.2.2 Interpretação do modelo da frequência dos acidentes por ano

A interpretação do modelo aqui apresentada vai incidir em cada variável individualmenteComece por notar-se que a maioria das variáveis consideradas são proporções que variam entre 0 e 1.

Note-se que se tem

$$Y_i = \text{número de acidentes por freguesia} \sim Pois(\lambda_i)$$
$$E[Y_i] = \lambda_i$$

$$Var[Y_i] = \lambda_i$$
$$\lambda_i = \frac{\mu_i}{l_i}.$$

Para o ano de 2004 a equação do modelo ajustado vem

$$\ln(\hat{\lambda}_i) = -3.986 - 0.007\text{Prop}_{\text{AUTOM}} + 4.979\text{Prop}_{\text{IN}} + 9.215\text{Prop}_{\text{SEMENS}} + 0.042\text{Hosp} + 0.176\text{Escolas} + 0.003\text{Enc\_mensais}.$$

Pode, então, dizer-se que para um aumento de 10% na proporção de população que utiliza automóvel, há uma redução de 0.07% na taxa/risco de acidentes (por unidade de tráfego) por freguesia ($e^{-0.007*0.1} = 0.9993$), supondo que todas as outras variáveis se encontram fixas. Não é, portanto, uma variável com muita influência no risco de acidentes. No entanto, para um aumento de 10% na proporção da população que trabalha na freguesia, há um aumento dessa mesma taxa de acidente de 64.53%. O mesmo aumento para a proporção de população sem nível de ensino produz um aumento do risco de acidente de 51.31%. Relativamente às variáveis numéricas (número de hospitais, escolas e encargos mensais por habitação), tem-se um aumento de 4.28% da taxa de acidentes por freguesia por cada hospital existente, e um aumento de 19.24% por escola. Um aumento de 100€ nos encargos mensais por habitação resulta num aumento da taxa de risco de 34.99%, por freguesia.

No ano de 2005, a equação do modelo ajustado é

$$\ln(\hat{\lambda}_i) = -3.665 - 0.008\text{Prop}_{\text{AUTOM}} + 4.637\text{Prop}_{\text{IN}} + 8.426\text{Prop}_{\text{SEMENS}} + 0.105\text{Hosp} + 0.168\text{Escolas} + 0.002\text{Enc\_mensais}.$$

Para o ano de 2005, a situação é bastante idêntica à do ano anterior, pelo que não será feita uma análise detalhada, a fim de evitar redundâncias.

Em 2006, tem-se a equação do modelo ajustado dada por

$$\ln(\hat{\lambda}_i) = -1.17 - 0.0099\text{Prop}_{\text{AUTOM}} + 2.388\text{Prop}_{\text{IN}} - 2.41\text{Prop}_{\text{IDOSO}} - 0.041\text{C\_Saúde} + 0.092\text{Hosp} + 0.179\text{Escolas}.$$

Para este ano, o risco de acidente associado ao aumento da proporçao de população que trabalha na freguesia é menor relativamente aos anos anteriores, verificando-se que um aumento de 10% nesta variável provoca um aumento de 26.97% no risco de acidente. Neste modelo são, também, consideradas duas variáveis que não estão incluídas nos modelos dos anos anteriores: a proporção de população idosa na freguesia e o número de centros de saúde. Um aumento de 10% na proporção de

população idosa provoca uma diminuição na taxa de acidente por freguesia de 21.42%. Por cada centro de saúde há uma diminuição do risco de acidente de 0.41%.

A equação do modelo ajustado para o ano de 2007 é dada por

$$\ln(\hat{\lambda}_i) = -2.108 - 0.009\text{Prop}_{\text{AUTOM}} + 4.045\text{Prop}_{\text{IN}} + 5.048\text{Prop}_{\text{SEMENS}} + 0.145\text{Hosp} + 0.136\text{Escolas} + 0.0014\text{Enc\_mensais} - 3.175\text{Prop}_{\text{IDOSO}}.$$

Esta equação é idêntica à do ano de 2004, pelo que não será feita uma análise tão detalhada. A única alteração é ser considerada a variável proporção de população idosa residente em cada freguesia, cujo aumento, por exemplo, de 10% resulta numa diminuição da taxa de acidentes em 27.2%.

3.8.2.3 Análise do número total de acidentes

Posteriormente, foi feita uma análise com o número de acidentes por freguesia no conjunto de todos os anos, resultando num modelo que envolve todas as variáveis. Os coeficientes associados são apresentados na Tabela 3.8.

Tabela 3.8 Coeficientes de regressão do modelo de Poisson, considerando os acidentes no total dos anos.

Variável	$\boldsymbol{\beta_i}$
Base	-1.242
$\text{Prop}_{\text{AUTOM}}$	-0.009
Prop_{IN}	3.983
$\text{Prop}_{\text{SEMENS}}$	5.102
Hosp	0.096
C_Saúde	-0.022
Escolas	0.166
Enc_mensais	0.0016
$\text{Prop}_{\text{IDOSO}}$	-1.508

Através da análise dos resíduos (Figura 3.7) conclui-se que existem quatro observações que se destacam por terem um valor de resíduo mais elevado: Campo Grande, Marvila, Santa Maria de Belém e São Paulo.

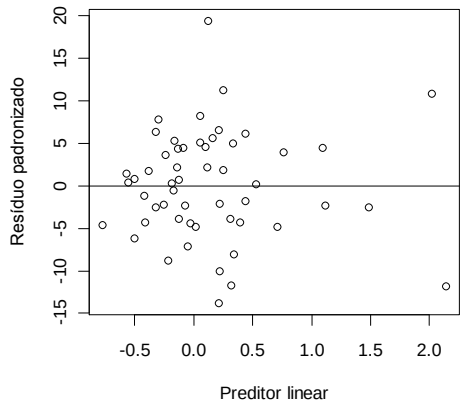


Figura 3.7 Resíduos padronizados do modelo de Poisson, no total dos anos.

A distância de Cook (Figura 3.8) sugere que as freguesias de Marvila, Santa Maria dos Olivais, São João de Brito e Santa Maria de Belém apresentam-se como discordantes, visto apresentarem um valor dessa distância superior a 5.

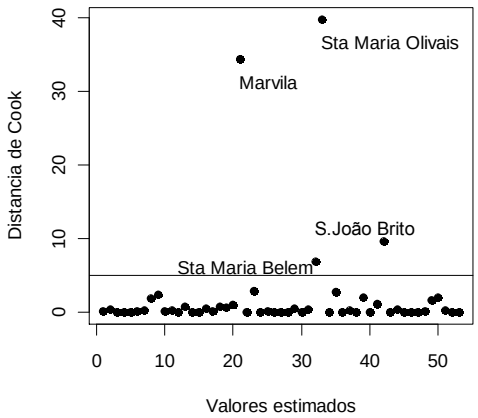


Figura 3.8 Distância de Cook dos valores estimados do modelo de Poisson, para todos os anos.

No entanto, o modelo resultante após descartar essas observações mantém-se inalterado.

3.8.2.4 Interpretação do modelo da frequência dos acidentes no conjunto dos quatro anos

Tal como feito para a análise por anos do número de acidentes por freguesia, a interpretação do modelo aqui efectuada baseia-se em cada covariável, admitindo as restantes como fixas.

Assim, tem-se a equação do modelo ajustado

$$\ln(\lambda_i) = -1.242 - 0.008\text{Prop}_{\text{AUTOM}} + 3.983\text{Prop}_{\text{IN}} - 1.508\text{Prop}_{\text{idosa}} + 5.102\text{Prop}_{\text{SEMES}} - 0.022\text{C\_Saúde} + 0.096\text{Hosp} + 0.166\text{Escolas} + 0.0016\text{Enc\_mensais}.$$

Desta equação resulta que um aumento de 10% na proporção de população que utiliza automóvel produz uma redução de 0.08% na taxa de acidente por unidade de tráfego, por freguesia. Relativamente à proporção da população que trabalha na freguesia, um aumento de 10% nesta proporção resulta num aumento de 48.93% no risco de acidente. Este aumento verifica-se, também, na proporção da população sem ensino, produzindo um aumento de 66.56% na taxa de acidente. Na proporção de população idosa, um aumento da mesma de 10% produz uma redução de 14% na taxa de acidente. Por cada hospital ou escola a mais na freguesia há uma taxa de acidente aumentada, acontecendo o mesmo num aumento dos encargos mensais por habitação. Já os centros de saúde, quantos mais, menor é a taxa de acidente.

3.9 Principais conclusões

Nesta secção foram estudados os efeitos dos factores considerados estatisticamente significativos quer na gravidade dos acidentes, quer no número de acidentes com vítimas na cidade de Lisboa por freguesia, usando os modelos lineares generealizados.

Relativamente à gravidade dos acidentes com vítimas, as variáveis que se mostraram significativas para esse estudo foram os factores atmosféricos, a luminosidade, a hora do dia, o funcionamento dos sinais luminosos, o sexo dos condutores, a existência ou não de peões e a natureza do acidente interagindo com a idade dos condutores. Concluiu-se que a probabilidade de um acidente ser grave é maior se for de noite e entre as 00h00 e as 6h59, assim como se os condutores envolvidos forem todos do sexo masculino, se forem todos ou adultos ou pelo menos um jovem e um idoso

envolvidos e se se despistarem. Condições atmosféricas menos favoráveis e um mau funcionamento dos sinais luminosos não parecem ter influência no aumento da gravidade dos acidentes em Lisboa, ao contrário da existência de peões envolvidos, que aumenta a probabilidade de um acidente ser grave.

Quanto ao número de acidentes com vítimas, estudado por freguesia, em todos os anos há quatro variáveis que são significativas na explicação dos mesmos: a proporção da população da freguesia que usa automóvel, a proporção da população da freguesia que trabalha na mesma, o número de hospitais e o número de escolas. A proporção da população sem ensino e os encargos mensais por habitação apenas não se mostraram relevantes no ano de 2006. Nos anos de 2006 e 2007 considera-se, ainda, a proporção de idosos em cada freguesia.

Quando se consideram todos os anos, todas estas variáveis são significativas para o modelo de Poisson que modela o número de acidentes com vítimas por freguesia.Conclui-se, também, que quanto maior a proporção da população que utiliza automóvel, menor o risco de ocorrer um acidente com vítima, apesar de ser uma redução bastante baixa. No entanto, quanto maior a proporção da população que trabalha na freguesia, maior o risco de acidente, assim como acontece para a proporção de população sem ensino. Um aumento no número de hospitais e escolas aumenta, também, essa probabilidade, ao contrário do aumento nos centros de saúde. A probabilidade de ocorrer um acidente com vítimas na freguesia aumenta, também, aquando do aumento na proporção de população idosa.

Há, portanto, um considerável número de factores que contribuem quer para um aumento da gravidade dos acidentes, quer para um aumento do seu número, quando consideramos uma análise por freguesias.

4 Processos Pontuais Espaciais

4.1 Introdução

A necessidade de interpretar e prever acontecimentos dependentes da localização espacial da sua ocorrência tem conduzido a uma área de conhecimento em desenvolvimento de formulação de modelos complexos assentes em teorias estatísticas que acomodem este carácter espacial das observações dos dados. Estes dados espaciais, cuja análise requer abordagens adequadas, são classificados em três categorias: dados por pontos, dados agregados ou padrões pontuais. Os primeiros estão relacionados com fenómenos que ocorrem em toda uma região no espaço, mas apenas se tem acesso a uma amostra num subconjunto de localizações; os segundos surgem de uma recolha dos dados de forma agregada em subdivisões da região em estudo e, por fim, os padrões pontuais surgem quando é a localização espacial do fenómeno que se pretende modelar. Estes últimos são objecto de estudo desta dissertação.

As primeiras aplicações deste tipo remontam a meados do século XX, na área da ecologia e das ciências florestais (Carvalho & Natário, 1 a 4 de Outubro de 2008). No entanto, a sua aplicação estendeu-se ao longo dos anos a outras áreas, tais como arqueologia, astronomia, epidemologia, geografia, medicina, entre outros.

A nível dos acidentes rodoviários têm sido propostas várias metodologias de cariz espacial em diversas aplicações ((Nicholson, 1998; Flahaut *et al*., 2003; Kim *et al*.,1996). Nicholson (1998) fez uma selecção dos acidentes por tipo de local em função da distribuição espacial dos mesmos, propondo uma classificação dos processos espaciais de acordo com a distribuição das agregações dos acidentes. Banos & Huguenin-Richard (2000) fizeram uma análise espacial dos acidentes com peões crianças, através do uso da estimativa kernel da sua intensidade e compararam a localização desses acidentes com a localização das escolas, a fim de averiguar dependência entre ambas. Flahaut *et al.* (2003) aplicam a autocorrelação espacial e o estimador de kernel para identificar zonas críticas de acidentes e verificam que ambas as técnicas ajudam na identificação de locais críticos.

Com o objectivo de modelar a distribuição aleatória da localização da ocorrência de certos acontecimentos, os processos pontuais espaciais surgem como modelos matemáticos que permitem descrever dados desta natureza. As localizações observadas designam-se de padrões pontuais espaciais, que se definem como um conjunto de pontos existentes numa área ou conjunto, sendo interpretados como uma *realização* de um processo pontual. Esses pontos identificam-se pelas suas coordenadas espaciais, mas podem, também, ter associadas características dos acontecimentos relevantes para o estudo dos mesmos. Pretende-se, assim, explorar a natureza espaço-temporal dos

acidentes, permitindo uma estimação da superfície de risco associada em função de factores extrínsecos ao acidente.

Na maior parte das aplicações práticas, a região em análise é planar, ou seja, a duas dimensões, podendo, no entanto, estender-se para regiões de dimensão maior. Nesta dissertação será considerado apenas o caso a duas dimensões como exemplo.

A análise espacial, apesar de distinta em muitos aspectos da análise estatística clássica, apresenta algumas influências desta, nomeadamente no que respeita à exploração dos dados, estimação dos parâmetros, ajustamento de modelos e testes de hipóteses. A teoria aqui apresentada será, basicamente, dentro desses contextos.

Um dos pontos essencias na modelação dos processos pontuais é a designada *hipótese de aleatoriedade espacial completa* (CSR[4]), em que se assume que as localizações dos acontecimentos numa dada região planar se distribuem uniformemente na área, condicional a um número fixo de acontecimentos, e o seu número segue uma distribuição Poisson de média proporcional à área em questão. É, assim, natural que a esta hipótese estejam associados os processos de Poisson homogéneos, descritos posteriormente.

4.2 Processos Pontuais Espaciais

Nesta secção serão definidos os processos pontuais espaciais, bem como as suas principais propriedades.

Considera-se que os dados constituem conjuntos de pontos não ordenados em $\mathbb{R}^2$.

Definição 4.1 Seja $D \subset \mathbb{R}^2$. Um processo pontual espacial X em D é uma aplicação do espaço de probabilidades $(\Omega, S, \mathcal{P})$ no espaço mensurável $(\mathbb{N}, \mathcal{N})$, tal que

$$\mathbb{N} = \{\boldsymbol{x} \subseteq D : n(\boldsymbol{x_A}) < \infty \quad A \subseteq D, \text{Borel, limitado}\}, \tag{4.1}$$

onde $n(\boldsymbol{x_A}) = N(A)$ é o número de pontos de $\boldsymbol{x} = \{x_1, x_2, \dots, x_n\}$ em A e $\mathcal{N}$ é a menor σ −álgebra de subconjuntos de $\mathbb{N}$ tal que a aplicação

$$\begin{aligned} \mathcal{N} &\to \mathbb{N}_0 \\ \boldsymbol{x} &\mapsto N_x(A) \end{aligned}$$

[4] *CSR* é a abreviatura do inglês para *Complete Spatial Randomness*.

é mensurável.

Note-se que a definição anterior implica que $N(A)$ é uma variável aleatória que toma valores finitos sempre que A for uma região limitada.

Os processos pontuais aqui definidos assumem-se como processos pontuais simples, isto é, $x_i \neq x_j, se\ i = j$, que implica que todos os pontos são não coincidentes.

A medida de probabilidade induzida em $\mathcal{N}$ designa-se a *distribuição de X* e é determinada pela probabilidade de X estar em U, ou seja,

$$P(X \in U), \qquad U \in \mathcal{N} \tag{4.2}$$

Alternativamente, as *distribuições de dimensão finita* de um processo pontual X permitem caracterizar a distribuição de um processo pontual X em $(\mathbb{N}, \mathcal{N})$ e existem infinitas distribuições deste tipo associadas a infinitas sequências de conjuntos A que descrevem um processo pontual. Estas são definidas considerando todas as sequências de conjuntos de Borel limitados $(A_1, A_2, \dots, A_m)$, com m finito, como as distribuições conjuntas de $\big(N(A_1), N(A_2), \dots, N(A_m)\big)$:

$$P(N(A_1) = n_1, \dots, N(A_m) = n_m).$$

De forma a ser possível uma ordenação clara dos acontecimentos com respeito à sua distância a um ponto arbitrário qualquer, exige-se o seguinte para os processos pontuais que serão considerados:

Definição 4.2 Seja $|A|$ a área da região A, os processos pontuais simples dizem-se *ordenados* se obedecem à seguinte condição

$$\lim_{|A| \to 0} \frac{P(N(A) > 1)}{|A|} = 0. \tag{4.3}$$

4.2.1 Estacionariedade e Isotropia

Os processos pontuais espaciais podem ser classificados como estacionários e/ou isotrópicos.

Definição 4.3 Um processo pontual espacial X diz-se *estacionário* se a sua distribuição é invariante sob translacções em $\mathbb{R}^2$, ou seja, para cada vector $v \in \mathbb{R}^2$, a distribuição do processo pontual $X + v$, obtido por translacção de cada ponto $x \in X$ em $x + v$, é identicamente distribuido a X.

Assim, num processo pontual estacionário, as suas propriedades podem ser estimadas recorrendo apenas a uma realização em D, visto que são as mesmas em sub-regiões diferentes de D.

Definição 4.4 X diz-se *isotrópico* se a sua distribuição é invariante sob rotações em $\mathbb{R}^2$. Neste caso uma rotação é descrita por um ângulo α entre $0°$ e $360°$, tal que, sendo $x = (x_1, x_2)$, então as coordenadas do ponto rodado são dadas por

$$x_{1_\alpha} = x_1 \cos(\alpha) + x_2 \sin(\alpha) \quad \text{e} \quad x_{2_\alpha} = -x_1 \sin(\alpha) + x_2 \cos(\alpha).$$

A estacionariedade e a isotropia juntas produzem invariância ao movimento.

4.2.2 Efeitos de fronteira

Na prática, os pontos observados de um processo pontual espacial estão, usualmente, limitados a uma região W, designada de *janela de observação*. Neste caso, apenas se observa $X \cap W$. Quando existe um ponto *u* perto do limite de *W*, o ponto de *X* mais próximo de *u* poderá cair fora da região. A nível inferencial, esta é uma questão bastante relevante.

Definição 4.5 Os problemas de dados faltantes devido a não serem observados os pontos em X/W designam-se efeitos de fronteira.

A fim de ultrapassar este problema, são necessárias correcções de fronteira. Estas correcções são importantes na medida em que muitos estimadores buscam informação na vizinhança dos pontos e, deste modo, ao considerar-se uma região limitada não é possível aceder-se a toda a informação das vizinhanças dos pontos próximos da fronteira. Com estes métodos é possível reduzir a influência da janela de observação W.

Entre estes métodos destaca-se o método da fronteira (*border method*), que assume apenas as vizinhanças com uma distância inferior a r como relevantes para cada ponto. Ou seja, a análise inferencial é feita com base numa redução da janela de observação , W_{-r}, definida como

$$W_{-r} = \{x \in W : b(x, r) \subseteq W\}. \tag{4.4}$$

Esta nova região, sub-região de W, é um subconjunto de pontos de W que estão no seu interior e têm uma distância à fronteira de W superior a r.

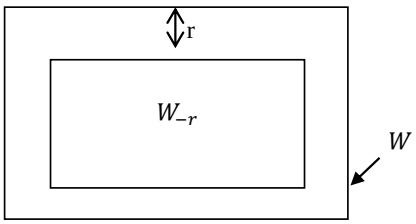


Figura 4.1 Janela de observação rectangular W e correspondente janela de redução W_{-r} de acordo com o método da fronteira.

4.2.3 Funções de distribuição de distâncias

A consideração de distâncias entre as localizações observadas dos acontecimentos ou a pontos arbitrários pode fornecer pistas quanto:

- Às características do processo que lhe está subjacente;
- Ao nível de estacionaridade e isotropia;
- À prevalência ou não da hipótese de aleatoriedade espacial completa;
- Descrições sumárias adicionais dos processos pontuais.

Neste contexto, é importante considerar as funções de distribuição associadas à distância entre um ponto qualquer e o acontecimento que lhe está mais próximo e entre dois acontecimentos próximos.

Definição 4.6 Seja X um processo pontual espacial. A função distribuição da distância de um ponto arbitrário o ao acontecimento que lhe está mais próximo, $F(t)$, é dada por

$$F(t) = 1 - P\big(N(b_{o,t}) = 0\big), \tag{4.5}$$

com $b_{o,t}$ um círculo de raio t centrado em o. Esta função é vulgarmente designada de *função distribuição de espaço vazio*.

Definição 4.7 A função distribuição da distância de um acontecimento arbitrário ao acontecimento que lhe está mais próximo, $G(t)$, designa-se função distribuição do vizinho mais próximo.

Uma função com especial relevância é a função J que combina as duas funções acima referidas:

$$J(t) = \frac{1 - G(t)}{1 - F(t)}, t \geq 0. \quad (4.6)$$

4.2.4 Processos pontuais marcados

Aos pontos de um processo pontual espacial podem estar associadas características relevantes dos mesmos, numéricas ou qualitativas, que são variáveis aleatórias designadas por *marcas*. Estas marcas fornecem, pois, informação adicional ao objecto em estudo representado por um ponto.

Considerando um ponto x, designa-se uma marca associada a esse ponto por m_x .

Definição 4.8 Seja X um processo pontual em $D \subset \mathbb{R}^2$ e $\mathcal{M}$ um espaço tal que $m_x \in \mathcal{M}$ é uma marca associada a $x \in X$. Um processo pontual marcado no espaço D é um processo pontual Y em $D \times \mathcal{M}$.

O espaço das marcas pode ser um conjunto finito, um intervalo contínuo de números reais ou ainda conjuntos mais complexos como o conjunto de todos os polígonos convexos.

No caso contínuo, uma marca pode ser, por exemplo, o volume, a massa, o diâmetro, a altura, enquanto que no caso discreto uma marca é uma opção dentro das várias possíveis. Nos acidentes rodoviários, uma possível marca é a gravidade do acidente, como será aplicado posteriormente.

4.2.5 Momentos e propriedades de segunda ordem

Tal como para as varáveis aleatórias usuais, interessa definir os momentos de um processo pontual.

As propriedades de primeira ordem de um processo pontual, ou simplesmente momentos de primeira ordem, são descritas por uma função intensidade. Quanto às propriedades de ordem superior, a função $K(t)$ desempenha um papel fundamental, como será descrito posteriormente.

