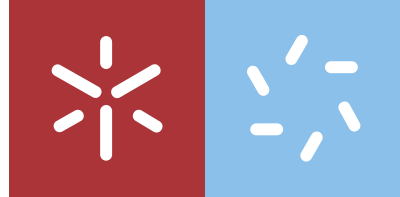




Aplicações de Modelos Lineares Generalizados
Mistos no estudo de Acidentes Rodoviários

Ana Maria Lavandeira Ferreira Maia

UMinho | 2024



Universidade do Minho
Escola de Ciências

Ana Maria Lavandeira Ferreira Maia

Aplicações de Modelos Lineares
Generalizados Mistos no estudo de
Acidentes Rodoviários

Outubro de 2024



Universidade do Minho
Escola de Ciências

Ana Maria Lavandeira Ferreira Maia

**Aplicações de Modelos Lineares Generalizados
Mistos no estudo de Acidentes Rodoviários**

Dissertação de Mestrado
Mestrado em Estatística para Ciência de Dados

Trabalho efetuado sob a orientação de
Doutora Susana Margarida Ferreira de Sá Faria
Doutora Elisabete Fraga de Freitas

Direitos de Autor e Condições de Utilização do Trabalho por Terceiros

Este é um trabalho académico que pode ser utilizado por terceiros desde que respeitadas as regras e boas práticas internacionalmente aceites, no que concerne aos direitos de autor e direitos conexos.

Assim, o presente trabalho pode ser utilizado nos termos previstos na licença abaixo indicada.

Caso o utilizador necessite de permissão para poder fazer um uso do trabalho em condições não previstas no licenciamento indicado, deverá contactar o autor, através do RepositóriUM da Universidade do Minho.

Licença concedida aos utilizadores deste trabalho:



CC BY-NC-ND

<https://creativecommons.org/licenses/by-nc-nd/4.0/>

Agradecimentos

”Quando se tem um propósito na vida, sacrifícios passam a ter sentido”.

Hoje, vivo um momento muito importante, mas foi necessário muito esforço físico e mental, além de muita perseverança, para chegar até aqui. Eu jamais conseguiria chegar até aqui sozinha.

Primeiramente, agradeço a Deus, que sempre esteve presente, a conduzir o meu caminho e não me deixar desistir.

Agradeço ao meu marido e aos meus filhos pelo incentivo, por estarem sempre ao meu lado quando precisei, pela paciência e amor durante meus momentos de ansiedade no Mestrado.

À minha amiga do Mestrado, Cármen Alves, pelo seu carinho especial e por me apoiar sempre nas dificuldades .

Agradeço de maneira especial à minha orientadora, a Prof.^a Dr.^a Susana Faria, por me auxiliar e partilhar os seus conhecimentos, cuja orientação foi essencial para conduzir esta pesquisa.

À Prof.^a Dr.^a Elizabete Fraga, pela sua orientação.

Minha profunda gratidão a todos que, de alguma forma, direta ou indiretamente, contribuíram para a realização deste projeto, para o meu crescimento pessoal e profissional.

Declaração de Integridade

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração.

Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Universidade do Minho, Braga, dezembro 2024

Ana Maria Lavandeira Ferreira Maia

Resumo

Os Modelos Lineares Generalizados Mistos (MLGM) são uma extensão dos Modelos Lineares Generalizados (MLG) que permitem incluir componentes de variabilidade não explicada pelos efeitos fixos, sendo adequados para modelar dados de contagem e analisar correlações em dados longitudinais. Estes modelos incorporam efeitos aleatórios no preditor linear, permitindo modelar a estrutura de correlação entre observações do mesmo indivíduo, o que é especialmente relevante em dados longitudinais.

Neste trabalho, demonstrou-se a flexibilidade destes modelos na análise de dados de acidentes rodoviários sendo, aplicados para compreender a influência de diversas variáveis na frequência de ocorrências de acidentes.

Os dados são referentes a sublanços das autoestradas A4 e A41 na Região do Grande Porto, Portugal, abrangendo um período de oito anos (2014-2021) e disponibilizados pela concessionária Ascendi.

A análise exploratória destacou a A41 como a autoestrada com maior volume de tráfego e maior do número de acidentes, nomeadamente, no sentido crescente. Em 2020, observou-se uma redução acentuada do número de acidentes devido à pandemia e ao confinamento. A irregularidade do pavimento aumentou entre 2014 e 2018, enquanto que a aderência à superfície melhorou.

O sublanço "Maia-Alfena" foi identificado como o sublanço de maior frequência de acidentes, em particular no sentido crescente.

A construção de modelos de regressão de *Poisson* com dois efeitos aleatórios permitiu capturar a heterogeneidade dos dados, revelando que o estado do pavimento, o número de vias e as características geométricas, tanto em planta como em perfil da estrada, influenciam a ocorrência de acidentes.

Palavras-chave: Acidentes Rodoviários, Modelação Estatística, Modelos Lineares Generalizados Mistos, Regressão de *Poisson*, Segurança Rodoviária.

Abstract

Generalized Linear Mixed Models (GLMMs) are an extension of Generalized Linear Models (GLMs) that allow the inclusion of components of variability not explained by fixed effects, and are suitable for modeling count data and analyzing correlations in longitudinal data. These models incorporate random effects in the linear predictor, allowing the modeling of the correlation structure between observations of the same individual, which is especially relevant in longitudinal data.

In this work, the flexibility of these models in the analysis of road accident data was demonstrated, being applied to understand the influence of several variables on the frequency of accident occurrences.

The data refer to subsections of the A4 and A41 motorways in the Greater Porto Region, Portugal, covering an eight-year period (2014-2021) and provided by the concessionaire Ascendi.

The exploratory analysis highlighted the A41 as the motorway with the highest volume of traffic and the highest number of accidents, particularly in the upward direction. In 2020, there was a sharp reduction in the number of accidents due to the pandemic and lockdown. Pavement unevenness increased between 2014 and 2018, while surface adhesion improved.

The "Maia-Alfena" subsection was identified as the subsection with the highest frequency of accidents, particularly in the ascending direction.

The *Poisson* Regression models with two random effects allowed capturing the heterogeneity of the data, revealing that the condition of the pavement, the number of lanes and the geometric characteristics, both in plan and in the profile of the road, influence the occurrence of accidents.

Keywords: Road Accidents, Statistical Modeling, Generalized Linear Mixed Models, *Poisson* Regression, Road Safety.

Conteúdo

Abreviaturas	1
1 Introdução	2
1.1 Objetivos	4
1.2 Estrutura da Tese	5
2 Revisão da Literatura	6
2.1 Modelos Estatísticos	6
2.2 Modelos Estatísticos para Acidentes Rodoviários	7
3 Metodologia	10
3.1 Dados Longitudinais	10
3.2 Modelo Linear Misto	11
3.3 Modelo Linear Generalizado Misto	13
3.3.1 Estrutura do Modelo Linear Generalizado Misto	13
3.3.2 Interpretação dos Parâmetros do Modelo	15
3.3.3 Estimação dos Parâmetros do Modelo	16
3.3.4 Inferência-Testes de Hipóteses e Seleção do Modelo	18
3.3.5 Análise dos Resíduos	19
4 Aplicações nos Acidentes Rodoviários	21
4.1 Base de Dados	21
4.2 Análise Exploratória dos Dados	28
4.2.1 Variável Resposta	28
4.2.2 Análise dos Acidentes nas Autoestradas A4 e A41: Uma Abordagem Comparativa	38
4.2.3 Análise Exploratória do Número de Acidentes no Sublanço “Maia-Alfena”	42
4.2.4 Estado do Pavimento	45

4.2.5	Características Geométricas de Planta e de Perfil	50
4.2.6	Volume de Tráfego	50
4.3	Modelos de Regressão para Dados de Contagem	55
4.3.1	Modelo Linear Generalizado Misto com Distribuição de <i>Poisson</i>	55
4.4	Análise de Diagnóstico do Modelo	63
4.5	Validação do Modelo	64
5	Conclusão e Trabalho Futuro	66