Considere-se dx como uma pequena região contendo x e A como sendo uma região limitada e não-vazia de $D \subseteq \mathbb{R}^2$.

Definição 4.9 A função intensidade que caracteriza as propriedades de primeira ordem de um processo pontual é dada por

$$\lambda(x) = \lim_{|dx|\to 0} \left\{ \frac{E[N(dx)]}{|dx|} \right\}. \tag{4.7}$$

Para um processo pontual estacionário tem-se que a medida de intensidade é invariante sob translacções. Neste caso, $\lambda(x) = \lambda = \frac{E[N(A)]}{|A|}$ é designada de intensidade de X e a interpretação da mesma é como sendo o número esperado de acontecimentos por unidade de área.

Definição 4.10 As propriedades de segunda ordem de um processo pontual espacial descrevem-se através de uma função de intensidade de segunda ordem, definida por

$$\lambda_2(x,y) = \lim_{|dx|,|dy|\to 0} \left\{ \frac{E[N(dx)N(dy)]}{|dx||dy|} \right\}. \tag{4.8}$$

Inerente a esta função de segunda ordem está o conceito de intensidade condicional, $\lambda_c(x|y) = \lambda_2(x,y)/\lambda(y)$, que corresponde à intensidade do ponto x sabendo informação do ponto y.

É de notar que para um processo estacionário, a intensidade de segunda ordem depende apenas da quantidade $y - x$, ou seja,

$$\lambda_2(x,y) \equiv \lambda_2(x-y)$$

Se o processo pontual for, ainda, isotrópico, então

$$\lambda_2(x-y) = \lambda_2(t),$$

com $t = \| x - y \|$ e $\|.\|$ é a distância euclideana usual.

Definição 4.11 A densidade de covariância é definida por

$$\vartheta(x,y) = \lambda_2(x,y) - \lambda(x)\lambda(y).$$

Salvo menção em contrário, serão considerados daqui por diante apenas os processos pontuais estacionários e isotrópicos (e ordenados).

Para estes processos, as propriedades de segunda ordem podem ser caracterizadas pela função de segundo momento reduzido, $K(t)$.

Definição 4.12 A função de segundo momento reduzido K(t) é dada por

$$K(t) = \frac{2\pi}{\lambda^2}\int_0^t \lambda_2(u)u du. \tag{4.9}$$

Umas das principais características da importância da função $K(t)$ na descrição de um padrão espacial advém da sua relação com a função $\lambda_2(t)$, na medida em que pode ser interpretado como o valor esperado de uma quantidade que se pode observar. De facto, a equação acima permite definir $\lambda_2(t)$ a partir da função $K(t)$:

$$\lambda K(t) = 2\pi\lambda^{-1}\int_0^t \lambda_2(u)u du \iff \lambda_2(t) = \frac{\lambda^2}{2\pi t}K'(t). \tag{4.10}$$

A partir das definições anteriores pode, ainda, definir-se a função K de outra forma, e que permite estabelecer algumas propriedades interessantes da mesma:

Definição 4.13 Sendo $N_0(t)$ o número de acontecimentos dentro de uma distância t distintos de um acontecimento arbitrário, então tem-se

$$\lambda K(t) = E[N_0(t)]. \tag{4.11}$$

Como consequência de (4.9) , para padrões cujos acontecimentos tenham forte probabilidade de estar rodeados de um espaço vazio, acontecimentos distintos destes dentro de uma distância pequena têm um valor esperado pequeno. No entanto, para acontecimentos que tendem a aparecer aglomerados esse valor esperado será mais elevado.

Outra consequência relevante do segunto momento reduzido e que permite caracterizar padrões regulares ou agregados é a seguinte:

$$K(t) = \pi t^2 + 2\pi\lambda^{-2}\int_o^t v(u)u\, du. \tag{4.12}$$

Portanto, πt^2 é uma marca importante na caracterização dos padrões pontuais. Para padrões agregados tem-se $K(t) > \pi t^2$, como se concluirá posteriormente, e o contrário para padrões regulares.

4.2.6 Estimação das propriedades de segunda ordem

A estimação das propriedades de segunda ordem incidem, essencialmente, na função K, pois esta é mais facilmente estimável que a função intensidade λ_2. De acordo com a equação (4.10), é possível, então, encontrar a correspondente estimativa de λ_2. Considere-se uma largura de banda h e seja $\widehat{K}(t)$ a estimativa de $K(t)$. Então, pode fazer-se a aproximação:

$$\widehat{K}'(t) \approx \frac{\widehat{K}(t+h) - \widehat{K}(t)}{h},$$

de onde se deduz

$$\hat{\lambda}_2(t) = \frac{\hat{\lambda}^2}{2\pi t}\widehat{K}'(t).$$

Esta última produz uma estimativa tipo histograma em intervalos h de $[0, t]$.

Uma estimativa óbvia da função K advém do facto de $\lambda K(t)$ poder ser interpretado como o valor esperado de $N_0(t)$, tal como visto anteriormente. Sendo, ainda, a intensidade λ o número médio de acontecimentos por unidade de área, então uma sua estimativa pode ser

$$\hat{\lambda} = \frac{n}{|W|}.$$

Designando por $E(t) = E[N_0(t)]$ o número esperado de acontecimentos dentro de uma distância t de um acontecimento arbitrário, este pode ser estimado considerando a distância $u_{ij} = \| x_i - x_j \|$ como

$$\tilde{E}(t) = \frac{1}{n}\sum_{i=1}^{n}\sum_{j\neq i}^{n} I(u_{ij} \leq t), \tag{4.13}$$

em que $I(.)$ é a função indicatriz.

Um dos problemas deste estimador prende-se com os efeitos de fronteira, que pode ser corrigido usando o método de fronteira descrito atrás.

Acontece, também, que sob amostragem, a distribuição de $\widehat{K}(t)$ é difícil de tratar analiticamente, excepto para padrões homogéneos.

4.2.7 Estimação das distribuições do vizinho mais próximo e do espaço vazio

A estimação das funções distribuição do espaço vazio e do vizinho mais próximo torna-se bastante útil, quer como complemento da estimação das propriedades de segunda ordem, quer em testes para a hipótese de *CSR*. As distribuições empíricas destas funções são um bom ponto de partida para uma boa aproximação das mesmas.

Considere-se uma região W, a janela de observação, e cubra-se a mesma com uma fina grelha de m pontos tal que t_i é a distância do i-ésimo desses pontos ao acontecimento que lhe está mais próximo. Segue, então, que a função distribuição empírica do espaço vazio pode ser dada por:

$$\tilde{F}(t) = \frac{\sum_{i=1}^{m} I(t_i \leq t)}{m}. \tag{4.14}$$

Devido aos efeitos de fronteira, este estimador é enviesado negativamente, e uma possível correcção para o mesmo é considerar-se apenas os pontos que se encontrem a uma distância da fronteira inferior ou igual a t. Assim, sendo e_i a distância do i-ésimo ponto ao ponto da fronteira que lhe está mais próximo, obtém-se o estimador

$$\hat{F}(t) = \frac{\sum_{i=1}^{m} I(e_i \geq t, t_i \leq t)}{\sum_{i=1}^{m} I(e_i \geq t)}. \tag{4.15}$$

Para a função distribuição do vizinho mais próximo utiliza-se, também, a função distribuição empírica e a correspondente correcção, mas considerando t_i como as distâncias do i-ésimo acontecimento ao acontecimento que lhe está mais próximo e e_i como as distâncias do i-ésimo ponto ao acontecimento que lhe está mais próximo na fronteira da janela W. Desta forma, considerando que se tem n acontecimentos, obtém-se:

$$\tilde{G}(t) = \frac{\sum_{i=1}^{n} I(t_i \leq t)}{n}. \tag{4.16}$$

$$\hat{G}(t) = \frac{\sum_{i=1}^{n} I(e_i \geq t, t_i \leq t)}{\sum_{i=1}^{n} I(e_i \geq t)}. \tag{4.17}$$

4.2.8 Processo de Poisson Homogéneo

Como foi referido anteriormente, um dos pontos mais importantes no estudo dos processos pontuais espaciais é o da hipótese de aleatoriedade espacial completa. Este conceito relaciona-se com os processos de Poisson homogéneos na medida em que, sob essa hipótese, assume-se que o número de ocorrências de certo acontecimento numa região do plano segue uma distribuição de Poisson, com média proporcional à área da região e ao número médio de acontecimentos por unidade de área.

Considere-se A como uma região do plano e N como definido anteriormente, isto é, $N(A)$ é a variável aleatória que conta o número de pontos em A.

Definição 4.14 Um processo de Poisson (homogéneo) é caracterizado por duas propriedades fundamentais:

(PP1) Para toda região do plano limitada e fechada A, $N(A)$ tem distribuição de Poisson com média $\lambda|A|$, em que $|A|$ designa a área da região A ;

(PP2) Seja $N(A) = n$, os n acontecimentos em A formam um processo pontual identicamente distribuidos a n pontos independentes e uniformemente distribuídos em A;

(PP2) Para quaisquer duas regiões A e B disjuntas, as variáveis aleatórias $N(A)$ e $N(B)$ são independentes.

A constante λ refere-se ao número esperado de pontos por unidade de área e designa-se de intensidade do processo. Deste modo,

$$E[N(A)] = \lambda.|A|, \text{ para toda a região finita planar } A.$$

A propriedade (PP1) implica que a intensidade dos eventos não varia no plano.

A última propriedade implica que não existem interacções entre os acontecimentos.

Estes processos, daqui em diante serão, por vezes, referidos apenas como processos de Poisson, remetendo sempre para processos de Poisson homógeneos.

Proposição 4.1 Os processos de Poisson são estacionários e isotrópicos.

Proposição 4.2 Para um processo de Poisson com intensidade λ tem-se que $\mathrm{N(A)}\sim\mathrm{Poisson}(\lambda.|\mathrm{A}|)$ e portanto $\lambda(\mathrm{x}) = \lambda$ é a função intensidade de primeira ordem descrita acima.

Proposição 4.3 Para um Processo de Poisson tem-se $\lambda_2(\mathrm{x},\mathrm{y}) = \lambda_2(\mathrm{t}) = \lambda^2,\ \mathrm{t} > 0$.

Proposição 4.4 Para um processo de Poisson tem-se que

$$K(t) = \pi t^2, \qquad t > 0.$$

Considere-se $\boldsymbol{Y}$ como um processo pontual marcado em D e marcas em $\mathcal{M}$ e seja S o espaço dos pontos sem as marcas. Pode estabelecer-se, assim, as seguintes relações entre os processos pontuais marcados e os processos de Poisson:

- S é um processo de Poisson em D com intensidade μ e as marcas associadas a S são independentes e identicamente distribuídas com distribuição $Q \in \mathcal{M}$;
- $\boldsymbol{Y}$ é um processo de Poisson em $D \times \mathcal{M}$.

Num processo pontual espacial marcado, as translacções e as rotações actuam apenas sobre os pontos, mantendo as marcas inalteradas.

No caso de um processo de Poisson homogéneo, uma análise exploratória das propriedades de segunda ordem pode ser feita através do gráfico de $\widehat{K}(t)$ e correspondentes limites de erros baseados na sua variância.

Existem várias formas de se definir essa expressão da variância; considera-se aqui apenas uma delas, especialmente útil como limite de erros para janelas rectangulares. Seja o número de acontecimentos em W fixo e igual a n, a variância de $\widehat{K}(t)$ pode ser dada pela forma exacta de Lotwick e Silverman:

$$v_{LS}(t) = \left(\frac{n-1}{n^3}\right)|W|^2\big(2b(t) - a_1(t) + (n-2)a_2(t)\big), \tag{4.18}$$

com

$$b(t) = \pi t^2 |W|^{-1} \left(1 - \frac{\pi t^2}{|W|}\right) + \frac{1}{|W|^2}(1.0716Pt^3 + 2.2375t^4);$$

$$a_1(t) = \frac{1}{|W|^2}(0.21Pt^3 + 1.3t^4)\ ;$$

$$a_2(t) = \frac{1}{|W|^3}(0.24Pt^5 + 2.62t^6),$$

sendo P o perímetro de W.

Estes limites não são, no entanto, fiáveis quando o padrão se afasta muito da aleatoriedade espacial completa. Neste caso, pode dividir-se W em sub-regiões iguais e estimar $K(t)$ para cada sub-região, separadamente. No entanto, este método só se espera eficiente para pequenos valores de t.

Podem, também, comparar-se as estimativas das funções distribuição do espaço vazio e do vizinho mais próximo com a correspondente função teórica para um processo de Poisson, a fim de averiguar se existe homogeneidade espacial. Para um processo de Poisson homogéneo com intensidade λ, o número de pontos que está dentro de um círculo centrado em u e de raio t, $b(u,t)$, tem distribuição Poisson com média $\mu = \lambda\ |b(u,t)| = \lambda\pi t^2$. Logo, a probabilidade da distância de um dado ponto ao acontecimento que lhe está mais próximo ser inferior a t é dada por

$$F_{pois}(t) = 1 - \exp(-\lambda\pi t^2). \tag{4.19}$$

Quanto à função distribuição do vizinho mais próximo para estes processos, $G_{pois}(t)$, esta toma o mesmo valor que $F_{pois}(t)$. Intuitivamente, este resultado verifica-se porque os pontos são independentes uns dos outros num processo de Poisson e o conhecimento de um acontecimento não afecta os outros pontos do processo, logo G e F são equivalentes.

A hipótese de *aleatoriedade espacial completa*, apesar de ser um conceito idealizado, é usualmente tido como modelo nulo e é o ponto de partida para a construção de modelos realísticos para os padrões espaciais.

4.2.9 Processo de Poisson não homogéneo

Os processos de Poisson não-homogéneos surgem quando a intensidade λ de um processo de Poisson varia espacialmente, ou seja, para toda a localização x, a intensidade é dada por $\lambda(x)$.

Definição 4.15 Um processo pontual é um processo de Poisson não homogéneo se:

(PP1') Dada uma região A, o número de acontecimentos que nela ocorrem , $N(A)$, segue uma distribuição de Poisson com média $\int_A \lambda(x)dx$, com $\lambda(x)$ uma função não-negativa.

(PP2') Sejam $N(A) = n$, os n acontecimentos são independentes e identicamente distribuidos com função densidade de probabilidade proporcional a $\lambda(x)$.

(PP3') Dadas duas regiões A e B disjuntas, as variáveis aleatórias $N(A)$ e $N(B)$ são independentes.

4.2.10 Interacção entre pontos

Na modelação de um padrão pontual observado, se a hipótese de aleatoriedade completa for rejeitada, interessa averiguar se há algum tipo de dependência entre os pontos. Neste caso, procura-se modelar um padrão que seja regular ou agregado.

Classicamente, sugerem-se três principais classes de padrões pontuais que assentam em *independência* (os processos de Poisson), *regularidade* (em que os pontos tendem a afastar-se uns dos outros) ou *agregação* (em que os pontos tendem a estar próximos).

A Figura 4.2 representa cada um desses casos, considerando-se que a janela de observação W é de forma quadrangular.

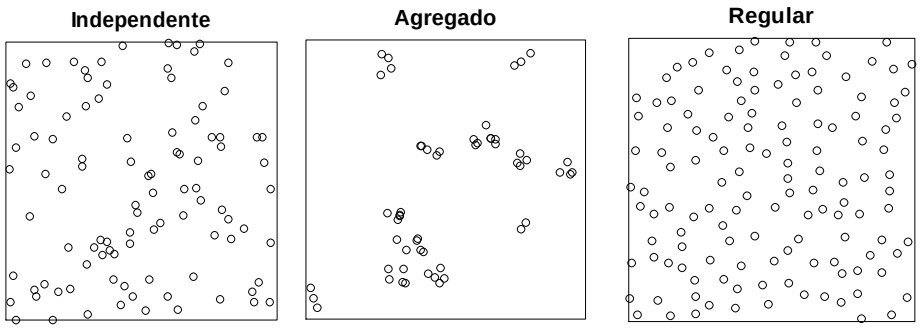


Figura 4.2 Simulações de padrões pontuais independente, agregado e regular.

4.3 Modelação e Inferência Estatística

Após uma análise exploratória dos dados, através da estimação das propriedades de primeira e segunda ordem e de outras descrições adicionais, é necessário formular um modelo estatístico e analisar o seu ajustamento aos dados. Numa primeira análise é importante verificar a hipótese de aleatoriedade espacial completa (CSR), onde os processos de Poisson homogéneo representam especial importância.

Na procura do modelo que melhor se ajuste aos dados, deve ter-se em conta quer a não-homogeneidade espacial, quer a dependência entre pontos ("interacção", onde se inclui a regularidade e aglomeração).

4.3.1 Testes à hipótese de aleatoriedade espacial completa

A hipótese de aleatoriedade espacial completa está, como já se viu, associada ao um processo de Poisson homogéneno. Deste modo, são feitos testes a esta hipótese usando estes processos como hipótese nula.

Estes testes baseiam-se, essencialmente, nas ferramentas que descrevem os padrões pontuais, tais como a função K e as funções distribuição do espaço vazio e do vizinho mais próximo.

Testes usando distâncias ao vizinho mais próximo permitem uma análise em pequena escala das interacções entre acontecimentos, visto analisarem-se as ocorrências que se encontram mais próximas no espaço. Assim, ao comparar-se a função G(t) com a correspondente função distribuição empírica (ou outra estimativa considerando os efeitos de fronteira), sob hipótese de aleatoriedade completa, se estas estiverem próximas, então a homogeneidade é aceite. O mesmo pode ser feito recorrendo à função distribuição do espaço vazio, $F(t)$.

A rejeição da hipótese de homogeneidade implica a procura de outros modelos para o ajustamento dos dados. Estes modelos podem estar associados, por exemplo, a covariáveis que tenham influência sobre a intensidade do processo.

Têm sido estudados vários métodos para a construcção destes testes, sendo o método de simulação de Monte Carlo o mais usado.

4.3.1.1 Testes da contagem de quadraturas

Um teste simples à hipótese de *CSR* é um teste de Qui-quadrado, baseado nas contagens observadas em quadraturas. Sendo W a janela de observação, o método consiste em dividir W em m sub-regiões ("quadraturas") de igual área e contar o número de pontos que se encontram dentro de cada uma delas, $n_i,\ \ i = 1, \dots, m$. Sob hipótese de *CSR*, n_j são variáveis aleatórias independentes e identicamente distribuídas com uma distribuição de Poisson com igual valor esperado n/m. Deste modo, aplica-se o teste de Qui-quadrado de Pearson e tem-se a estatística de teste

$$X^2 = \sum_{i=1}^{m} \frac{\left(n_i - \frac{n}{m}\right)^2}{\frac{n}{m}}. \qquad (4.20)$$

Sob hipótese nula, X^2 segue uma distribuição χ^2_{m-1}.

4.3.1.2 Testes usando a distribuição do vizinho mais próximo e do espaço vazio

Estes testes permitem estudar a hipótese de *CSR* por comparação da distribuição das distâncias $G(t)$ ou $F(t)$ sob hipótese de aleatoriedade completa com as correspondentes distribuições empíricas, $\tilde{G}(t)$ (5.19 ou 5.20) ou $\tilde{F}(t)$ (5.17 ou 5.18), respectivamente. Esta comparação pode ser feita graficamente, esperando obter-se linearidade no caso de valer a hipótese nula.

Uma abordagem muito útil na análise destas funções é através do método de simulação Monte Carlo.

O método aqui considerado é designado de teste de Monte Carlo simultâneo, dado que os invólucros superior e inferior são construídos simultaneamente.

Considerando o teste para a distribuição do vizinho mais próximo, $G(t)$, suponha-se que se consideram $(s-1)$ simulações de n acontecimentos independentes e distribuidos uniformemente em W, obtendo-se as funções estimadas $\hat{G}_i(t), i = 2, \dots s$ e $\hat{G}_1(t) = \tilde{G}(t)$. Para cada simulação, compara-se a curva simulada com a curva teórica e calcula-se o desvio máximo entre elas, tal que

$$D_i = \max_t |\hat{G}_i(t) - \tilde{G}(t)|, \; i = 2, \dots s.$$

Finalmente, obtém-se o valor máximo dos D_i,

$$D_{\max} = \max_i D_i,$$

e calculam-se os invólucros tal que

$$\begin{aligned} L(t) &= \tilde{G}(t) - D_{max} \\ G(t) &= \tilde{G}(t) + D_{max}. \end{aligned} \tag{4.21}$$

Neste caso, a função $\tilde{G}(t)$ excede os invólucros se o valor do desvio considerado para os dados exceder o valor D_{max}. Sob a hipótese de aleatoriedade completa, isto acontece com probabilidade $\alpha = \frac{1}{s}$, sendo este o tamanho do teste.

Considerando a função distribuição do espaço vazio, o procedimento é análogo ao anterior.

4.3.1.3 Testes usando a função K

Para testes de aleatoriedade completa usando a função K, basta considerar-se invólucros tal como descrito no teste anterior. Assim, após k simulações de aleatoriedade espacial completa da função K, essa hipótese não se rejeita se a função K empírica se inserir dentro desses invólucros.

4.3.2 Modelos

Como já referido, se a hipótese de aleatoriedade completa for verdadeira pode, então, descrever-se o padrão observado através de um processo de Poisson homogéneo. Este é, pois, o ponto de partida para a modelação de padrões pontuais espaciais.

No entanto, há casos em que ao ocorrer um acontecimento num dado local, a probabilidade de ocorrer um outro acontecimento próximo deste é maior, resultando um padrão *agregado*. Por outro lado, pode também suceder que um dado acontecimento é mais provável de estar rodeado por um espaço vazio, resultando num padrão *regular*. Ambos os cenários acometem a situações de dependência entre pontos, sendo vulgarmente descritas como "interacções" (Baddeley e Turner (2000)).

Comentário [R1]: Bibliografia- pasta 'Teoria' com este nome

4.3.3 Processos de Poisson agregados

Os processos de Poisson agregados foram introduzidos nos anos 50 por Neyman e Scott e estão relacionados com a modelação de padrões espaciais agregados.

Definição 4.16 Os processos de Poisson agregados são definidos por:

(PPA1) Acontecimentos pais formando um processo de Poisson homogéneo com intensidade ρ;

(PPA2) Cada pai produz um número aleatório de descendentes S, independente e identicamente distribuídas com distribuição de probabilidade p_S.

(PPA3) As posições dos descendentes relativamente aos seus pais são independentes e identicamente distribuidos com função densidade de probabilidade bivariada $h(.)$.

De seguida serão apresentadas algumas propriedades destes processos, que são úteis no estudo da modelação do padrão observado, permitindo, também, obter estimativas preliminares dos parâmetros.

Proposição 4.5 Os processos de Poisson agregados são estacionários com intensidade dada por $\lambda = \rho\mu$, em que $\mu = E[S]$. A isotropia vale no caso de, em (PPA3), a função densidade probabilidade ser radialmente simétrica.

Proposição 4.6 A função K de um processo de Poisson agregado é dada por

$$K(t) = \pi t^2 + \frac{E[S(S-1)]H_2(t)}{\rho\mu^2}, \tag{4.22}$$

sendo $H_2(t)$ é a função distribuição do vector da diferença entre as posições de dois descendentes do mesmo pai.