Lista de Figuras

Figura 1	Evolução da Sinistralidade Rodoviária em Portugal Continental entre 2010 e 2023	3
Figura 2	Sinistralidade Rodoviária em Portugal Continental por mês	4
Figura 3	Frequência do número de acidentes por segmento	29
Figura 4	Distribuição do número de acidentes por ano	30
Figura 5	Número de acidentes nos sublanços	30
Figura 6	Evolução do número total de acidentes ao longo dos anos	31
Figura 7	Percentagem do número total de acidentes por ano	32
Figura 8	Percentagem do número de acidentes por sublanços em todos os sentidos e vias	33
Figura 9	Evolução do número de acidentes por Ano/Sublanço	34
Figura 10	Percentagem do número de acidentes por sublanço e por sentido	35
Figura 11	Número de acidentes por sublanço e sentido	36
Figura 12	Distribuição percentual do número de acidentes por sublanço e número de vias .	36
Figura 13	Distribuição do número de acidentes por sublanço, sentido e número de vias . .	37
Figura 14	Número de acidentes por sublanço e tipo de via	37
Figura 15	Número de acidentes por sublanço e entradas e saídas de veículos	38
Figura 16	Número de acidentes por ano e autoestrada	39
Figura 17	Número de acidentes por autoestrada e sentido	40
Figura 18	Número de acidentes por autoestrada e via	40
Figura 19	Número de acidentes por autoestrada, sentido e via	41
Figura 20	Distribuição do número de acidentes por autoestrada, <i>sentido e Ano</i>	41
Figura 21	Número total de acidentes por sentido no sublanço Maia-Alfena	42
Figura 22	Percentagem de acidentes por segmento de estrada e sentido no sublanço . . .	43
Figura 23	Número de acidentes no sentido crescente por ano no sublanço Maia-Alfena . .	44
Figura 24	Distribuição do número de acidente por segmentos	44
Figura 25	Distribuição das variáveis <i>IRI_md</i> e <i>IRI_max</i> por ano e sentido	46
Figura 26	Distribuição das variáveis <i>RD_md</i> e <i>RD_max</i> por ano e sentido	47

Figura 27	Distribuição das variáveis <i>MPD_md</i> e <i>MPD_min</i> por ano e sentido	48
Figura 28	Distribuição das variáveis <i>ATRITO_md</i> e <i>ATRITO_min</i> por ano e sentido	49
Figura 29	Distribuição do volume de tráfego médio diário anual total	51
Figura 30	Volume de Tráfego Médio Diário Anual de veículos ligeiros por ano e sublanço . .	52
Figura 31	Volume de Tráfego Médio Diário Anual de veículos pesados por ano e sublanço .	52
Figura 32	Gráfico QQ dos resíduos Quantílicos	63
Figura 33	Gráfico dos resíduos quantílicos vs valores preditos	64
Figura 34	Gráfico dos valores observados e valores previstos	65

Lista de Tabelas

Tabela 1	Descrição das variáveis da base de dados inicial	22
Tabela 2	Descrição das variáveis da base de dados final	27
Tabela 3	Identificação dos sublanços, extensão e autoestrada	28
Tabela 4	Estatísticas descritivas da variável resposta: <i>Acd_t</i>	29
Tabela 5	Distribuição do número total de acidentes por ano e sublanço nos dois sentidos e em todas as vias	32
Tabela 6	Distribuição do número de acidentes pela variável <i>ENT_S</i>	38
Tabela 7	Número de acidentes por autoestrada e por ano	39
Tabela 8	Média das variáveis referentes ao estado do pavimento	45
Tabela 9	Estatísticas descritivas das variáveis <i>IRI_md</i> e <i>IRI_max</i> por ano e sentido	45
Tabela 10	Estatísticas descritivas das variáveis <i>RD_md</i> e <i>RD_max</i> por ano e sentido	47
Tabela 11	Estatísticas descritivas das variáveis <i>MPD_md</i> e <i>MPD_min</i> por ano e sentido	48
Tabela 12	Estatísticas descritivas das variáveis <i>ATRITO_md</i> e <i>ATRITO_min</i> por ano e sentido	49
Tabela 13	Estatísticas descritivas de características geométricas em planta e em perfil	50
Tabela 14	Volume de tráfego médio anual de veículos ligeiros e pesados	53
Tabela 15	Estatísticas descritivas das variáveis <i>TMDAligMEDIA</i> , <i>TMDApesMEDIA</i> , <i>TMDAt-</i> <i>MEDIA</i> por sentido	54
Tabela 16	Modelos nulos com efeitos aleatórios na interseção e no declive da variável <i>Ano</i>	56
Tabela 17	Teste de Razão de Verossimilhança entre os Modelos III e VI	57
Tabela 18	MLGM simples das variáveis referentes ao estado de pavimento	57
Tabela 19	MLGM para as variáveis referentes às características geométricas	58
Tabela 20	Estimativas dos parâmetros para os efeitos fixos, erro padrão, estatística de teste e valores de prova do modelo completo	59
Tabela 21	Estimativas dos parâmetros para os efeitos fixos, erro padrão, estatística de teste e valores de prova do Modelo VII	59
Tabela 22	Estimativas dos parâmetros dos efeitos aleatórios do Modelo VII	60