Tem-se, ainda, que o segundo termo de (4.22) é não-negativo e monótono não-decrescente, e que $K(t) - \pi t^2$ se aproxima de uma constante c, quando $t \to +\infty$. Essa constante é dada por

$$c = \frac{E[S(S-1)]}{\rho\mu^2}.$$

Proposição 4.7 A intensidade de segunda ordem é dada por:

$$\lambda_2(t) = \lambda^2 + \rho E[S(S-1)]h_2(t), \tag{4.23}$$

em que $h_2(t)$ é a função densidade de probabilidade do vector diferença referido no ponto anterior.

A Figura 4.3 representa uma realização de um processo de Poisson agregado, considerando 20 pais em média por unidade quadrado e 5 descendentes em média por cada pai, com um raio de cada aglomeração não excedendo 0.05.

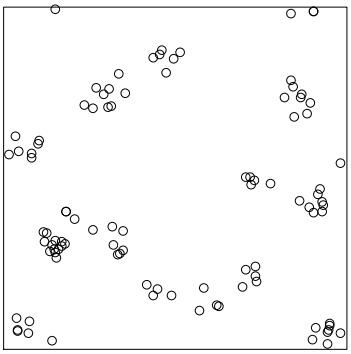


Figura 4.3 Realização de um processo de Poisson agregado.

Comentário [R2]: retirado de Baddeley R

4.3.4 Processos de Cox

Estes processos são uma classe de modelos que derivam dos processos de Poisson não homogéneos com funções intensidade estocásticas.

Definição 4.17 Um processo de Cox X é definido pelas seguintes propriedades:

1. $\{\Lambda(x): x \in \mathbb{R}^2\}$ é um processo estocástico que toma apenas valores não negativos

2. Para X condicional apenas a uma realização de Λ(x), com $\{\Lambda(x) = \lambda(x): x \in \mathbb{R}^2\}$, os acontecimentos formam um processo de Poisson não homogéneo com função de intensidade $\lambda(x)$.

Proposição 4.8 A intensidade de segunda ordem de um processo de Cox é dada por

$$\lambda_2(x, y) = E[\Lambda(x)\Lambda(y)]. \tag{4.24}$$

Proposição 4.9 Um processo de Cox **X** é estacionário ou isotrópico se e só se o processo das intensidades Λ(x) é estacionário ou isotrópico, respectivamente.

Proposição 4.10 No caso de um processo de Cox estacionário, a intensidade é $\lambda = \mathrm{E}[\Lambda(\mathrm{x})]$. No caso de estacionariedade e isotropia, a intensidade de segunda ordem é dada por

$$\lambda_2(\| x - y \|) = \lambda^2 + Cov\big(\Lambda(x), \Lambda(y)\big). \tag{4.25}$$

Proposição 4.11 A função K de um processo de Cox define-se recorrendo a (5.10), tal que

$$\lambda K(t) = \frac{2\pi}{\lambda} \int_0^t \lambda_2(x)x \, dx.$$

Uma classe especialmente útil quando se pretende modelar a superfície da intensidade com recurso a covariáveis é o dos processos de Cox log-Gaussianos. Neste caso, modela-se o logaritmo do

processo de intensidade recorrendo a covariáveis e acrescentando-se um erro que é um processo gaussiano.

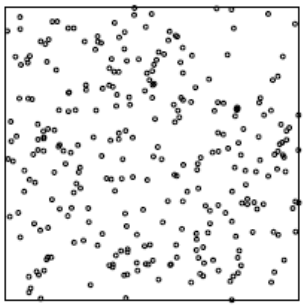


Figura 4.4 Realização de um processo de Cox.

4.3.5 Outros processos

Outros processos são os processos de inibição simples, para modelar padrões mais regulares quando o padrão de aleatoriedade espacial completa não é satisfeito devido a condições impostas. É o caso, por exemplo, quando existem árvores a competir entre si por luz, impondo-se uma distância mínima admissível.

Os processos pontuais de Markov servem para modelar padrões agregados ou regulares.

4.4 Ajustamento de modelos

Nos processos pontuais espaciais, recorre-se a métodos de inferência baseados nas medidas resumo teóricas, bem como outros métodos não-paramétricos. O uso destes métodos deve-se, essencialmente, à dificuldade em trabalhar com as funções de verosimilhança dos modelos de interesse, devido à sua complexidade.

4.4.1 Usando a função $K(t)$

Um dos modelos de inferência consiste no uso da função $K(t)$, cujo valor teórico é conhecido para alguns casos. A estimação desta função, $\widehat{K}(t)$, pode, então, ser comparada com o seu valor teórico, $K(t;\boldsymbol{\theta})$**.** O valor estimado $\widehat{\boldsymbol{\theta}}$ escolhido é o que minimiza o valor de $\boldsymbol{\theta}$ para a discrepância entre as duas funções:

$$D(\boldsymbol{\theta}) = \int_0^{t_0} w(t)\left(\left(\widehat{K}(t)\right)^c - \left(K(t;\boldsymbol{\theta})\right)^c\right)^2 dt, \tag{4.26}$$

escolhendo-se adequadamente as constantes t_0 e c e a função $w(t)$ dos pesos.

A distribuição do parâmetro $\widehat{\boldsymbol{\theta}}$ pode ser, também, obtida através do método de Monte Carlo.

Se, no entanto, não existir uma forma explícita ou numérica da função $K(t,\boldsymbol{\theta})$, pode substituir-se a mesma em $D(\boldsymbol{\theta})$ pela média de *s* suas estimativas pontuais através da simulação de *s* realizações do modelo que se está a tratar.

4.4.2 Por maximização da função verosimilhança

Como já referido, a função de verosimilhança para grande parte dos modelos dos processos pontuais espaciais é de tratamento matemático difícil.

No entanto, se se considerar um processo de Poisson não homogéneo com função intensidade $\lambda(t)$, a função de verosimilhança assume uma forma de fácil manipulação. Assim, seja $\boldsymbol{x} = \{x_1, x_2, \dots, x_n\}$ uma realização de um processo desta natureza numa região finita A, a correspondente função log-verosimilhança é dada por:

$$L(\lambda) = \sum_{i=1}^{n} \log\big(\lambda(x_i)\big) - \int_A \lambda(x)dx. \tag{4.27}$$

Esta função resulta da factorização da região A no produto de uma distribuição Poisson com média $\mu = \int_A \lambda(x)dx$ para o número de acontecimentos n com um conjunto de localizações x_i independentes com densidade $\frac{\lambda(x)}{\mu}$.

Uma implementação expedita na estimação da máxima verosimilhança numericamente, foi proposta por Berman e Turner em 1992, usando programação adequada que se baseia num modelo linear generalizado com respostas Poisson. Neste caso, a função log-verosimilhança é aproximada por

uma soma finita idêntica à log-verosimilhança de um modelo linear generalizado com respostas Poisson.

Para outros processos, uma forma de ultrapassar o problema da complexidade da função verosimilhança é através do recurso à função pseudo-verosimilhança (Diggle, 2003).

4.5 Acidentes rodoviários com vítimas em Lisboa

Neste capítulo analisar-se-á o padrão de ocorrências dos acidentes com vítimas em Lisboa, entre 2004 e 2007, através dos processos pontuais espaciais. Esta análise foi efectuada recorrendo ao pacote do software estatístico **R**, *spatstat* (Baddeley & Turner, 2005, 2006), que permite manipular, criar gráficos de padrões espaciais, analisar e ajustar modelos e simular modelos de processos pontuais.

Os acidentes foram georreferenciados, permitindo, assim, fazer-se uma análise espacial dos mesmos. No entanto, não foi possível obter essa georeferenciação para 1025 acidentes, que corresponde a cerca de 11% da amostra inicial que foram excluidos da análise. Considera-se que a amostra dos restantes 8238 acidentes permite análises potencialmente conclusivas.

Comentário [WU3]: Confirmar!

Para uma análise mais precisa, considerou-se ainda que os acidentes estavam limitados não apenas ao concelho de Lisboa, mas às próprias vias de trânsito. Desta forma, considerou-se como janela de observação um polígono compacto contendo todas as vias de trânsito com uma certa largura para as mesmas (Figura 4.5).

Inicialmente é feita uma análise espacial separada por anos (2004, 2005, 2006 e 2007) e, posteriormente, considerando-se o conjunto de todos os anos, por forma a reduzir o efeito aleatório específico de cada ano.

Figura 4.5 Janela de observação formada pelo polígono compacto fechado que delimita as artérias da cidade de Lisboa

4.5.1 Análise espacial anual dos acidentes

Nesta secção será feita uma análise espacial dos acidentes considerando-se cada ano separadamente.

Na Tabela 4.1 é indicado o número de acidentes georreferenciados em cada ano.

Tabela 4.1 Número de acidentes georreferenciados, por ano

Ano	2004	2005	2006	2007
Nº de acidentes	2040	2108	2143	1947

As Figura 4.6 a Figura 4.9 representam a localização dos acidentes, limitada pela correspondente janela de observação considerada, para cada ano.

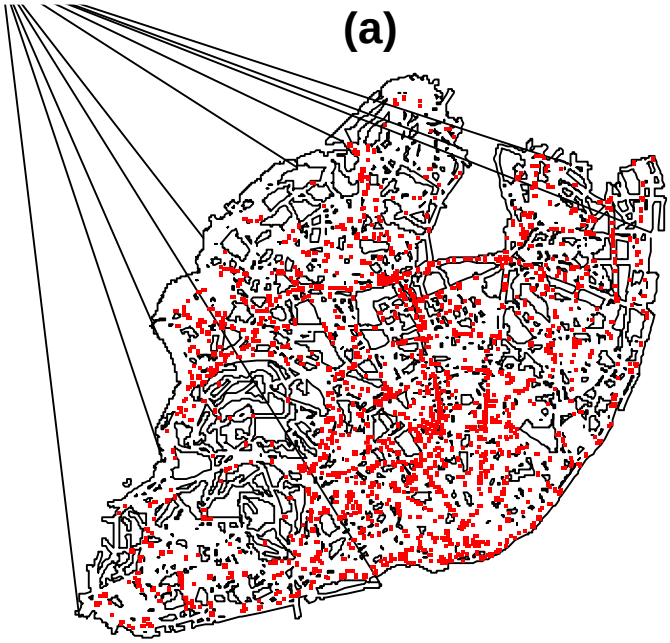


Figura 4.6 Localização dos acidentes na cidade de Lisboa no ano de 2004.

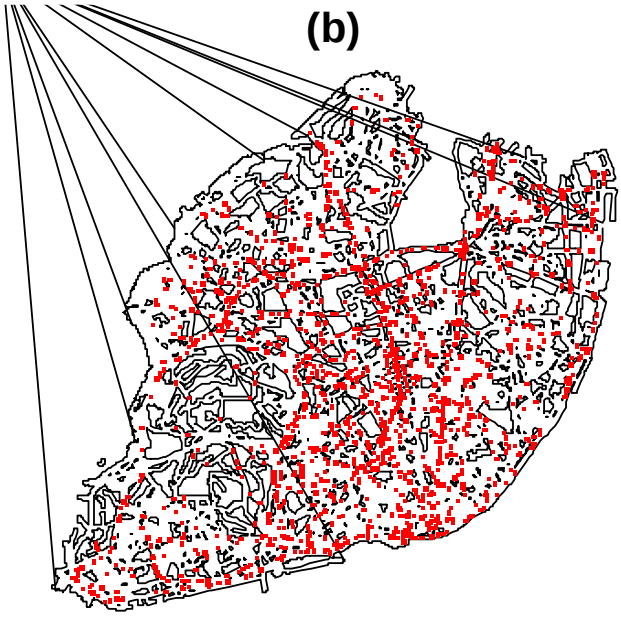


Figura 4.7 Localização dos acidentes na cidade de Lisboa no ano de 2005.

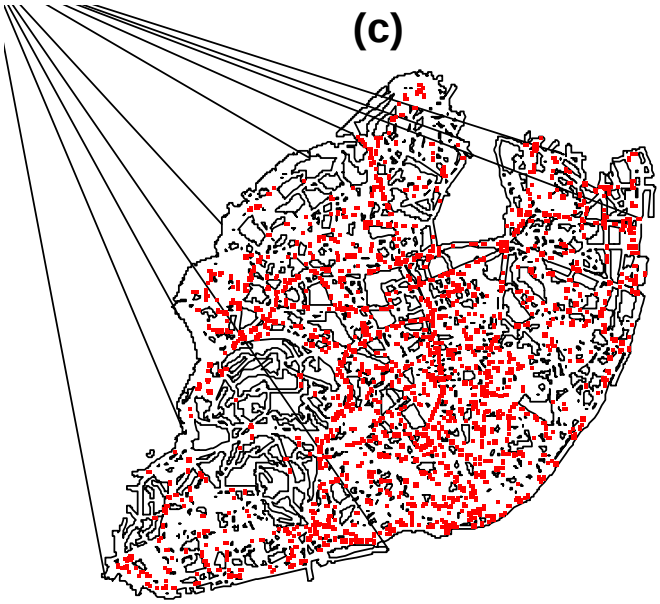


Figura 4.8 Localização dos acidentes na cidade de Lisboa no ano de 2006.

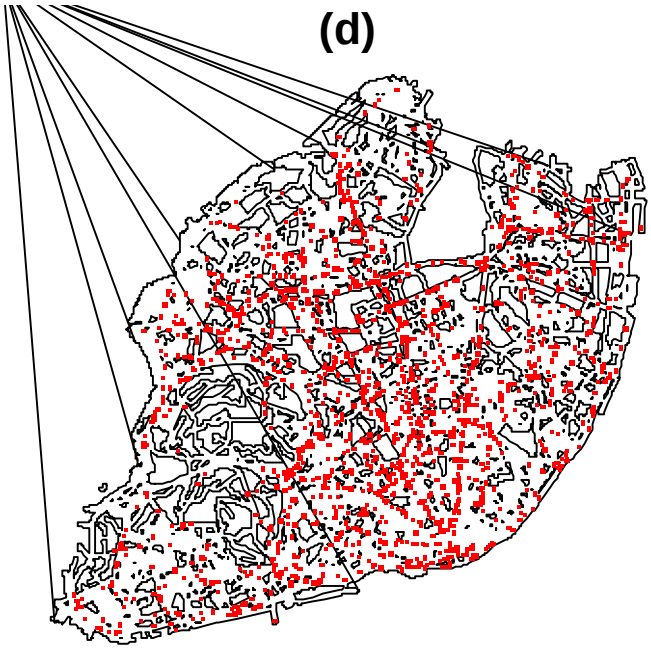


Figura 4.9 Localização dos acidentes na cidade de Lisboa no ano de 2007.

Analisando as figuras, parece verificar-se que a localização dos acidentes não varia muito de acordo com os anos, havendo uma tendência para ocorrerem mais na zona Centro Este de Lisboa.

Esta facto pode ser analisado por uma avaliação da homogeneidade espacial, através de gráficos das estimativas kernel da intensidade dos padrões pontuais observados (Diggle, 2003), em cada um dos anos (Figura 4.10 a Figura 4.13).

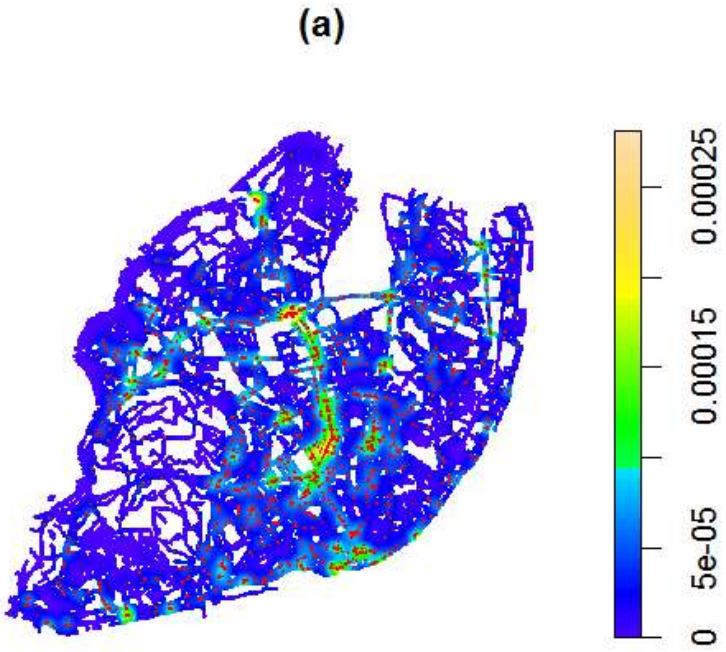


Figura 4.10 Estimativa kernel da intensidade dos padrões espaciais observados e respectiva localização da ocorrência dos acidentes na cidade de Lisboa (a vermelho) no ano de 2004.

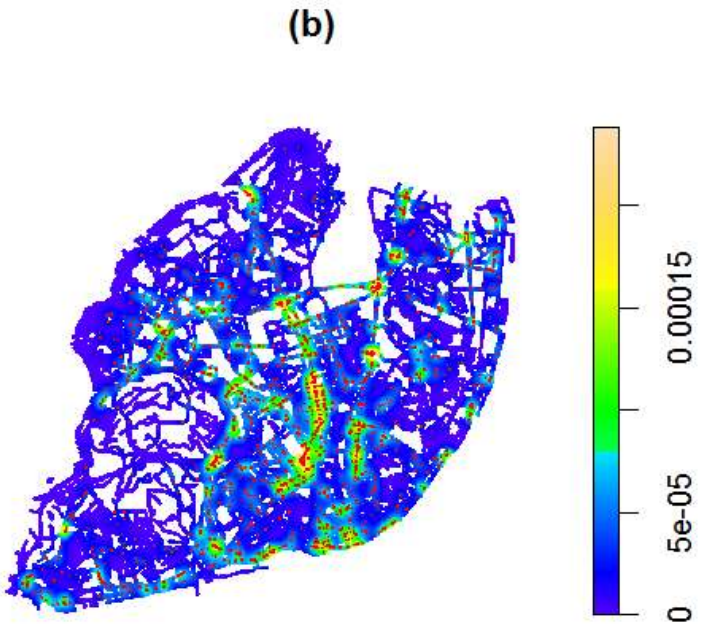


Figura 4.11 Estimativa kernel da intensidade dos padrões espaciais observados e respectiva localização da ocorrência dos acidentes na cidade de Lisboa (a vermelho) no ano de 2005.

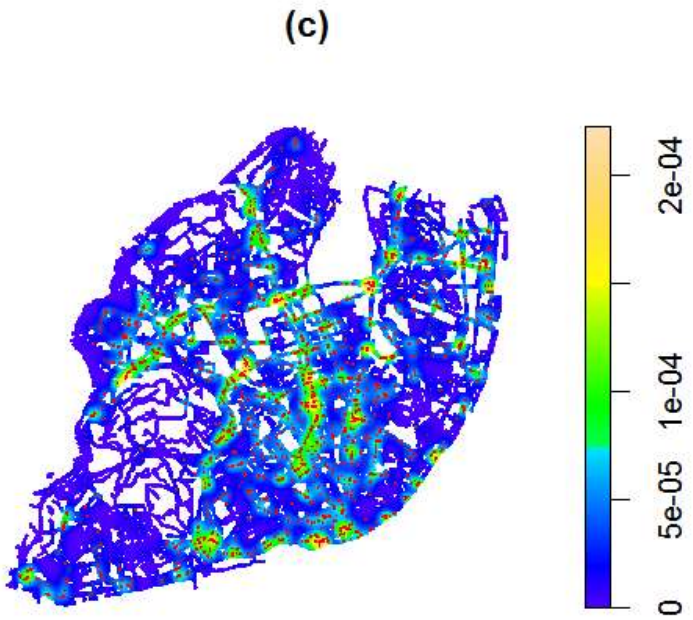


Figura 4.12 Estimativa kernel da intensidade dos padrões espaciais observados e respectiva localização da ocorrência dos acidentes na cidade de Lisboa (a vermelho) no ano de 2006.

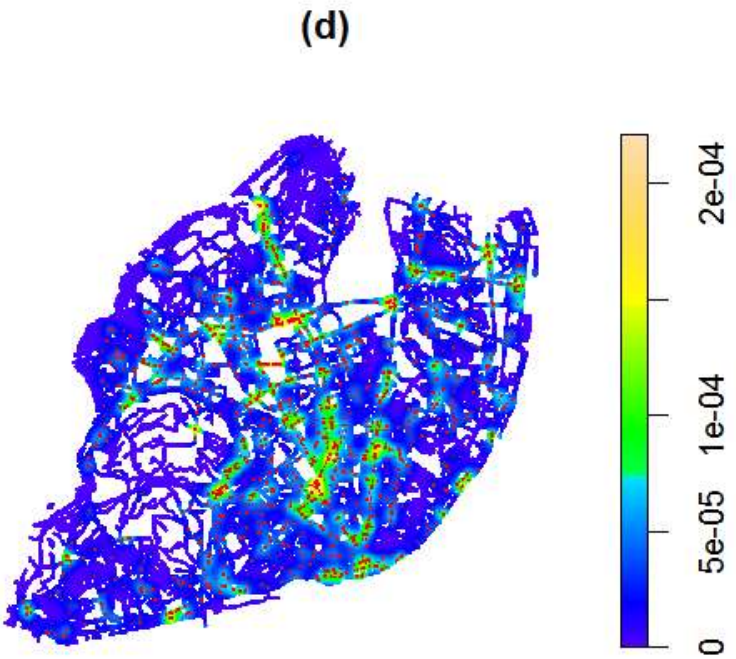


Figura 4.13 Estimativa kernel da intensidade dos padrões espaciais observados e respectiva localização da ocorrência dos acidentes na cidade de Lisboa (a vermelho) no ano de 2007.

Pelas figuras acima verifica-se, de facto, uma concentração dos acidentes em certas zonas (tons mais próximo do amarelo), parecendo haver falta de homogeneidade na distribuição espacial da ocorrência dos mesmos, o que limita a análise exploratória relativamente às estatísticas resumo padrão. Interessa, assim, verificar se a heterogeneidade existente resulta de algum factor externo, como por exemplo o tráfego, e a partir daí tentar estimar um modelo que se ajuste aos dados.

Assumindo homogeneidade, a estimativa da intensidade das ocorrências dos acidentes por unidade de área (Equação (4.6)) é dada na Tabela 4.2, verificando-se pouca variabilidade ao longo dos anos, apesar de em 2007 apresentar um valor um pouco mais baixo.

Tabela 4.2 Intensidade por unidade de área, por ano.

Ano	2004	2005	2006	2007
Intensidade	3.03×10^{-5}	3.13×10^{-5}	3.18×10^{-5}	2.89×10^{-5}

Seguidamente, será feita uma análise à hipótese de aleatoriedade espacial, usando os métodos descritos anteriormente.

4.5.1.1 Testes à hipótese da aleatoriedade espacial completa

Numa primeira análise menos formal de hipótese de aleatoriedade completa, foram analisadas graficamente as estimativas das funções K e as distribuições do espaço vazio e do vizinho mais próximo. Pretende-se comparar as estimativas destas funções com os respectivos valores teóricos sob a hipótese *CSR*.

A

Figura 4.14 apresenta os gráficos das estimativas da função K, tendo em conta a correcção de fronteira de acordo com o método da fronteira (secção 4.2.2), contra a correspondente função teórica sob validade da hipótese de aleatoriedade espacial completa (a tracejado), por ano. Esta é uma forma de verificar a validade de hipótese de *CSR*, ao verificar se as duas funções se afastam muito uma da outra.

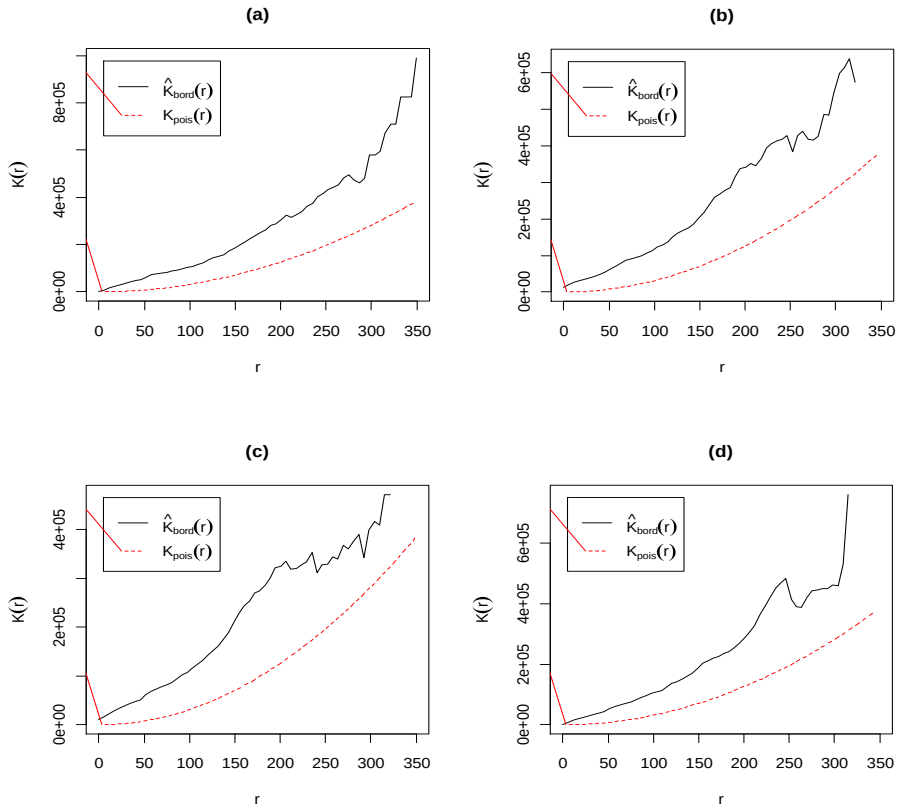


Figura 4.14 $\widehat{K}(t)$, (linha sólida), com correcção de fronteira, e correspondente função teórica sob hipótese de CSR ($K(t) = \pi t^2$), nos anos de (a) 2004, (b) 2005, (c) 2006 e (d) 2007.

A análise dos gráficos leva-nos a considerar a rejeição da hipótese de aleatoriedade espacial completa em todos os anos, dada a discrepância entre as duas funções. Esta discrepância incide num valor sobre-estimado da função K relativamente ao seu valor teórica sob hipótese de homogeneidade, remetendo para um padrão pontual agregado, como consequência da

Definição 4.13.

Na Figura 4.15 apresentam-se os gráficos das estimativas da função K e a correspondente função teórica, assumindo que há falta de homogeneidade. Mais uma vezes foi considerado o método da fronteira para correcção dos efeitos de fronteira. Note-se que na construção desta função, a função intensidade foi estimada a partir das estimativas alisadas kernel da intensidade do padrão pontual observado (Baddeley & Turner, 2006, p. 150).

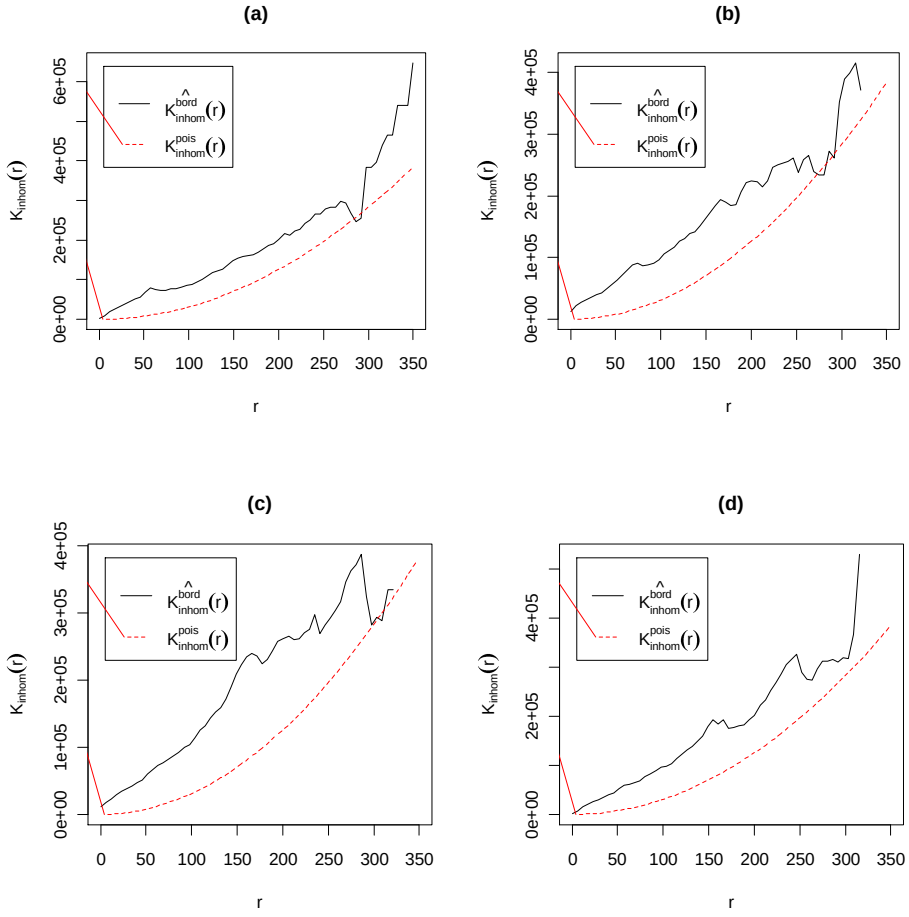


Figura 4.15 Estimativa da função K (linha sólida) e correspondente função teórica (linha a tracejado), assumindo falta de homogeneidade, nos anos de (a) 2004, (b) 2005, (c) 2006 e (d) 2007.

Os gráficos acima sugerem que há, de facto, uma melhor aproximação da função K estimada com a correspondente função teórica quando se assume falta de homogeneidade.

Também as estimativas das funções distribuição do espaço vazio e do vizinho mais próximo foram estimadas e contrastadas, graficamente, com os respectivos valores teóricos, sob hipótese de aleatoriedade espacial completa (Figura 4.16 e Figura 4.17). Considerou-se, mais uma vez, o método da fronteira para corrigir os efeitos fronteira.

A Figura 4.16 apresenta as estimativas das funções distribuição do espaço vazio, para cada ano, com as correspondentes funções teóricas sob hipótese de aleatoriedade completa, $F(t) = 1 - \exp(-\lambda\pi t^2)$.

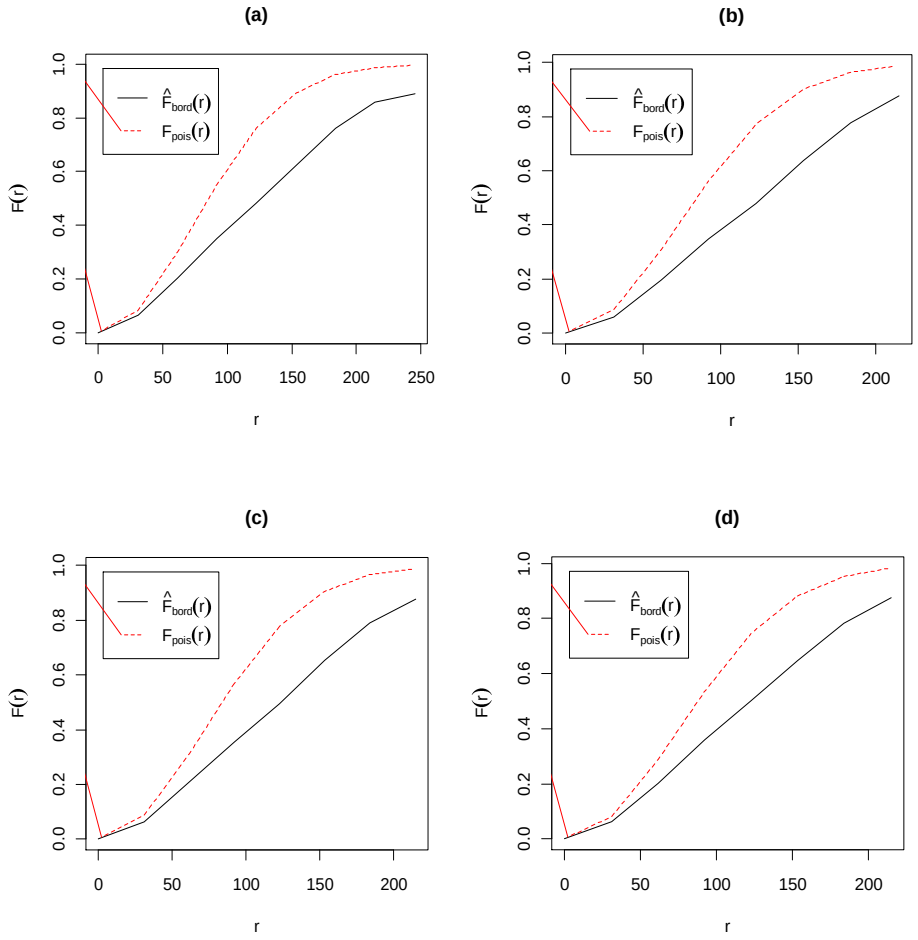


Figura 4.16 Estimativas das funções distribuição de F com correcção de fronteira (a cheio), com a correspondente função teórica sob hipótese de CSR, (a tracejado), para os anos de (a) 2004, (b) 2005, (c) 2006 e (d) 2007.

Verifica-se que há uma discrepância notável entre as funções, levando a concluir, novamente, que não há homogeneidade dos dados.

Ao analisar-se a função distribuição do vizinho mais próximo, as conclusões mantêm-se, tal como se pode verificar pela Figura 4.17.

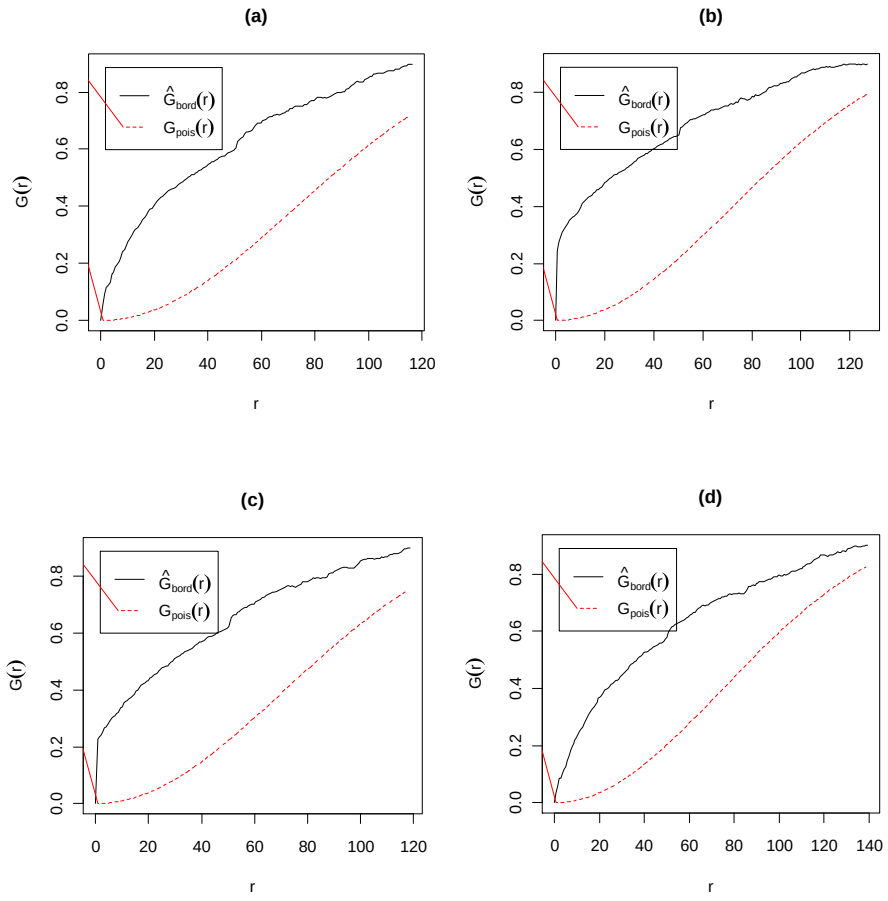


Figura 4.17 Estimativas das funções distribuição de G com correcção de fronteira (a cheio), com a correspondente função teórica sob hipótese de CSR, (a tracejado) para os anos de (a) 2004, (b) 2005, (c) 2006 e (d) 2007.

Teste da contagem de quadraturas

Como se viu na Secção 4.3.1.1, um dos testes que se pode realizar à hipótese *CSR* é o teste da contagem de quadraturas.

Para a aplicação do teste, a janela de observação foi dividida em 23 sub-regiões, em cada uma delas foi aplicado o teste de Qui-quadrado de Pearson. O teste resultou nas estatísticas e teste e valores-p indicados na Tabela 4.3

Tabela 4.3 Valores observados da estatística de teste χ^2 e correspondentes valores-p, por ano.

	2004	2005	2006	2007
χ^2_{obs}	424.31	441.89	394.22	353.42
Valor-p	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

Este teste sugere uma forte rejeição da hipótese de aleatoriedade completa, visto os *valores-p* resultantes do teste de Qui-quadrado de Pearson serem muito pequenos.

Este teste será utilizado novamente mais tarde, considerando-se uma divisão de regiões assente no tráfego rodoviário.

Testes baseados no método de Monte Carlo

Como referido nas secções 4.3.1.2 e 4.3.1.3, podem ser feitos testes à hipótese de aleatoriedade completa usando o método de Monte Carlo. Este método será aplicado às funções do espaço vazio e do vizinho mais próximo.

Relativamente à função distribuição do espaço vazio, foram calculados os invólucros superior e inferior, de acordo com a equação (4.21), com base em 99 simulações. Desta forma, obteve-se um teste de tamanho $\alpha = 1\%$. O teste rejeita a hipótese nula se o gráfico da função observada excede o invólucro para todo o valor de r. A Figura 4.18 representa as funções espaço vazio com correcção de fronteira pelo método da fronteira (Secção 4.2.2) e os correspondentes invólucros superior e inferior, contra as distribuições teóricas sob hipótese de *CSR*.

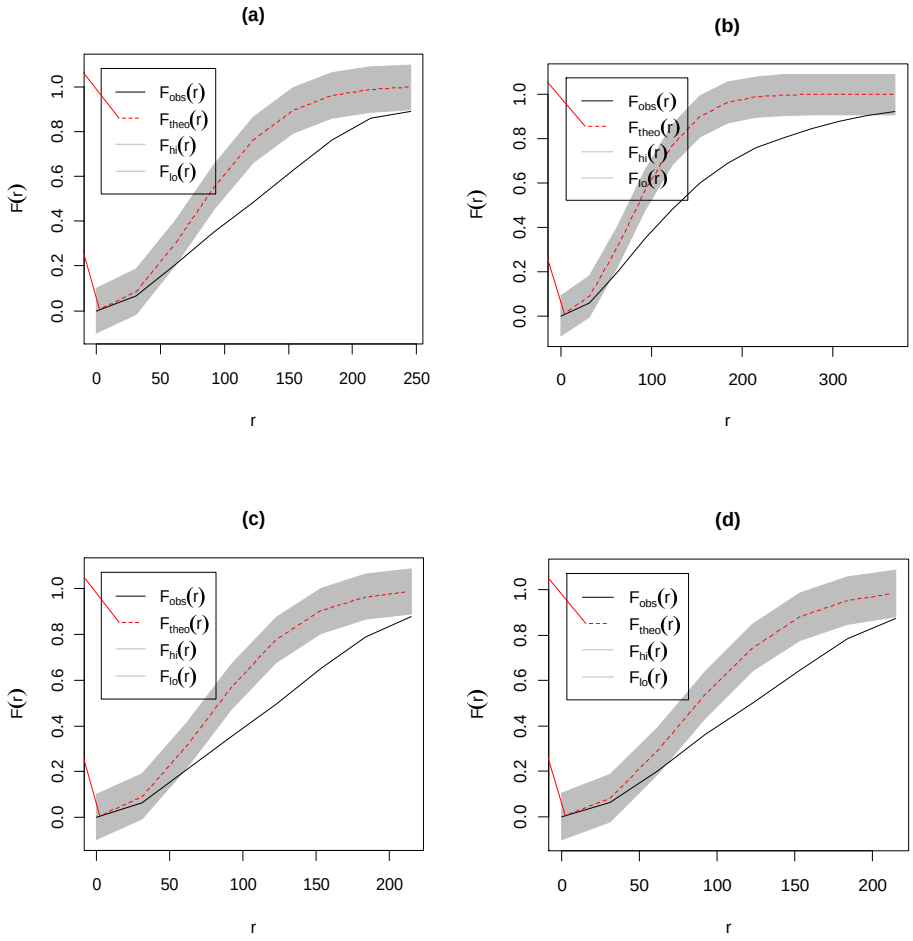


Figura 4.18 Estimativas da função F (a cheio) e invólucros, contra a correspondente distribuição teórica sob hipótese de CSR (a tracejado), nos anos de (a) 2004, (b) 2005, (c) 2006 e (d) 2007.

Conclui-se que se rejeita a hipótese de aleatoriedade espacial completa ao nível de 1%, apesar de para pequenos valores de r a função ainda se encontrar dentro dos invólucros. No entanto, a função começa a ter um desvio acentuado para valores os restantes valores. Logo, há fortes evidências para rejeitar a homogeneidade espacial através deste teste, como tem sido concluido até agora.

A mesma análise foi realizada com a função distribuição do vizinho mais próximo, mas com um tamanho de teste $\alpha = 5\%$, obtendo-se os gráficos da Figura 4.19. Conclui-se, novamente, que parece não haver homogeneidade espacial dos dados.

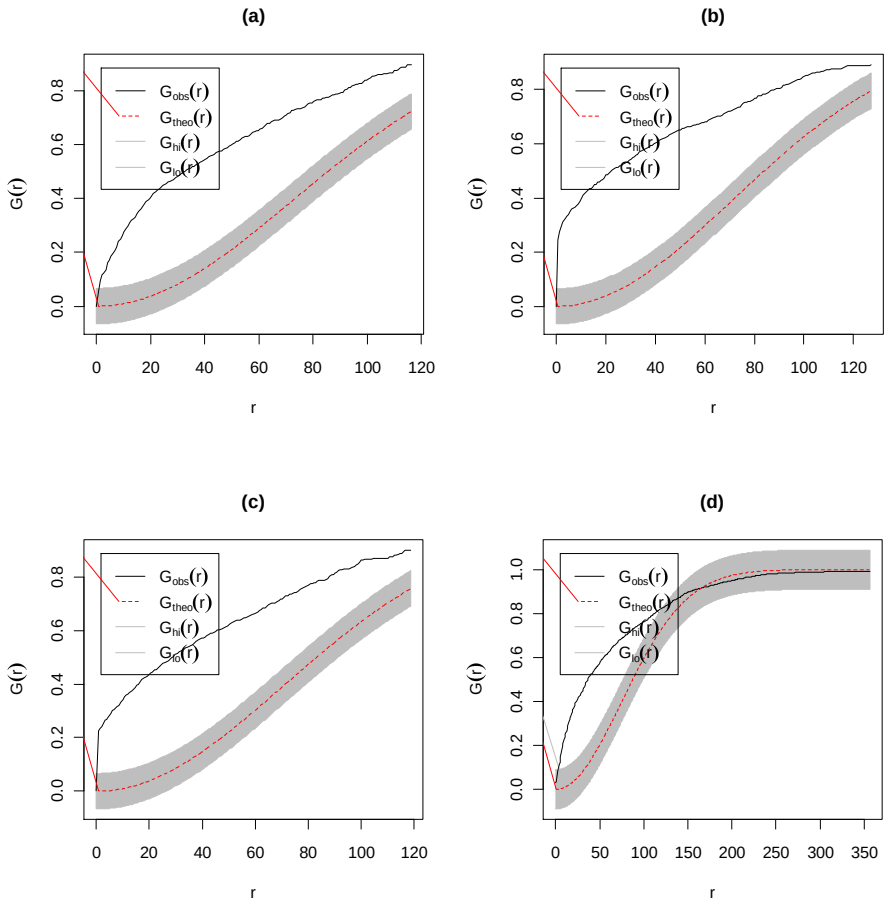


Figura 4.19 Estimativas da função G (a cheio) e invólucros, contra a correspondente distribuição teórica sob hipótese de CSR (a tracejado), nos anos de (a) 2004, (b) 2005, (c) 2006 e (d) 2007.

4.5.1.2 Análise da influência do tráfego na falta de homogeneidade espacial dos acidentes rodoviários

Como já referido para a análise envolvendo os modelos lineares generalizados de Poisson (Capítulo 3), o tráfego é uma varivável importante quando se pretende analisar o número de acidentes ocorridos. Desta forma, e sendo uma variável espacial, foi considerada nesta análise de padrões

espaciais. Pretende-se verificar se a falta de homogeneidade provém do tráfego e, deste modo, relacionar a função intensidade do processo não homogéneo com esta variável.

A variável tráfego está definida para cada localização da actual janela de observação, correspondendo ao Tráfego Médio Diário de dia Útil (TMDU)[5], cuja unidade é veículos ligeiros equivalentes, no ano de 2008. Existe um desfazamento temporal entre os dados de acidentes (2004 a 2007) e os do tráfego (2008) rodoviários em análise. Os dados de tráfego rodoviário são frequentemente inexistentes, pelo que se considera os dados de 2008 uma *proxy* para o tráfego dos anos que lhe antecedem.

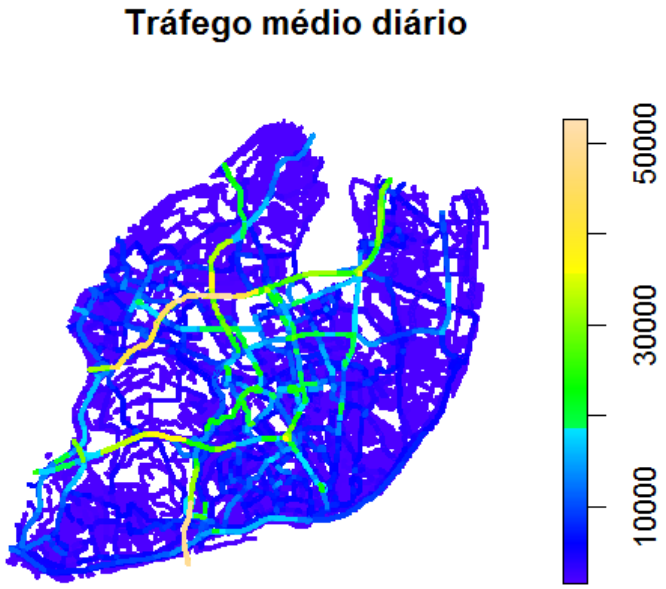


Figura 4.20 Tráfego Médio Diário em dia Útil em Lisboa em 2008.

De forma a normalizar o tráfego quanto à sua variabilidade, obtém-se a seguinte medida do mesmo:

[5] O Tráfego Médio Diário de dia Útil é uma medição do tráfego e a sua unidade é em veículos ligeiros equivalente, i.e., $TMDU(vle)$. Esta unidade foi modelada com um factor de 2.5, tal que, por exemplo, 10 veículos ligeiros e dois veículos pesados correspondem a $15vle = 10(\text{ligeiros}) + 2.5 \times 2(\text{pesados})$.

$$TMDU_{std} = \frac{TMDU}{S(TMDU)}. \quad (4.28)$$

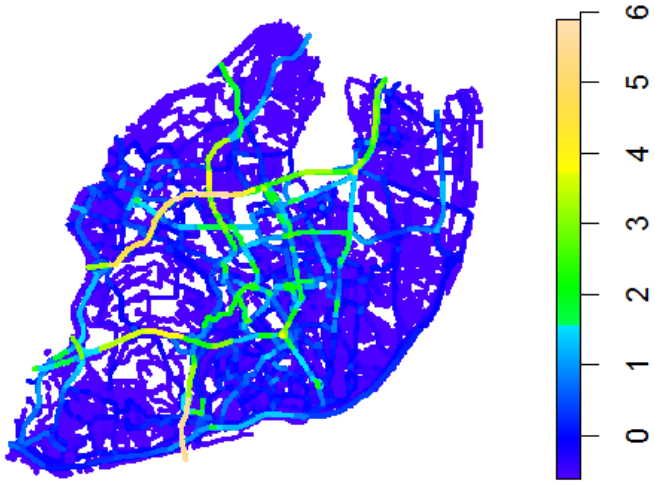


Figura 4.21 TMDU normalizado.

Os valores de tráfego com que se realizarão as análises seguintes é o normalizado (Figura 4.21) que será referido simplesmente como "tráfego" (apesar de não corresponder à definição formal do mesmo).

Tal como mencionado na secção anterior, uma forma de analisar se esta variável contribui para a explicação da distribuição espacial dos acidentes, é através do método da contagem de quadraturas. A diferença agora reside na divisão das regiões, que é feita pelos níveis desta variável. Assim, dividiu-se o tráfego em cinco partes iguais, obtendo-se o seguinte esquema dos pontos:

Tabela 4.4 Divisão do tráfego em cinco partes e correspondentes número de pontos em cada divisão da janela.

Tráfego	**]0.197, 1.5]**	**]1.5, 2.8]**	**]2.8, 4.1]**	**]4.1, 5.4]**	**]5.4, 6.7]**
2004	1184	498	236	62	60

2005	1187	554	241	70	56
2006	1231	536	238	76	62
2007	1174	467	200	62	44

Aplicando, agora, o teste de Qui-quadrado de Pearson para verificar a independência de cada uma das divisões da janela, obtêm-se os seguintes resultados:

Tabela 4.5 Valores observados da estatística de teste χ^2 e correspondentes valores-p, por ano.

	2004	2005	2006	2007
χ^2	1206.286	1368.396	1295.072	933.2061
Valor-p	< 2.2e-16	< 2.2e-16	< 2.2e-16	< 2.2e-16

Verifica-se, portanto, que o valor-p é muito pequeno, indicando a rejeição da hipótese nula de homogeneidade espacial. Não se poderá, contudo, concluir que se aceita o facto de a heterogeneidade espacial poder ser explicada pelo tráfego, apesar de não se rejeitar esta hipótese (alternativa).

Ao assumir-se que a intensidade do padrão observado é uma função da covariável tráfego, então tem-se

$$\lambda(x) = \rho(Z(x)), \tag{4.29}$$

com $Z(x)$ o valor do tráfego em toda a localização u da janela de observação, e ρ uma função que interessa investigar, indicando como a intensidae depende do valor da covariável.

Na estimação da função ρ podem ser usadas as estimativas suavizadas kernel, usando métodos do risco relativo, como explicado em (Baddeley, Chang, Song, & Turner, Submitted for publication.).

A figura abaixo representa a estimativa de intensidade como função do tráfego.

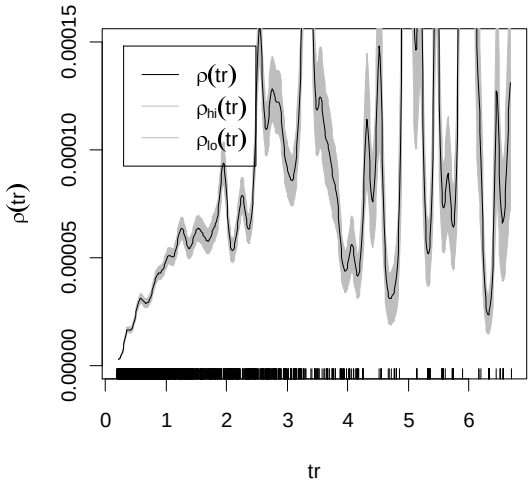


Figura 4.22 Estimativa da intensidade $\rho(tr)$ como função do tráfego (tr).

Analisando o gráfico da Figura 4.22 que representa uma estimativa para a função intensidade ρ, não se consegue inferir facilmente sobre a possível relação entre a mesma e a variável tráfego, embora o gráfico sugira uma tendência cúbica.

4.5.1.3 Modelação e Inferência

Após uma análise exploratória dos dados, onde se analisou a hipótese de aleatoriedade espacial completa (secção 4.5.1.1) e a sua relação da sua falta de homogeneidade com o tráfego rodoviário (secção 4.5.1.2), interessa formular um modelo que explique a distribuição espacial dos dados, isto é, acidentes rodoviários em Lisboa. Da análise inicial concluiu-se que os dados não estão distribuídos de forma homogénea no espaço, o que leva ao recurso a modelos não-homogéneos.

Deste modo, os modelos aqui considerados serão processos de Poisson não-homogéneos.

Para o processo de Poisson não-homogéneo, será considerada a variável do tráfego na modelação da intensidade do processo, de forma a averiguar se, de facto, há uma relação entre a mesma e a localização dos acidentes.

Neste caso, o modelo nulo considerado será o Processo de Poisson homogéneo, que traduz a hipótese de homogeneidade (*CSR)*, que servirá de comparação com os anteriores.

Os modelos propostos referentes à situação de não homogeneidade consideram a intensidade como:

1. Função log-linear do tráfego;
2. Função log-quadrática no tráfego;
3. Função log-cúbica no tráfego e
4. Função linear ao tráfego (proporcionalidade).

- *Modelo com log-intensidade linear no tráfego*

Neste caso, ajusta-se um modelo cuja intensidade é log-linear no tráfego, tal que

$$\lambda(x) = \exp(\beta_0 + \beta_1 Z(x)), \tag{4.30}$$

com β_0 e β_1 os parâmetros e $Z(x)$ o tráfego para cada localização x da janela de observação.

Este modelo foi implementado para cada um dos anos, resultanto nos seguintes valores estimados dos parâmetros (Tabela 4.6):

Tabela 4.6 Parâmetros estimados do modelo (5.28) e erros padrão associados, para cada ano.

Ano	β_0	Erro padrão β_0	β_1	Erro padrão β_1
2004	-10.913	0.030	0.416	0.0304
2005	-10.890	0.030	0.415	0.012
2006	-10.856	0.029	0.409	0.012
2007	-10.891	0.031	0.377	0.013

- *Modelo com log-intensidade quadrática no tráfego*

Nesta hipótese de se considerar a intensidade do processo como função quadrática do tráfego, o modelo ajustado é da forma

$$\lambda(x) = \exp(\beta_0 + \beta_1 Z(x) + \beta_2 Z^2(x), \quad (4.31)$$

com β_0, β_1 e β_2 os parâmetros e $Z(x)$ o tráfego na localização x.

A Tabela 4.7 apresenta o valor desses parâmetros, considerando o modelo ajustado para cada um dos anos.

Tabela 4.7 Parâmetros estimados do modelo (4.31) e erros padrão associados, para cada ano.

Ano	β_0	Erro padrão β_0	β_1	Erro padrão β_1	β_2	Erro padrão β_2
2004	-11.553	0.047	1.397	0.049	-0.191	0.010
2005	-11.553	0.046	1.421	0.049	-0.195	0.009
2006	-11.429	0.045	1.289	0.047	-0.169	0.009
2007	-11.439	0.046	1.245	0.050	-0.172	0.010

- *Modelo com log-intensidade cúbica no tráfego*

$$\lambda(x) = \exp(\beta_0 + \beta_1 Z(x) + \beta_2 Z^2(x) + \beta_3 Z^3(x)), \quad (4.32)$$

com $\beta_0, \beta_1, \beta_2$ *e* β_3 os parâmetros e $Z(x)$ o tráfego na localização x.

Tabela 4.8 Parâmetros estimados do modelo (4.32) e erros padrão associados, para cada ano.

Ano	β_0	Erro padrão β_0	β_1	Erro padrão β_1	β_2	Erro padrão β_2	β_3	Erro padrão β_3
2004	-12.030	0.065	2.558	0.109	-0.741	0.046	0.064	0.005
2005	-12.012	0.064	2.535	0.108	-0.722	0.046	0.062	0.005
2006	-11.825	0.061	2.269	0.104	-0.632	0.045	0.054	0.005
2007	-11.835	0.063	2.253	0.109	-0.657	0.048	0.057	0.005

- *Modelo com intensidade linear no tráfego*

Este modelo toma a intensidade como sendo proporcional ao tráfego, ou seja,

$$\lambda(x) = \beta\, Z(x), \tag{4.33}$$

que equivale a

$$\log \lambda(x) = \log \beta + \log Z(x).$$

Os parâmetros estimados para cada ano são dados pela Tabela 4.9, tal que

$$\lambda(x) = e^{\beta} Z(x). \tag{4.34}$$

Tabela 4.9 Parâmetros estimados para o modelo (4.34) e respectivos erros padrão, por ano.

Ano	β	Erro padrão β
2004	-10.261	0.022
2005	-10.240	0.021
2006	-10.216	0.021
2007	-10.310	0.023

Adequabilidade do modelo

Tal como para os GLM, também na análise espacial se aplica a análise dos desvios na selecção de modelos.

Considera-se como modelo nulo o processo de Poisson homogéneo (CSR) e comparam-se com cada um dos modelos acima descritos, de modo a inferir sobre a sua adequabilidade.

Os valores-p resultantes do teste do Qui-quadrado para a comparação de modelos, provenientes da análise do desvio[6], foram inferiores a 2.2×10^{-16}, para todos os anos e para todos os modelos considerados.

Assim, para ambos os modelos acima analisados, se verifica a rejeição da hipótese de aleatoriedade completa a favor de cada uma das alternativas, pois os valores-p são bastante reduzidos.

Tendo em conta que na análise de desvio ANOVA a hipótese nula tem de ser um sub-modelo da hipótese alternativa, então esta não pode ser aplicada para o modelo (4.34) em que a intensidade é proporcional ao tráfego.

Assim, será usado outro critério, bastante útil na selecção dos modelos ajustados, e já discutido anteriormente: o Critério da Informação de Akaike (AIC). Calculando o AIC de cada um dos modelos ajustados, pode concluir-se qual deles se adequa melhor aos dados. Quanto menor esse valor, mais adequado é o modelo.

Na Tabela 4.10 representam-se os valores da Informação de Akaike (AIC) para os três modelos propostos.

Tabela 4.10 Valores da Informação de Akaike para cada um dos modelos considerados, por ano.

Ano	Modelo (4.30)	Modelo (4.31)	Modelo (4.32)	Modelo (4.34)
2004	45218.31	44743.87	**44599.32**	44702.78
2005	46265.99	45749.04	**45613.01**	45712.25
2006	47262.3	46846.36	**46735.75**	46847.64
2007	43990.92	43630.73	**43524.09**	43678.79

Conclui-se que o menor AIC favorece o modelo (4.32) para todos os anos, ou seja, a estimativa da log-intensidade que melhor se adapta os dados é cúbica no tráfego.

[6] A metodologia da análise dos desvios (ANOVA) é descrita na Secção 3.5.1.

Após a escolha do melhor modelo, é possível calcular uma estimativa da função K não homogénea com as estimativas da função intensidade os valores ajustados do melhor modelo proposto para cada ano. A Figura 4.23 representa os gráficos das estimativas dessa função (a cheio) contra o valor teórico da mesma (a tracejado).

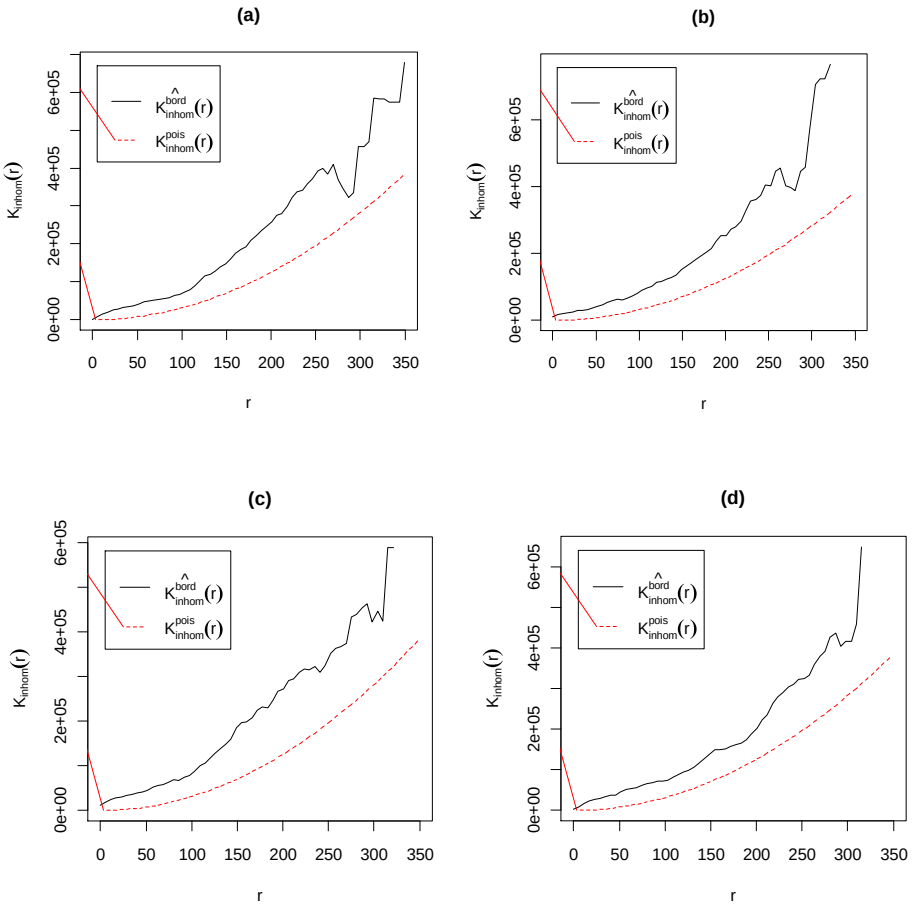


Figura 4.23 Função $\hat{K}$ não homogénea (linha a cheio) com intensidade estimada pelos valores ajustados do modelo (4.32), contra a correspondente função teórica (a tracejado), para os anos de (a) 2004 , (b) 2005, (c) 2006 e (d) 2007.

Os gráficos da figura acima sugerem que, mesmo considerando a dependência da localização dos acidentes com o tráfego, a ocorrência dos mesmos ainda parece estar aglomerada, dado o afastamente

das duas funções. Isto pode dever-se, por exemplo, à existência de outras variáveis que influenciem o padrão espacial e que não estão a ser consideradas nos modelos propostos ou de outros modelos mais adequados para descrever a dependência da intensidade dos acidentes com o tráfego.

Qualidade do ajustamento de modelos

Na secção anterior concluiu-se que dos modelos não-homogéneos propostos, o que melhor se ajusta aos dados nos anos de 2004 e 2005 é o modelo cujo logaritmo da intensidade é função cúbica do tráfego. É sob esse modelo que será analisada a qualidade do ajustamento.

Como foi referido anteriormente, há um grande condicionamento para a aplicação dos métodos de inferência clássicos, dado que as funções de verosimilhança são de difícil tratamento matemático.

Os métodos considerados includem testes de hipótese, tais como o teste de Qui-quadrado para a bondade do ajustamento, o teste de Kolmogorov Smirnov (não paramétrico) e testes de Monte Carlo baseados nos invólucros já mencionados. Será, ainda, realizada uma análise dos resíduos, apesar de algumas dificuldades de interpretação, como será mencionado.

- *Bondade do ajustamento via métodos não paramétricos*

Um teste para avaliar a qualidade do ajustamente é o teste qui-quadrado baseado na contagem de quadraturas, à semelhança do que foi feito para testar a hipótese de *CSR* . Sob hipótese nula, o número de ocorrências dos acidentes nas divisões efectuadas são variáveis independentes Poisson com média calculada pelo modelo ajustado. Mais uma vez, a janela de observação foi dividida em 23 regiões de igual área.

O teste resultou num valor-p inferior a 2.2×10^{-16} para todos os anos, rejeitando o modelo proposto para cada um.

Um outro teste disponível é uma versão adaptada do *teste de Kolmogorov-Smirnov* para processos pontuais, que compara a distribuição dos valores observados do tráfego nas localizações dos acidentes com a distribuição ajustada dos mesmos sob o modelo em causa. Os valores-p resultantes, bem como a estatística de teste D, são apresentados na Tabela 4.11 para todos os anos, inferiores a 2.2×10^{-16}, rejeitando, assim, a hipótese nula dos dados serem ajustados pelo modelo (4.32).

Tabela 4.11 Valores da estatística de teste D do teste de Kolmogorov-Smirnov e respectivos valores-p, por ano.

Anos	***D***	***Valor-p***
2004	0.0654	5.215×10^{-8}
2005	0.0714	9.531×10^{-10}
2006	0.0632	7.26×10^{-8}
2007	0.081	1.612×10^{-11}

Os valores-p ínfimos resultantes levam à rejeição da hipótese nula de os dados serem bem ajustados pelo modelo log-cúbico (4.32).

- *Bondade do ajustamento usando funções distribuição do espaço vazio e vizinho mais próximo*

Tal como para testar a hipótese de aleatoriedade completa, uma forma de verificar se o modelo proposto se ajusta aos dados é através de métodos usanto testes de Monte Carlo.

Neste caso, ao invés de se efectuarem n simulações de acontecimentos de um processo de Poisson homogéneo (*CSR)*, serão realizadas n simulações de realizações do processo pontual do modelo ajustado. A estimação da função teórica é tida como a média de outras n realizações simuladas do modelo em questão.

Usando a função do espaço vazio e considerando um teste de tamanho $\alpha = 1\%$, obtiveram-se os invólucros e respectiva estimativa da função F, baseados nesse teste.

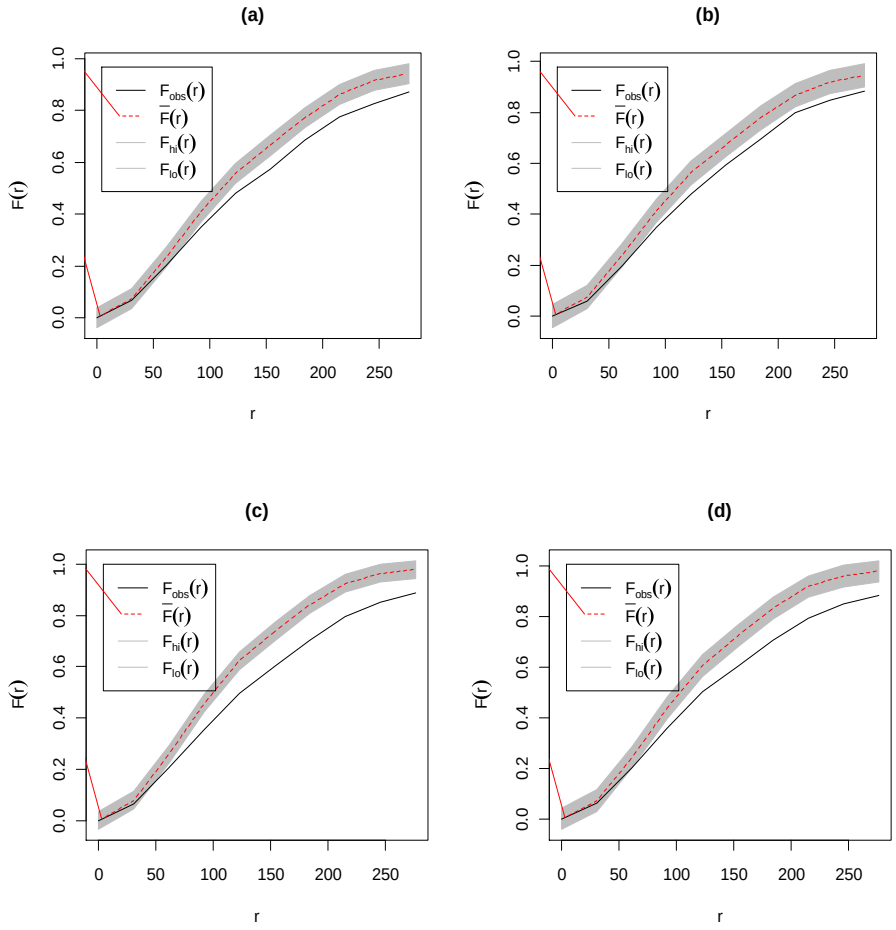


Figura 4.24 Estimativas das funções *F* (a cheio) e invólucros com base no modelo proposto para cada ano, contra a correspondente estimativa da distribuição teórica (a tracejado), para os anos de (a) 2004, (b) 2005, (c) 2006 e (d) 2007.

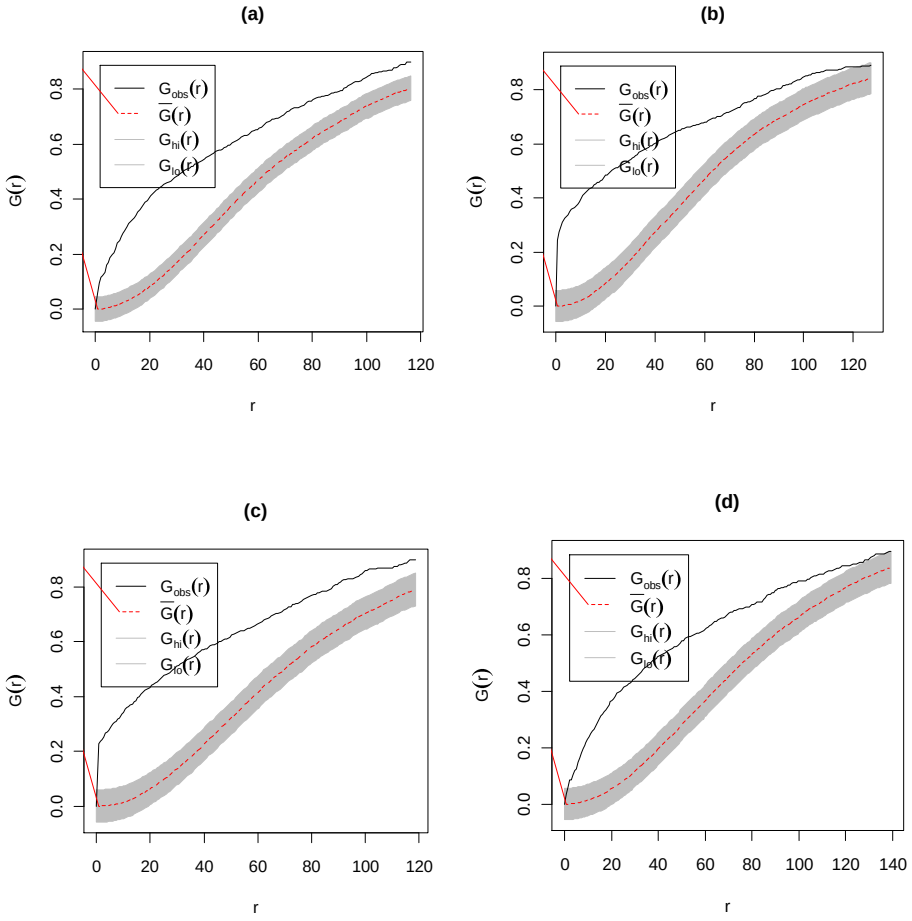


Figura 4.25 Estimativas das funções G (a cheio) e invólucros com base no modelo proposto para cada ano, contra a correspondente estimativa da distribuição teórica (a tracejado), para os anos de (a) 2004, (b) 2005, (c) 2006 e (d) 2007.

Da análise dos gráficos dos invólucros conclui-se que, apesar de uma ligeira melhoria relativamente à aplicação deste método assumindo homogeneidade, as estimativas das funções F e G ainda caiem fora dos invólucros, sugerindo um ajustamento mal conseguido.

- *Análise dos resíduos*

A análise dos resíduos é um ferramenta de diagnóstico da adequabilidade de um modelo bastante útil em várias áreas da Estatística, sendo que na Estatística espacial está a ser, apenas, actualmente desenvolvida.

Deste modo, será, aqui, feita apenas uma breve análise dos resíduos para os dados dos acidentes.

Para cada região B, os resíduos são definidos como

$$R(B) = n(\boldsymbol{x} \cap B) - \int_B \hat{\lambda}(u)du, \tag{4.35}$$

com $\boldsymbol{x}$ o padrão pontual observado e $n(\boldsymbol{x} \cap B)$ o número de pontos de $\boldsymbol{x}$ que estão na região B.

Uma forma de fazer uma análise neste âmbito é considerar um alisamento dos resíduos (Figura 4.26), dado por

$$s(x) = \hat{\lambda}(x) - \lambda^{\natural}(x), \tag{4.36}$$

onde $\hat{\lambda}(x)$ é a estimativa (não paramétrica) kernel da itensidade na localização x e $\lambda^{\natural}(x)$ a estimativa alisada da itensidade de acordo com o modelo ajustado. O modelo está bem ajustado se a diferença (4.36) for aproximadamente zero.

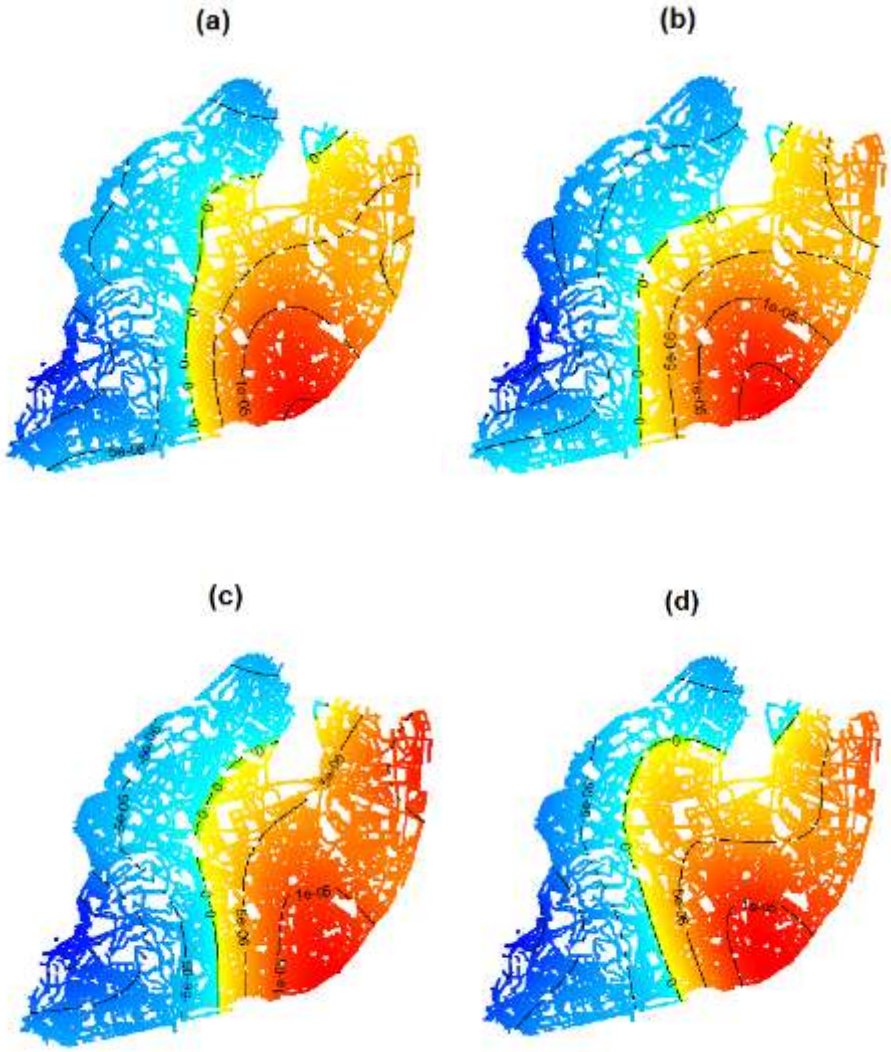


Figura 4.26 Alisamento residual do modelo ajustado para os anos de (a) 2004, (b) 2005, (c) 2006 e (d) 2007.

A Figura 4.26 sugere que há uma subestimação da intensidade da ocorrência dos acidentes na zona Este da cidade de Lisboa, e uma sobrestimação na zona Oeste, em qualquer dos anos.

4.5.2 Análise espacial do total dos acidentes

Nesta secção será feita uma análise muito semelhante à que foi realizada na secção anterior considerando o conjunto dos acidentes dos quatro anos. Desta forma, pretende minimizar-se os efeitos aleatórios que caracteristicas específicas anuais possam fazer emergir, tornando a amostra menos representativa da realidade actual.

A análise exploratória efectuada será do mesmo âmbito da considerada na análise por anos. Interessa visualizar a distribuição da ocorrência dos acidentes, limitada pela janela de observação, tal como mostra a Figura 4.27.

Figura 4.27 Localização da ocorrência dos acidentes, limitados à janela de observação considerada.

Verifica-se que existe uma concentração dos acidentes na parte Centro Este de Lisboa, e um número consideravelmente menor na periferia.

Este facto é constatado, mais uma vez, pelas estimativas Kernel para a intensidade dos padrões pontuais dos acidentes (Figura 4.28).

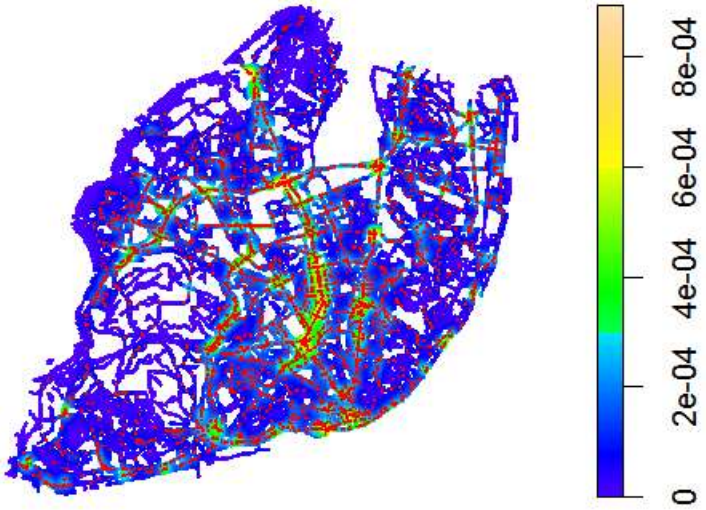


Figura 4.28 Estimativas kernel da intensidade dos padrões espaciais observados para os quatro anos considerados.

A figura revela uma maior intensidade na região centro (em tons de amarelo) e uma menor nas periferias (em tons de azul), sugerindo alguma falta de homogeneidade na distribuição dos acidentes, como se verificou anteriormente para cada um dos anos.

No caso de se considerar homogeneidade, a intensidade (média) estimada é de 0.000122 acidentes or metro quadrado.

4.5.2.1 Testes à hipótese de aleatoriedade completa

Tal como na análise por anos, serão aplicados testes que permitam testar a hipótese de aleatoriedade espacial completa.

Numa primeira análise analisou-se o gráfico da função K estimada, comparando-a com a correspondente teórica no caso de homogeneidade ($K(t) = \pi t^2$).

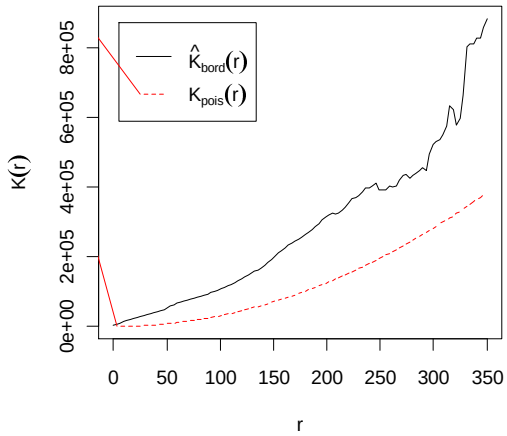


Figura 4.29 $\widehat{K}(t)$, (linha sólida), com correcção de fronteira pelo método da fronteira, e correspondente função teórica sob hipótese de CSR.

A Figura 4.29 sugere alguma falta de homogeneidade dos dados, a favor duma aglomeraçãodos mesmos, visto que a função K estimada se afasta da correspondente teórica por cima.

Da mesma forma, foram construidos os gráficos das funções distribuição do espaço vazio (Figura 4.30) e do vizinho mais próximo (Figura 4.31), contra as correspondentes funções teóricas sob hipótese de aleatoriedade espacial completa. Mais uma vez se considerou o método da fronteira para correcção dos efeitos de fronteira.

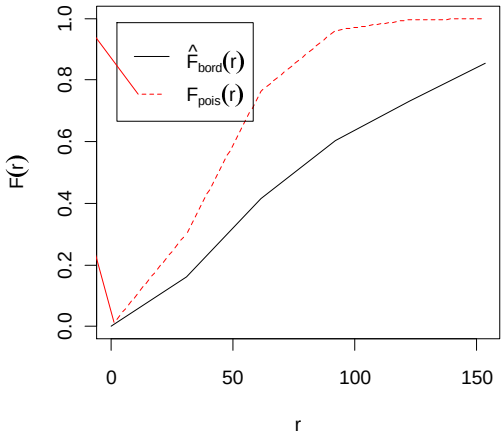


Figura 4.30 Estimativa da função distribuição de F com correcção de fronteira (a cheio), com a correspondente função teórica sob hipótese de CSR, (a tracejado).

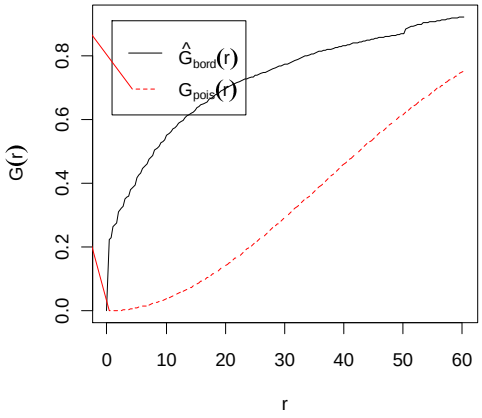


Figura 4.31 Estimativa da função distribuição de G com correcção de fronteira (a cheio), com a correspondente função teórica sob hipótese de CSR, (a tracejado).

Note-se que, em ambos os casos, os gráficos sugerem a rejeição da hipótese de aleatoriedade completa, visto que as estimativas das funções distribuição se afastam significativamente das correspondentes funções teóricas sob essa hipótese.

Um dos testes à CSR é o da contagem de quadraturas, em que a janela de observação é dividida em regiões e aplica-se o teste de Qui-quadrado de Pearson de forma a averiguar a independência entre elas. Mais uma vez, a janela foi dividida em 23 regiões de igual área.

O teste resultou num valor-p ínfimo inferior a 2.2×10^{-16} , remetendo para a rejeição da hipótese nula, ou seja, a hipótese de aleatoriedade espacial completa.

- *Teste de Monte Carlo*

O método de Monte Carlo aplicado ao teste da hipótese de aleatoriedade completa foi, mais uma vez, realizado, quer para a função distribuição do espaço vazio, quer para a do vizinho mais próximo.

Para a função distribuição do espaço vazio foram realizadas 99 simulações da hipótese de *CSR*, resultando um teste de tamanho $\alpha = 1\%$. A Figura 4.32 representa os invólucros resultantes e a função F estimada (linha a cheio), bem como a correspondente distribuição teórica sob hipótese de aleatoriedade espacial completa (a tracejado).

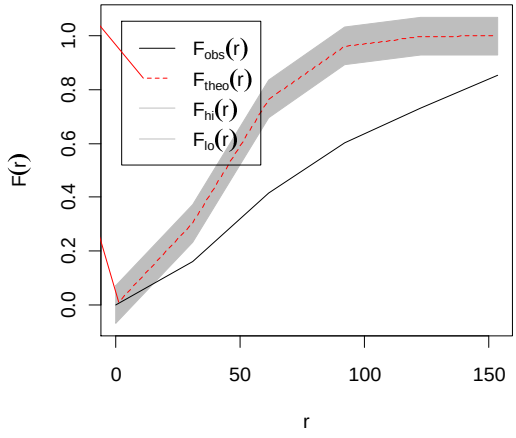


Figura 4.32 Estimativa da função F (a cheio) e invólucros contra a correspondente distribuição teórica sob hipótese de CSR (a tracejado).

Foi seguido o mesmo procedimento para a função distribuição do vizinho mais próximo, mas considerando um teste de tamanho $\alpha = 5\%$.

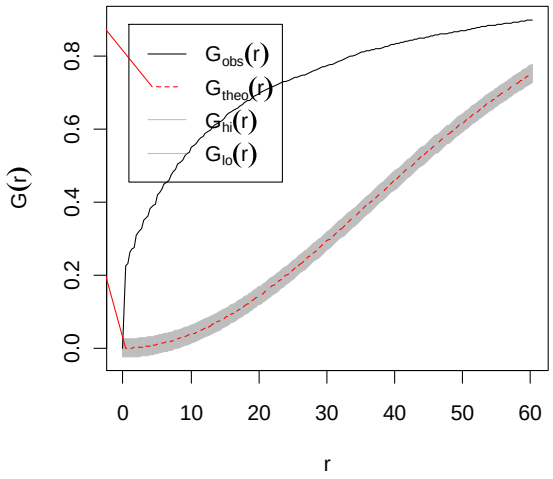


Figura 4.33 Estimativa da função G (a cheio) e invólucros contra a correspondente distribuição teórica sob hipótese de CSR (a tracejado).

Os gráficos das figuras acima sugerem a não homogeneidade dos dados, dado o afastamento de cada uma das funções dos correspondentes valores teóricos sob hipótese de aleatoriedade espacial completa.

4.5.2.2 Análise da falta de homogeneidade

Considerando, novamente, a variável tráfego para a análise da nhomogeneidade, foram ajustados modelos que a consideram, sob diversas formas. O tráfego considerado para as análises foi o uniformizado (Equação (4.28)) e, tal como anteriormente, será simplesmente invocado como *tráfego.*

Um dos métodos que permite analisar a dependência do tráfico na densidade dos acidentes é o método da contagem de quadraturas. Mais uma vez, a janela de observação foi dividida em regiões irregulares de acordo com o tráfego, para consequente aplicação do teste do qui-quadrado a cada uma delas.

A divisão do tráfego foi efectuada em cinco partes obtendo-se os seguintes valores:

Tabela 4.12 Divisão do tráfego em cinco partes e correspondentes número de pontos em cada divisão da janela.

Tráfego	]0.197, 1.5]	]1.5, 2.8]	]2.8, 4.1]	]4.1, 5.4]	]5.4, 6.7]
Nº acidentes	4776	2055	915	270	222

O teste resultou num valor observado $\chi^2 = 4224.257$ e um valor-p inferior a 2.2×10^{-16}, sugerindo a rejeição da hipótese de homogeneidade.

Um outro teste, mais potente, é o de Kolmogorov-Smirnov, onde se usa o tráfego como variável contínua, não havendo a necessidade de o transformar em factores. Neste teste o valor $T(x, y)$ do tráfego é calculado para cada localização dos acidentes e, posteriormente, compara-se a distribuição empírica dos valores de T com a distribuição dos valores ajustados de T sob a hipótese de *CSR*.

Este teste resultou num valor da estatística de teste $D = 0.3747$ e um valor-p inferior a 2.2×10^{-16}, indicando a rejeição da hipótese de *CSR*.

4.5.2.3 Modelação e Inferência

No conjunto de todos os dados, foram novamente propostos quatro modelos para a estimativa da intensidade:

1. Função log-linear do tráfego;

2. Função log-quadrática no tráfego;
3. Função log-cúbica no tráfego e
4. Função linear ao tráfego (proporcionalidade).

- *Modelo com log- intensidade linear no tráfego*

Para este modelo obteve-se a seguinte estimativa da intensidade:

$$\hat{\lambda}(x) = \exp\big(-9.501 + 0.420 * Z(x)\big). \quad (4.37)$$

- *Modelo com log-intensidade quadrática no tráfego*

Por outro lado, se a intensidade se relacionar quadraticamente com o tráfego, então virá que

$$\hat{\lambda}(x) = \exp(\ -10.102 + \ 1.349\ Z(x) - 0.181\ Z^2(x)), \quad (4.38)$$

- *Modelo com log-intensidade cúbica no tráfego*

$$\hat{\lambda}(x) = \exp(\ -10.542 + \ 2.439\ Z(x) - 0.697\ Z^2(x) + 0.060 Z^3(x)), \quad (4.39)$$

- *Modelo com intensidade proporcional ao tráfego*

Para este modelo a intensidade estimada vem

$$\hat{\lambda}(x) = e^{-8.836} Z(x). \quad (4.40)$$

Adequabilidade do modelo

Através da análise do desvio, foi realizado o teste da razão de verosimilhanças da hipótese nula do modelo se ajustar a um processo de Poisson homogéneo contra a hipótese de se ajustar aos restantes modelos considerados.

O teste resultou num valor-p extremamente pequeno, inferior a 2.2×10^{-6}, para qualquer um dos modelos, excepto o modelo em que a intensidade é proporcional ao tráfego, indicando a rejeição da hipótese de aleatoriedade espacial completa a favor de cada um dos modelos alternativos propostos.

Tendo em conta que na análise do desvio ANOVA a hipótese nula tem de ser um sub-modelo da hipótese alternativa, então esta não pode ser aplicada para o modelo em que a intensidade é proporcional ao tráfego.

Assim, será usado outro critério, bastante útil na selecção dos modelos ajustados, e já discutido anteriormente: o Critério da Informação de Akaike (AIC). Para o modelo (4.34) este valor é de 159701.9 e para o modelo nulo de 164925.7. Como o primeiro é inferior ao segundo, então prevalece o modelo linear (4.34).

A tabela abaixo apresenta os valores do AIC para os modelos acima descritos, de forma a escolher aquele que melhor se ajusta aos dados.

Tabela 4.13 Valores do AIC para cada um dos modelos propostos.

Modelo Log-linear	Modelo Log-quadrático	Modelo Log-cúbico	Modelo proporcional
161643.4	159861.4	**159331.1**	159701.9

Os valores do AIC sugerem que qualquer um dos modelos de poisson não homogéneos são preferíveis ao modelo nulo, favorecendo, aínda, o modelo com a estimativa do logaritmo da intensidade como função cúbica do tráfego (4.32).

Conclui-se, então, que em ambos os casos se rejeita a hipótese de aleatoriedade completa, a favor de modelos de Poisson não homogéneos com intensidade dependente do tráfego.

De acordo com o modelo resultante (4.32) pode, então, analisar-se a função K não-homogénea, considerando-se a estimativa da intensidade resultante do mesmo. A Figura 4.34 representa essa estimativa (linha a cheio), com a correcção de fronteira usual, contra a correspondente distribuição teórica da mesma.

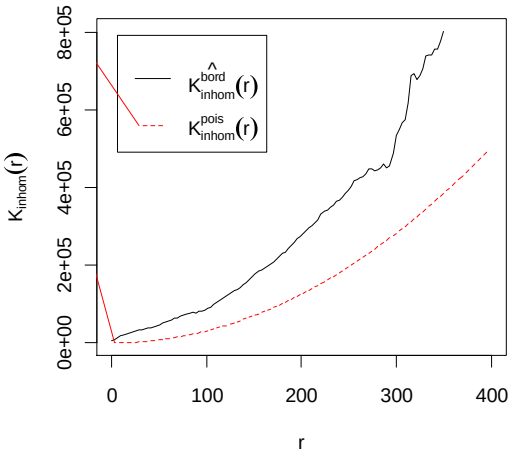


Figura 4.34 Função $\hat{K}$ não homogénea (linha a cheio) com intensidade estimada pelos valores ajustados do modelo (4.32), contra a correspondente função teórica (a tracejado).

O gráfico acima sugere que mesmo após considerar-se a dependência no tráfego, os acidentes parecem estar aglomerados.

4.5.2.4 Ajustamento de modelos

Tal como feito na análise considerando cada ano separadamente, a qualidade do ajustamento do modelo resultante será analisada através de métodos não pramétricos (testes de contagem de quadraturas e teste de Kolmogorov-Smirnov), assim como através de testes usando o método de Monte Carlo.

- *Bondade do ajustamento via métodos não paramétricos*

Na aplicação do teste da contagem de quadrados para o modelo (4.32) resultou um valor-p mínimo inferior a 2.2×10^{-16}, indicando a rejeição do modelo proposto, tal como tinha sido verificado para a análise anual.

O mesmo resultou com o teste de Kolmogorov-Smirnov, com um valor da estatística de teste $D = 0.0675$ e um valor-p inferior a 2.2×10^{-16}, indicando, mais uma vez, a rejeição do modelo.

- *Bondade do ajustamente via método de Monte Carlo*

Os testes usando invólucros já conhecidos, foram novamente aplicados a este modelo, através das distribuições das funções do espaço vazio e do vizinho mais próximo.

O gráficos das estimativas dessas funções e dos invólucros baseados em n simulações de realizações do modelo proposto, contra as correspondentes funções teóricas, estão representados nas Figura 4.35 e Figura 4.36.

Note-se que, mais uma vez, a estimativa da distribuição teórica é calculada como a média de outras n simulações de realizações do modelo em questão. Para a função distribuição do espaço vazio o tamanho do teste foi de $\alpha = 1\%$ e para a função distribuição do vizinho mais próximo de $\alpha = 5\%$.

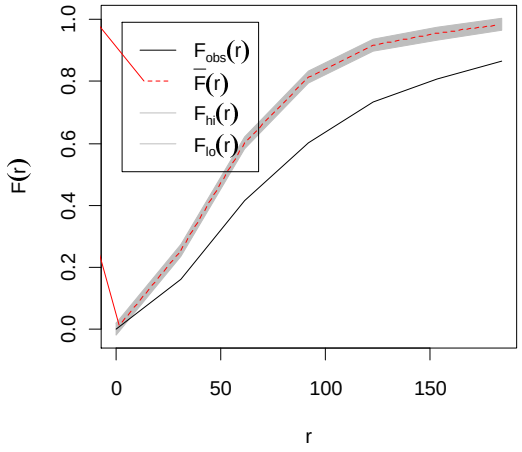


Figura 4.35 Estimativas da função F (a cheio) e invólucros com base no modelo (5.30), contra a correspondente estimativa da distribuição teórica (a tracejado), com base em 99 simulações.

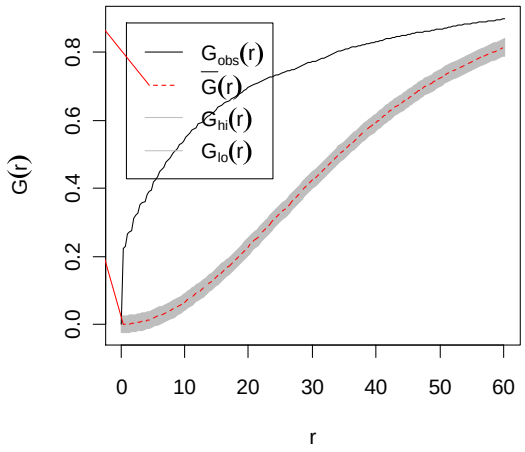


Figura 4.36 Estimativas da função G (a cheio) e invólucros com base no modelo (5.30), contra a correspondente estimativa da distribuição teórica (a tracejado), com base em 19 simulações.

Da análise dos gráficos acima conclui-se que, apesar de uma ligeira melhoria relativamente à aplicação deste método assumindo homogeneidade, as estimativas das funções F e G ainda caiem fora dos invólucros, sugerindo um mau ajustamento do modelo aos dados.

4.6 Conclusões

A análise espacial dos acidentes rodoviários com vítimas em Lisboa envolvendo processos pontuais permitiu concluir que estes não se distribuem uniformemente pela cidade, rejeitando-se, deste modo, a hipótese de aleatoriedade espacial completa. Assim, desenvolveu-se um modelo não-homogéneo de Poisson que inclui a variável do tráfego médio, com o objectivo de explicar a heterogeneidade encontrada. Após a consideração de várias hipóteses para a dependência da intensidade dos acidentes no tráfego, concluiu-se que a que melhor se ajustava aos dados é aquela em que o logaritmo da intensidade é cúbico na variável espacial considerada.

No entanto, após uma análise através de testes de ajustamento, do método de Monte Carlo usando as funções sumárias usuais e análise dos resíduos, o modelo ajustado não se mostrou adequado. Tal pode dever-se à não inclusão de outras variáveis importantes na explicação da distribuição dos acidentes, que se torna, assim, um desafio para investigações futuras neste contexto.

5 Conclusões e desenvolvimentos futuros

5.1 Conclusões

Os acidentes rodoviários têm sido alvo de vários estudos, na tentativa de reconhecer os principais factores que contribuem para os números elevados de sinistralidade rodoviária em Portugal. Actualmente, a Estratégia Nacional de Segurança Rodoviária (ENSR) prevê a implementação de várias Acções Chave (ANSR&ISCTE, 2009, p.61-77) para fazer face a este problema, enquadradas em objectivos operacionais específicos delimitados pela estratégia.

Esta dissertação teve como finalidade estudar os acidentes com vítimas na cidade de Lisboa, procurando as principais causas que estão na origem do seu elevado número e gravidade e caracterizando a distribuição espacial do risco da ocorrência dos mesmos.

Para esse estudo, foi considerada uma base de dados de 9263 acidentes com vítimas em Lisboa entre os anos de 2004 a 2007, que inclui um conjunto de características associadas aos mesmos. Desses acidentes, 8238 foram georreferenciados.

Numa primeira fase foi efectuada uma análise exploratória das várias características dos acidentes, a fim de percepcionar quais as variáveis que parecem interferir mais na ocorrência e gravidade dos mesmos. Essas variáveis foram usadas no processo de desenvolvimento de um modelo logístico Binomial, a fim de averiguar quais os factores que mais contribuem para a gravidade dos acidentes na cidade de Lisboa. Mostraram-se importantes oito variáveis, incluídas no modelo: factores atmosféricos, luminosidade, hora do dia da ocorrência do acidente, funcionamento dos sinais luminosos, sexo dos condutores, existência ou não de peões e natureza do acidente interagindo com a idade dos condutores. Constatou-se que é mais provável a ocorrência de um acidente grave se se tratar de um despiste, se for de noite e entre as 00h00 e as 6h59. O estudo sugere ainda que a existência de peões envolvidos em acidentes conduz a uma maior gravidade dos mesmos, confirmando percepções anteriores que justificaram a inclusão da Acção Chave 11.6.2 na ENSR: "Realização de um estudo pormenorizado de acidentes envolvendo peões e ciclistas em meio urbano" (ANSR&ISCTE, p.68). Da mesma forma se concluiu que são relevantes várias características dos condutores envolvidos, justificando medidas no âmbito da mesma estratégia relativamente a um conhecimento mais profundo das características dos condutores, conforme previsto na Acção Chave 5.2 (ANSR&ISCTE, 2009, p.64). No entanto, o estudo sugere que condições atmosféricas menos favoráveis e um mau funcionamento dos sinais luminosos não aumentam a probabilidade da existência de acidentes graves, e medidas direccionadas nesse sentido poderão não surtir efeito na redução da sinistralidade rodoviária.

Posteriormente, foi feita uma análise dos factores que mais contribuem para aumentar o número de acidentes com vítimas, considerando-se um estudo por freguesias. Neste estudo, o modelo de Poisson foi o que pareceu mais adequado, com a inclusão de um factor *offset*, o tráfego médio por freguesia, por forma a tornar comparável o número de acidentes em cada uma delas. No processo de desenvolvimento do modelo, contou-se com oito variáveis de exposição. Todas estas variáveis se mostraram importantes na explicação do número de acidentes quando se consideraram os dados agregados nos quatro anos em estudo. A par desse estudo foi realizada uma análise para cada ano separadamente, destacando-se como factores a incluir no modelo a proporção da população da freguesia que usa automóvel, a proporção da população da freguesia que trabalha na mesma, o número de hospitais e o número de escolas. Estas foram as variáveis que em todos os anos faziam parte do modelo. Concluiu-se, ainda, que quanto maior a proporção da população que utiliza automóvel, menor o risco de ocorrer um acidente com vítima, apesar de ser uma redução pequena. No entanto, quanto maior a proporção da população que trabalha na freguesia, maior o risco de acidente, assim como acontece para a proporção de população sem ensino (analfabeta). Um aumento no número de hospitais e escolas aumenta, também, essa probabilidade, ao contrário do aumento dos centros de saúde. A probabilidade de ocorrer um acidente com vítimas na freguesia aumenta aquando do aumento na proporção de população idosa. No entanto, medidas de segurança rodoviária direccionadas para este tipo de utilizador rodoviário vulnerável, não estão previstas explicitamente na ENSR.

Numa última fase foi realizada uma análise espacial da ocorrência dos acidentes, com a finalidade de perceber a sua ocorrência no espaço e estimar uma superfície de risco da ocorrência dos mesmos, associada a factores externos – variáveis espaciais. Concluiu-se, essencialmente, que os acidentes não ocorrem uniformemente no espaço, rejeitando-se, assim, a hipótese da designada aleatoriedade espacial completa. Com base neste resultado, foi levado a cabo o desenvolvimento de um modelo não-homogéneo, onde se incluiu a variável espacial do tráfego, na tentativa de explicar a falta de homogeneidade. No entanto, o modelo resultante não se mostrou adequado, provavelmente por serem necessárias mais variáveis a incluir no mesmo, como por exemplo a distância às escolas e hospitais mais próximos, e que não foram consideradas por impossibilidade de os adquirir em tempo útil.

Em suma, aos acidentes rodoviários com vítimas na cidade de Lisboa estão associados diversos factores de natureza ambiental, demográfica e urbana e algumas variáveis de exposição que contribuiem para o aumento da ocorrência e gravidade dos mesmos. Na distribuição espacial dos acidentes, rejeitou-se a hipótese de estes estarem localizados uniformemente no espaço, a favor da falta de homogeneidade na sua localização pela cidade de Lisboa.

5.2 Desenvolvimentos futuros

No estudo da sinistralidade rodoviária é crucial a obtenção de informação detalhada das características dos acidentes. Essas variáveis permitem explicar a ocorrência dos mesmos, assim como a gravidade, permitindo definir um conjunto de medidas que actuem de forma eficaz na redução da sinistralidade. Uma das dificuldades encontradas ao longo desta análise foi a escassez de informação que está, por vezes, associada a cada acidente, devido, essencialmente, à não recolha da mesma por parte das autoridades locais. Desta forma, é preciso destacar a importância que um relatório completo e rigoroso tem no estudo dos factores que contribuiem para a ocorrência dos acidentes com vítimas, possibilitando uma análise o mais próxima possível da realidade. Recomenda-se, portanto, uma forte consciencialização por parte das autoridades locais para que os relatórios dos acidentes efectuados pelas mesmas sejam o mais preciso e detalhado possível, tendo em conta o importante contributo que fornece às equipas de investigação dos acidentes rodoviários. Esta consciencialização está, de facto, já presente na ENSR, nomeadamente através de medidas propostas como a Acção Chave 23.4.3, descrita como "Elaboração de um Manual Técnico e de Boas Práticas para o registo dos acidentes de viação (preenchimento do BEAV)" .

De futuro, seria interessante uma continuação da análise espacial aqui apresentada, no sentido de obter uma explicação plausível para a falta de homogeneidade encontrada na distribuição da localização dos acidentes rodoviários com vítimas na cidade de Lisboa. Essa análise passaria por uma introdução de mais variáveis espaciais num modelo não-homogéneo, que em conjunto permitissem percepcionar essa distribuição. Alternativamente, poder-se-ia verificar se a hipótese de agregação de acidentes é válida e se resulta não da falta de homogeneidade da intensidade mas de uma interacção entre as localizações das ocorrências dos acidentes e, a partir daí, desenvolver modelos que a incorporassem. Seria, também, relevante levar a cabo um estudo envolendo a gravidade dos acidentes, através dos processos pontuais marcados, cuja marca seria a gravidade de cada um deles. A sistemática georreferenciação da localização da ocorrência dos acidentes seria consideravelmente benéfica para este tipo de análise, como é, aliás, introduzido na ENSR como medida a tomar, através de acções como a Acção Chave 23.5: "Implementação do projecto de georeferenciação da sinistralidade rodoviária".

Estudos sistemáticos como o realizado nesta dissertação seriam uma importante ferramenta para a redução da sinistralidade rodoviária, de modo a ser adquirido um conhecimento mais profundo dos acidentes com vítimas, contribuindo para a implmentação de medidas mais eficazes direccionadas para a realidade existente, evitando a alocação de recursos em acções que não se enquadrem no contexto nacional.

Este estudo contribui, assim, de forma decisiva para o avanço do conhecimento na área da segurança rodoviária, contando com a utilização de ferramentas recentes para análises inovadoras dos dados. É importante que a aplicação de ferramentas de análise como as que aqui foram realizadas sejam mais frequentes ao nível de investigação dos acidentes rodoviários com vítimas nos meios urbanos, para que Portugal possa continuar a contribuir, como se comprometeu, para um descréscimo consistente do número e da gravidade dos acidentes na Europa.

Bibliografia

(2010). *Relatório de Sinistralidade- Biénio 2009-2010- Distrito de Lisboa.* Governo Civil de Lisboa; Instituto Geográfico do Exército.

(Julho de 2011). Obtido de Consultores em Transportes, Inovação e Sistemas, S.A.: www.tis.pt

Al-Ghamdi, A. S. (2002). Using logistic regression to estimate the influence of accident factors on accident severity. *Accident Analysis & Prevention, 34*, 729-741.

ANSR. (2009). *Guia para a elaboração de Planos Municipais de Segurança Rodoviária.* Autoridade Nacional de Segurança Rodoviária.

ANSR, & ISCTE. (Março 2009). *Estratégia Nacional de Segurança Rodoviária 2008-2015.*

Baddeley, A. &. (2000). Practical maximum pseudolikelihood for. *Australian and New Zealand Journal of Statistics, 42*, 283-322.

Baddeley, A. (2007). Spatial Point Processes and Their Applications. In A. Baddeley, I. Bárány, & R. Schneider, *Stochastic Geometry: Lectures given at the C.I.M.E. Summer School held in Martina Franca, Italy, September 13–18, 2004* (Vol. 1892/2007, pp. 1-75). Berlin / Heidelberg: Springer.

Baddeley, A. (December 2010). Analysing spatial point patterns in R. *Workshop Notes.* Australia: CSIRO and University of Western Australia.

Baddeley, A., & Turner, R. (2005). Spatstat: an R package for analysing spatial point patterns. *Journal of Statistics Software, 12*, 1-42.

Baddeley, A., & Turner, R. (2006). Modelling spatial point patterns in R. In P. G. A. Baddeley, *Case Studies in Spatial Point Process Modeling* (pp. 23-74). New York: Springer Lecture Notes in Statistics 185, Springer-Verlag.

Baddeley, A., Chang, Y.-M., Song, Y., & Turner, R. (Submitted for publication.). Diagnostics for transformation of covariates in spatial Poisson point process models.

Banos, A., & Huguenin-Richard., F. (2000). Spatial distributions of road accidents in the vicinity of point sources applications to child pedestrians accidents. (E. Elsevier, Ed.) *Geography and Medicine*, 54-64.

Câmara Municipal de Lisboa. (s.d.). Obtido em Junho de 2011, de www.cm-lisboa.pt

Carvalho, M. L., & Natário, I. C. (1 a 4 de Outubro de 2008). Análise de Dados Espaciais. *Congresso Anual.* Vila Real: Sociedade Portuguesa de Estatística.

Daley, D., & Vere-Jones, D. (2003). *An introduction to the theory of point processes: elementary theory and methods* (Vol. 1). New York ; Berlin ; Heidelberg, DE: Springer-Verlag.

Diggle, P. J. (2003). *Statistical Analysis of Spatial Point Patterns* (2nd ed.). London: Hodder Arnold.

Eluru, N., Bhat, C. R., & Hensher, D. A. (2008). A mixed generalized ordered response model for examining pedestrian and bicyclist injury severity level in traffic crashes. *Accident Analysis & Prevention, 40*, 1033-1054.

EU. (2001). *Developing Urban Management and Safety.* DUMAS.

Flahaut, B., Mouchart, M., Martin, E. S., & Thomas, I. (2003). The local spatial autocorrelation and kernel method for identifying black zones: A comparative approach. *Accident Analysis & Prevention, v.35*, 991-1004.

http://pt.wikipedia.org/wiki/Ficheiro:Lisboa_-_Bairros_e_Fregueisas.png. Obtido em Janeiro de 2011

http://www.ansr.pt/Portals/0/Guia_PMSR_2009.pdf. Obtido em Outubro de 2010

Illian, J., Penttinen, A., Stoyan, H., & Stoyan, D. (2008). *Statistical Analysis and Modelling of Spatial Point Patterns.* Chichester: John Wiley & Sons.

Kim, K., Lawrence, N., Richardson, J., & Li, L. (1996). Modelling fault among bicyclists and drivers involved in collisions in Hawaii 1986-1991. In *Transportation Reearch Record 1538, TRB* (pp. 75-80). Washington, DC: National Research Council.

Lord, D., & Mannering, F. (2010). The Statistical Analysis of Crash-Frequency Data: A Review and Assessment of Methodological Alternatives. *Transportation Research Part A: Policy and Practice, 44*, 291-305.

Manning, C. (2007). Logistic Regression (with R).

Mercier, C., Shelley, M., Rimkus, J., & Mercier, J. (1997). Age and gender as predictors of injury severity in head-on highway vehicular collisions. In *Transportation Research Record 1581, TRB.* Washington, DC: National Research Council.

Moller, J., & Waagepetersen, R. P. (2007). Modern statistics for spatial point process. *Scandinavian Journal of Statistics, 34*, 643-684.

Nelder, J. A., & McCullagh, P. (1989). *Generalized Linear Models* (2nd ed.). London: Chapman and Hall.

Nicholson, A. (1998). Analysis of spatial distributions of accidents. *Safety Science, 31*, 71-91.

Quimby, A., Hills, B., Baguley, C., & Fletcher, J. (2003). *Urban Safety Management: Guidelines for Developing Countries.* Berkshire: TRL Limited.

Sando, T., Mussa, R., Sobanjo, J., & Sapinhour, L. (2005). Advantages and disadvantages of different crash modeling techniques. *Journal of Safeti Research, 36*, 485-487.

Turkman, M. A., & Silva, G. (2000). *Modelos Lineares Generalizados - da teoria à prática.* SPE.

Wedderburn, R., & Nelder, J. (1972). Generalized linear models. *Journal of the Royal Statistical Society Series A, 135*, 370-384.

Anexos

Anexo 1: Boletim Estatístico de Acidentes de Viação (BEAV).

As variáveis estudadas referentes às características dos acidentes rodoviários com vítimas na cidade de Lisboa, foram as obtidas a partir do Boletim Estatístico de Acidentes de Viação (BEAV), preenchido pelas forças de segurança -Guarda Nacional Republicana (GNR) e Polícia de Segurança Pública (PSP)- ao tomarem conhecimento de um acidente rodoviário, e complementadas com as coordenadas geográficas obtidas no âmbito dum projecto desenvolvido no LNEC.

(Reservado ao carimbo da unidade)

DIRECÇÃO-GERAL DE VIAÇÃO
Ministério da Administração Interna

BOLETIM ESTATÍSTICO DE ACIDENTES DE VIAÇÃO

Nº Boletim

Entidade Fiscalizadora

Instrumento de notação registado no I.N.E., sob o nº 2018, válido até 31/12/2004

A – a preencher em todos os acidentes **B e seguintes – a preencher apenas em acidentes com vítimas**

A – IDENTIFICAÇÃO DO ACIDENTE

A1 DATA/HORA
Ano Mês Dia — Hora Min.

A2 LOCALIZAÇÃO
1. ☐ Fora das localidades
☐ Dentro das localidades
2. Distrito
Concelho
Freguesia
Povoação (ou a mais próxima)
Coordenadas GPS
3. Designação da via
Km
Arruamento
n.º
4. Se houver separador central indique em que sentido
1 ☐ Crescente
2 ☐ Decrescente

A3 TIPO DE ACIDENTE
1 ☐ Acidente só com danos materiais
2 ☐ Acidente com vítimas
Mortos
Feridos graves
Feridos leves

A4 NATUREZA DO ACIDENTE
1 ☐ Despiste
2 ☐ Colisão
3 ☐ Atropelamento

A5 NÚMERO DE VEÍCULOS INTERVENIENTES
Ciclomotor e motociclo
Veículo ligeiro
Veículo pesado
Outros

A6 CONDUTORES INTERVENIENTES
1. SEXO
A B C
1 ☐☐☐ Masculino
2 ☐☐☐ Feminino
2. DATA DE NASCIMENTO
Ano Mês Dia
A
Ano Mês Dia
B
Ano Mês Dia
C

B – CIRCUNSTÂNCIAS EXTERNAS

B1 CARACTERÍSTICAS TÉCNICAS DA VIA
1. ESTRADA COM SEPARADOR
1 Auto-estrada – nº de vias de trânsito no sentido
2 Outra via – nº de vias de trânsito no sentido
2. ESTRADA SEM SEPARADOR – nº de vias no sentido
3. VIA DE TRÂNSITO
1 ☐ Esquerda
2 ☐ Direita
3 ☐ Central

B2 TRAÇADO DA VIA
1. EM PLANTA
1 ☐ Recta
2 ☐ Curva
2. EM PERFIL
1 ☐ Em patamar
2 ☐ Com inclinação
3 ☐ Em lomba
3. 1 ☐ Sem berma ou impraticável
2 ☐ Berma não pavimentada
3 ☐ Berma pavimentada
4. SITUAÇÃO DO ACCIDENTE
1 ☐ Em plena via
2 ☐ Na berma
3 ☐ No passeio
4 ☐ Em via ou pista reservada
5 ☐ Em parque de estacionamento
5. INTERSECÇÃO DE VIAS
1 ☐ **Fora da intersecção**
Em intersecção de nível
2 ☐ Em cruzamento
3 ☐ Em entroncamento
4 ☐ Em rotunda
5 ☐ Em passagem de nível
Em intersecção desnivelada
6 ☐ Em via de aceleração
7 ☐ Em via de desaceleração
8 ☐ Em ramo de ligação – entrada
9 ☐ Em ramo de ligação – saída
6. ACIDENTE EM OBRAS DE ARTE
1 ☐ Túnel
2 ☐ Viaduto/Ponte
3 ☐ Passagem estreita

B3 REGIME DE CIRCULAÇÃO
1. FAIXA DE RODAGEM COM
1 ☐ Sentido único
2 ☐ Dois sentidos
3 ☐ Reversível
2. VELOCIDADE PERMITIDA NO LANÇO
Limite geral _____ Km/h
Limite local _____ Km/h

B4 PAVIMENTO
1. TIPO DE PISO
1 ☐ Terra batida
2 ☐ Betuminoso
3 ☐ Betão de cimento
4 ☐ Calçada
2. ESTADO DE CONSERVAÇÃO
1 ☐ Em bom estado
2 ☐ Em estado regular
3 ☐ Em mau estado
3. OBSTÁCULOS OU OBRAS
1 ☐ Inexistentes
2 ☐ Não sinalizados
3 ☐ Insuficientemente sinalizados
4 ☐ Correctamente sinalizados
4. CONDIÇÕES DE ADERÊNCIA
1 ☐ Seco e limpo
2 ☐ Húmido
3 ☐ Molhado
4 ☐ Com água acumulada na faixa de rodagem
5 ☐ Com gelo, geada ou neve
6 ☐ Com lama
7 ☐ Com gravilha ou areia
8 ☐ Com óleo

B5 SINALIZAÇÃO
1. MARCAS NO PAVIMENTO
1 ☐ Sem marcas rodoviárias ou pouco visíveis
2 ☐ Com marcas – separadoras de sentido de trânsito
3 ☐ Com marcas – separadoras de sentido e de vias de trânsito
2. SINALIZAÇÃO LUMINOSA
1 ☐ Inexistente
2 ☐ A funcionar normalmente
3 ☐ Intermitente
4 ☐ Desligada
3. SINAIS
1 ☐ Stop
2 ☐ Cedência de passagem
3 ☐ Proibição de ultrapassagem
4 ☐ Passagem de peões
5 ☐ Outros

B6 LUMINOSIDADE
1 ☐ Em pleno dia
2 ☐ Sol encandeante
3 ☐ Aurora ou crepúsculo
4 ☐ Noite, sem iluminação
5 ☐ Noite, com iluminação

B7 FACTORES ATMOSFÉRICOS
1 ☐ Bom tempo
2 ☐ Chuva
3 ☐ Vento forte
4 ☐ Nevoeiro
5 ☐ Neve
6 ☐ Nuvem de fumo
7 ☐ Granizo

C – NATUREZA DO ACIDENTE

DESPISTE
1 ☐ Despiste simples
Com transposição do separador central
2 ☐ Com dispositivo de retenção
3 ☐ Sem dispositivo de retenção
4 ☐ Com transposição do dispositivo de retenção lateral
5 ☐ Com capotamento
6 ☐ Com colisão com veículo imobilizado ou obstáculo
7 ☐ Com fuga
COLISÃO
8 ☐ Frontal
9 ☐ Traseira com outro veículo em movimento
10 ☐ Lateral com outro veículo em movimento
11 ☐ Com veículo ou obstáculo na faixa de rodagem
12 ☐ Choque em cadeia
13 ☐ Com fuga
14 ☐ Outras situações
ATROPELAMENTO
15 ☐ De peões
16 ☐ De animais
17 ☐ Com fuga
Inatalia posterior. A B C ☐☐☐ A preencher no caso de se verificar

D – VEÍCULOS INTERVENIENTES

D1 CATEGORIA/CLASSE
1. VEÍCULOS A, B e C
A B C
1 ☐☐☐ Velocípede
2 ☐☐☐ Velocípede com motor
3 ☐☐☐ Ciclomotor
4 ☐☐☐ Motociclo cilindrada ≤ 50 cc
5 ☐☐☐ Motociclo cilindrada > 50 cc
6 ☐☐☐ Motociclo cilindrada > 50 cc > 25 kW potência/peso > 0,16 kW/kg
7 ☐☐☐ Automóvel ligeiro
8 ☐☐☐ Automóvel pesado
9 ☐☐☐ Veículo agrícola
10 ☐☐☐ Máquina industrial
11 ☐☐☐ Veículo sobre carris
12 ☐☐☐ Veículo de tracção animal
13 ☐☐☐ Desconhecido
2. Se o veículo for ciclomotor ou motociclo, especificar no caso de ser:
A B C
1 ☐☐☐ Triciclo
2 ☐☐☐ Quadriciclo
3. Se for automóvel ligeiro ou pesado, indicar o tipo:
A B C
1 ☐☐☐ Passageiros
2 ☐☐☐ Mercadorias
3 ☐☐☐ Misto
4 ☐☐☐ Tractor
5 ☐☐☐ Veículo especial. Qual?

4. A B C
1 ☐☐☐ Sem semi-reboque/reboque
2 ☐☐☐ Com semi-reboque/reboque

D2 **TIPO DE SERVIÇO**
A B C
1 ☐☐☐ Particular
2 ☐☐☐ Público

D3 **ANO DE MATRÍCULA**
A ☐☐☐☐ B ☐☐☐☐ C ☐☐☐☐

D4 **INSPECÇÃO PERIÓDICA**
A B C
1 ☐☐☐ Não obrigatória
2 ☐☐☐ Válida
3 ☐☐☐ Sem validade

D5 **CERTIFICADO ADR**
1. Preencher apenas no caso de transporte de mercadorias perigosas
A B C
1 ☐☐☐ Válido
2 ☐☐☐ Sem validade
3 ☐☐☐ Inexistente
2. MATÉRIA/OBJECTO PERIGOSO TRANSPORTADO ☐☐☐☐

D6 **CARGA/LOTAÇÃO/PNEUS**
1. CARGA/LOTAÇÃO
A B C
1 ☐☐☐ Sem carga
2 ☐☐☐ Com excesso de carga
3 ☐☐☐ Carga bem acondicionada
4 ☐☐☐ Carga mal acondicionada
5 ☐☐☐ Com lotação excedida
2. PNEUS
A B C
1 ☐☐☐ Sem deficiência
2 ☐☐☐ Com deficiência
3. TACÓGRAFO
A B C
1 ☐☐☐ Sem tacógrafo ou desactivado
2 ☐☐☐ Com tacógrafo

D7 **SEGURO**
A B C
1 ☐☐☐ Com seguro
2 ☐☐☐ Sem seguro
3 ☐☐☐ Isento

E – CONDUTORES INTERVENIENTES

E1 **CARACTERÍSTICAS DA HABILITAÇÃO DE CONDUÇÃO**
1. LICENÇA/CARTA DE CONDUÇÃO
A B C
1 ☐☐☐ Com licença/carta adequada ao veículo
2 ☐☐☐ Com licença/carta não adequada ao veículo
3 ☐☐☐ Em situação de instrução/exame
4 ☐☐☐ Cadocada/suspensa
5 ☐☐☐ Sem licença/carta
6 ☐☐☐ Não necessária ao veículo que conduz
2. PAÍS DE EMISSÃO
A B C
1 ☐☐☐ Portugal
2 ☐☐☐ Outro(s) A ☐☐ B ☐☐ C ☐☐
3. ANO DA HABILITAÇÃO
Relativamente ao veículo que conduzia
A ☐☐☐☐ B ☐☐☐☐ C ☐☐☐☐
4. CERTIFICADO ADR
A B C
1 ☐☐☐ Válido
2 ☐☐☐ Sem validade
3 ☐☐☐ Inexistente

E2 **CONDIÇÕES PSÍCO/FÍSICAS**
1. CONTRÔLO DO NÍVEL DE ALCOOLEMIA
A B C
1 ☐☐☐ **Submetido ao teste de alcoolemia**
Não submetido por
2 ☐☐☐ Doença
3 ☐☐☐ Lesão ou morte decorrente do acidente
4 ☐☐☐ Condutor não contactado na altura do acidente
5 ☐☐☐ Fuga
6 ☐☐☐ Recusa
7 ☐☐☐ Outra

2. TAXA DE ALCOOLEMIA
A ☐☐,☐☐ B ☐☐,☐☐ C ☐☐,☐☐
3. OUTROS FACTORES
A B C
1 ☐☐☐ Normal
2 ☐☐☐ Droga por despistagem
3 ☐☐☐ Sono/sonolência
4 ☐☐☐ Distracção
5 ☐☐☐ Doença súbita
6 ☐☐☐ Fadiga
4. TEMPO DE CONDUÇÃO CONTINUADA
A B C
1 ☐☐☐ Menos de 1 hora
2 ☐☐☐ De 1 a 3 horas
3 ☐☐☐ De 3 a 5 horas
4 ☐☐☐ Mais de 5 horas
5 ☐☐☐ Ignorada

E3 **ACÇÕES E MANOBRAS ANTES DO ACIDENTE**
1. A B C
1 ☐☐☐ Início de marcha
2 ☐☐☐ Saída de parqueamento ou via particular
3 ☐☐☐ Em marcha normal
4 ☐☐☐ Ultrapassagem pela esquerda
5 ☐☐☐ Ultrapassagem pela direita
6 ☐☐☐ Mudança de direcção para a esquerda
7 ☐☐☐ Mudança de direcção para a direita
8 ☐☐☐ Marcha atrás
9 ☐☐☐ Circulação em sentido oposto ao estabelecido
10 ☐☐☐ Travagem brusca
11 ☐☐☐ Parado ou estacionado
12 ☐☐☐ Inversão do sentido de marcha
13 ☐☐☐ Trânsito em filas paralelas
14 ☐☐☐ Mudança de via de trânsito para a esquerda
15 ☐☐☐ Mudança de via de trânsito para a direita
16 ☐☐☐ Desvio brusco/saída de fila de trânsito
17 ☐☐☐ Atravessando a via
2. Esquema ☐

E4 **INFORMAÇÃO COMPLEMENTAR A ACÇÕES E MANOBRAS**
A B C
1 ☐☐☐ Desrespeito da sinalização vertical
2 ☐☐☐ Desrespeito das marcas rodoviárias
3 ☐☐☐ Desrespeito da sinalização semafórica
4 ☐☐☐ Manobra irregular
5 ☐☐☐ Velocidade excessiva para as condições existentes
6 ☐☐☐ Não sinalização da manobra
7 ☐☐☐ Desrespeito das distâncias de segurança
8 ☐☐☐ Circulação afastado da berma ou passeio
9 ☐☐☐ Rebentamento pneumático
10 ☐☐☐ Queda de carga ou objecto
11 ☐☐☐ Falha mecânica do veículo
12 ☐☐☐ Ausência de luzes quando obrigatórias
13 ☐☐☐ Obstáculo imprevisto na faixa de rodagem
14 ☐☐☐ Abertura de porta
15 ☐☐☐ Encandeamento
16 ☐ Não identificado

E5 **ACESSÓRIOS DE SEGURANÇA**
A B C
1 ☐☐☐ Capacete
2 ☐☐☐ Cinto de segurança
3 ☐☐☐ Sem uso de cinto/capacete
4 ☐☐☐ Isento

F – CONSEQUÊNCIAS DO ACIDENTE

F1 **CONDUTORES VÍTIMAS**
1. GRAU DE GRAVIDADE DAS LESÕES
A B C
1 ☐☐☐ Morto
2 ☐☐☐ Ferido grave
3 ☐☐☐ Ferido leve

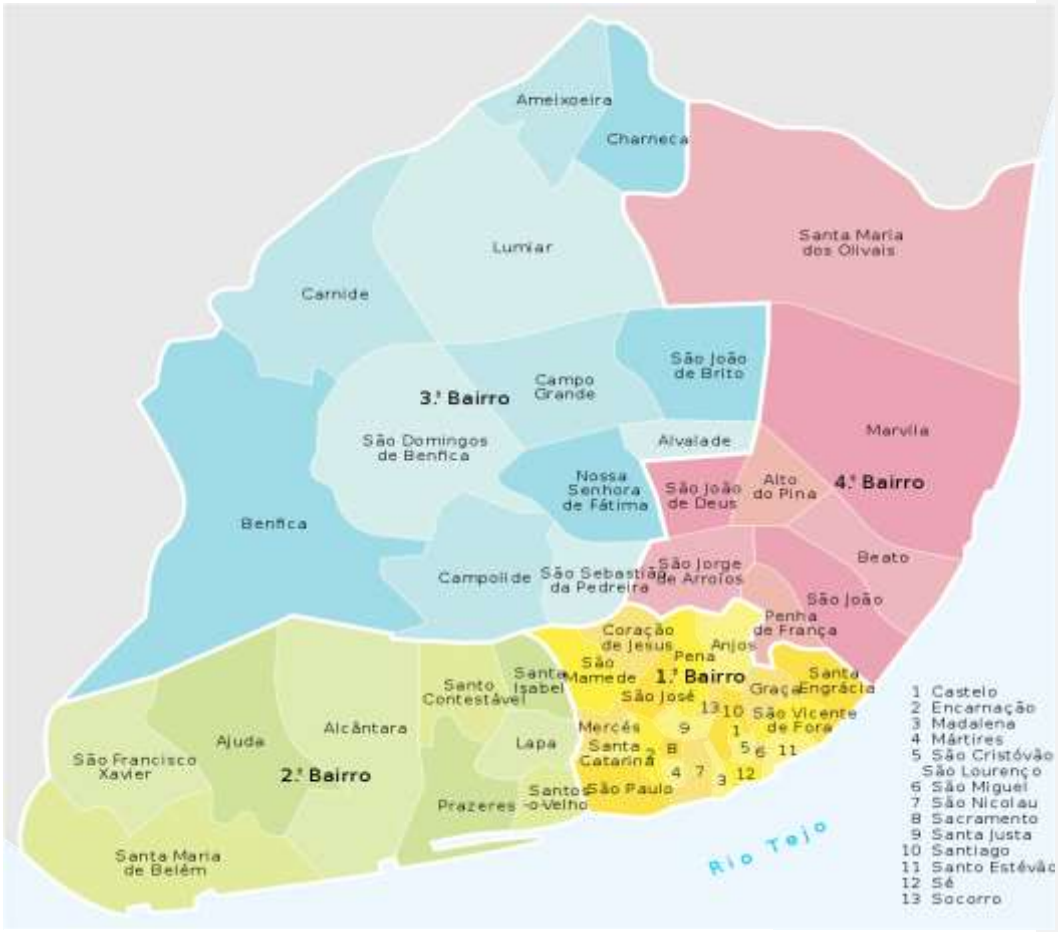
F2 **PASSAGEIROS VÍTIMAS**
Veículo A Veículo B Veículo C
1. SEXO
a b c d | i j l m | r s t u
1 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ Masculino
2 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ Feminino
2. IDADE
a ☐☐ b ☐☐ | i ☐☐ j ☐☐ | r ☐☐ s ☐☐
c ☐☐ d ☐☐ | l ☐☐ m ☐☐ | t ☐☐ u ☐☐
3. POSIÇÃO NO VEÍCULO
a b c d | i j l m | r s t u
1 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ À frente
2 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ À rectaguarda
3 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ Desconhecido
4. USO DE ACESSÓRIOS DE SEGURANÇA
a b c d | i j l m | r s t u
1 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ C/ capacete/cinto segurança
2 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ C/ sistema retenção de crianças
3 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ S/ uso capacete/cinto segurança
4 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ S/ sistema retenção de crianças
5. GRAU DE GRAVIDADE DAS LESÕES
a b c d | i j l m | r s t u
1 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ Morto
2 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ Ferido grave
3 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ Ferido leve
4 ☐☐☐☐ | ☐☐☐☐ | ☐☐☐☐ Ileso

F3 **PEÕES VÍTIMAS**
1. SEXO
a b c d
1 ☐☐☐☐ Masculino
2 ☐☐☐☐ Feminino
2. a b c d
1 ☐☐☐☐ Peão isolado
2 ☐☐☐☐ Peões em grupo
3 ☐☐☐☐ Conduzindo à mão velocípedes, carros de crianças ou de deficientes físicos
4 ☐☐☐☐ Deslocando-se sobre patins, trotinetes ou outros
3. IDADE
a ☐☐ b ☐☐ c ☐☐ d ☐☐
4. CONDIÇÕES PSÍCO-FÍSICAS
a b c d
1 ☐☐☐☐ Sem restrições
2 ☐☐☐☐ Com visão deficiente
3 ☐☐☐☐ Com audição deficiente
4 ☐☐☐☐ Com deficiência motora
Influenciado pelo álcool
a ☐☐,☐☐ b ☐☐,☐☐ c ☐☐,☐☐ d ☐☐,☐☐
5
5. ACÇÕES
a b c d
1 ☐☐☐☐ A sair ou entrar num veículo
2 ☐☐☐☐ Surgindo inesperadamente na faixa de rodagem de trás de um obstáculo
3 ☐☐☐☐ Em plena faixa de rodagem
4 ☐☐☐☐ Em trabalhos na via
5 ☐☐☐☐ Atravessando fora da passagem de peões, a menos de 50 m de uma passagem
6 ☐☐☐☐ Atravessando fora da passagem de peões a mais de 50 m de uma passagem ou quando não existe passagem
7 ☐☐☐☐ Atravessando em passagem sinalizada
8 ☐☐☐☐ Atravessando em passagem sinalizada com desrespeito da sinalização semafórica
9 ☐☐☐☐ Em ilhéu ou refúgio na via
10 ☐☐☐☐ Transitando pela direita da faixa de rodagem
11 ☐☐☐☐ Transitando pela esquerda da faixa de rodagem
12 ☐☐☐☐ Transitando pela berma ou passeio
6. UTILIZAÇÃO DE MATERIAL REFLECTOR
a b c d
1 ☐☐☐☐ Sim
2 ☐☐☐☐ Não
7. GRAVIDADE DAS LESÕES
a b c d
1 ☐☐☐☐ Morto
2 ☐☐☐☐ Ferido grave
3 ☐☐☐☐ Ferido leve

DATA ___/___/___
Número de boletins utilizados neste acidente ☐☐
Nome _____
(Posto) _____

Anexo 2: Mapa das freguesias da cidade de Lisboa

As freguesias da cidade de Lisboa referentes aos anos anteriores a 2010 estão representadas na figura abaixo. Na secção 2.1 é realizada uma análise exploratória tendo em conta o número de acidentes por freguesia, e na secção 3.6.2 a análise de regressão é feita por freguesia da cidade de Lisboa.



Fonte: *http://pt.wikipedia.org/wiki/Ficheiro:Lisboa_-_Bairros_e_Fregueisas.png*.

Figura I: Mapa das freguesias de Lisboa, antes de 2010.

Anexo 3: Área (m^2) e População residente de cada freguesia da cidade de Lisboa, em 2006.

Na Secção 2.1 o número de acidentes com vítimas ocorridos na cidade de Lisboa é analisado por área da freguesia e por 1000 habitantes, tendo em conta os valores apresentados na tabela abaixo.

Freguesia	Área (m^2)	População	Freguesia	Área (m^2)	População
Ajuda	2929958.414	17958	Santa Catarina	208908.837	4081
Alcantara	4399583.064	14443	Santa Engrácia	521741.921	5860
Alto do pina	824790.273	10253	Santa Isabel	625329.007	7270
Alvalade	591722.974	9620	Santa Justa	250192.649	700
Ameixoeira	1598658.838	9644	Santa Maria de Belém	3389858.340	9756
Anjos	489871.945	9738	Santa Maria dos Olivais	10811129.165	46410
Beato	1559039.888	14241	Santiago	64704.647	857
Benfica	7982610.382	41368	Santo Condestável	1038834.204	17553
Campo grande	2462157.845	11148	Santo Estevão	199044.102	2047
Campolide	2747871.949	15927	Santos-o-Velho	490402.762	4013
Carnide	3985960.798	18989	São Cristovão a São Lourenço	74257.601	1612
Castelo	54156.684	587	São Domingos de Benfica	4332291.395	33678
Charneca	1918791.154	10509	São Francisco Xavier	2225641.486	8101
Coração de Jesus	557234.247	4319	São João	1516120.131	17073
Encarnação	182442.582	3182	São João De Brito	2276644.521	13449
Graça	353381.082	6960	São João De Deus	925012.960	10782
Lapa	742096.545	8670	São Jorge De Arroios	1150252.411	17404
Lumiar	6088507.684	37693	São Jose	338506.265	3278
Madalena	113956.284	380	São Mamede	605176.677	6004
Mártires	98341.610	341	São Miguel	52017.281	1777
Marvila	6326370.279	38767	São Nicolau	259704.395	1175
Mercês	272565.721	5093	São Paulo	424642.098	3521
N.sra de Fátima	1926867.040	15291	São Sebastiao da Pedreira	1080507.407	5871
Pena	513658.120	6068	São Vicente de Fora	310385.145	4267
Penha de França	705600.825	13722	Sé	121174.509	1160
Prazeres	1558935.922	8492	Socorro	113318.607	2675
Sacramento	84963.444	880			

Anexo 4: Observações com repercussão elevada e respectivas características.

Obs.	GRAV	IDD_COND	SINAIS	NATUR	HORA	LUMIN	METEO	SEXOC	PEAO
384	Ligeiro	Existe pelo menos um idoso envolvido mas nenhum jovem.	A funcionar normalmente	Despiste	[16:00; 20:59]	Noite	Bom tempo	Todos os condutores são do sexo masculino	Não
1487	Ligeiro	Existe pelo menos um idoso envolvido mas nenhum jovem.	A funcionar normalmente	Despiste	[21:00; 23:59]	Noite	Bom tempo	Pelo menos um condutor do sexo feminino e um condutor do sexo masculino	Não
4355	Ligeiro	Existe pelo menos um idoso envolvido mas nenhum jovem.	Inexistentes	Despiste	[0:00; 6:59]	Noite	Bom tempo	Pelo menos um condutor do sexo feminino e um condutor do sexo masculino	Não
4741	Grave	Existe pelo menos um idoso envolvido mas nenhum jovem.	A funcionar normalmente	Despiste	[16:00; 20:59]	Noite	Bom tempo	Todos os condutores são do sexo masculino	Sim
6221	Não grave	Existe pelo menos um idoso envolvido mas nenhum jovem.	Inexistentes	Despiste	[0:00; 6:59]	Noite	Bom tempo	Todos os condutores são do sexo masculino	Não
1766	Grave	Existe pelo menos um jovem envolvido mas nenhum idoso	Falha/Intermitente	Atropelamento	[0:00; 6:59]	Noite	Outros	Todos os condutores são do sexo masculino	Sim

Anexo 5: Probabilidades de ocorrência de um acidentes grave em diferentes situações de acidente.

METEO	IDD_COND	NATUREZA	HORA	SINAIS	LUMIN	SEXOC	PEÃO	*p*
Bom tempo	Existe pelo menos um jovem envolvido mas nenhum idoso	Despiste	[7:00; 11:00]	A funcionar normalmente	Em pleno dia	Todos os condutores são do sexo feminino	Não	0.057
Bom tempo	Existe pelo menos um idoso envolvido mas nenhum jovem	Colisão	[16:00;21:00]	A funcionar normalmente	Em pleno dia	Todos os condutores são do sexo masculino	Não	0.0727
Bom tempo	Existe pelo menos um jovem e um idoso	Despiste	[21:00;24:00]	A funcionar normalmente	Em pleno dia	Todos os condutores são do sexo masculino	Não	0.101
Bom tempo	Existe pelo menos um jovem envolvido mas nenhum idoso	Atropelamento	[16:00;21:00]	A funcionar normalmente	Em pleno dia	Todos os condutores são do sexo masculino	Sim	0.204
Outros	Existe pelo menos um jovem envolvido mas nenhum idoso	Despiste	[0:00;6:00]	Inexistentes	Noite	Todos os condutores são do sexo masculino	Não	0.175