Tabela 23	Estimativas dos parâmetros para os efeitos fixos, erro padrão, estatística de teste e valores de prova do modelo completo	61
Tabela 24	Estimativas dos parâmetros para os efeitos fixos, erro padrão, estatística de teste e valores de prova do Modelo VIII	61
Tabela 25	Estimativas dos parâmetros dos efeitos aleatórios do Modelo VIII	62

Abreviaturas

A4	Autoestrada de Trás-os-Montes e Alto Douro
A41	Circular Regional Exterior do Porto
AcV	Acidente com vítimas
AIC	Critério de Informação Akaike
ANSR	Autoridade Nacional Segurança Rodoviária
AR	Alinhamento Reto
BIC	Critério de Informação Bayesiano
BN	Binomial Negativa
BNIZ	Binomial Negativa Inflacionada de Zeros
CC	Curva Circular
CCV	Curva Côncava
CCX	Curva Convexa
CL	Curva em Clotóide
EEG	Equações de Estimação Generalizadas
GNR	Guarda Nacional Republicana
IRI	Índice Internacional de Irregularidade
MLG	Modelo Linear Generalizado
MLGM	Modelo Linear Generalizado Misto
MLM	Modelo Linear Misto
MPD	Profundidade Média de Perfil
PSP	Polícia de Segurança Pública
RD	Profundidade das Rodeiras
T	Trainel
VC	Via Central
VD	Via Direita
VE	Via Esquerda

Capítulo 1

Introdução

Em Portugal, como noutros países da União Europeia, a sinistralidade rodoviária é identificada como um problema social atual, de dimensão preocupante, que justifica uma séria intervenção. Associada a uma questão de atitudes e de comportamentos, esta situação reflete uma realidade, com pesadas consequências humanas, sociais e económicas.

Mesmo em termos teóricos, poder-se-á afirmar que, no contexto da circulação rodoviária, é improvável (ou até impossível) eliminar completamente a ocorrência dos acidentes rodoviários, pois não é viável, do ponto de vista humano, eliminar totalmente o risco, mas sim reduzi-lo ao máximo possível.

Os acidentes ocorrem devido a uma combinação de diversos fatores, o que pode levar a uma avaliação incorreta das suas probabilidades. Normalmente, as probabilidades subjetivas diferem das probabilidades objetivas, e quanto maior é essa diferença, maior será o risco na avaliação do risco real.

De acordo com os dados emitidos pela Autoridade Nacional da Segurança Rodoviária (ANSR), responsável pela gestão da segurança rodoviária, apresenta-se na Figura 1 a evolução da sinistralidade rodoviária com vítimas (AcV) em Portugal Continental entre os anos de 2010 e 2023 [1].

Consideram-se os acidentes que tenham ocorrido na via pública ou que nela tenham origem, envolvendo pelo menos um veículo em movimento, e que sejam de conhecimento das entidades fiscalizadoras, como a Guarda Nacional Republicana (GNR) e Polícia de Segurança Pública (PSP), e dos quais resultem pelo menos uma vítima.

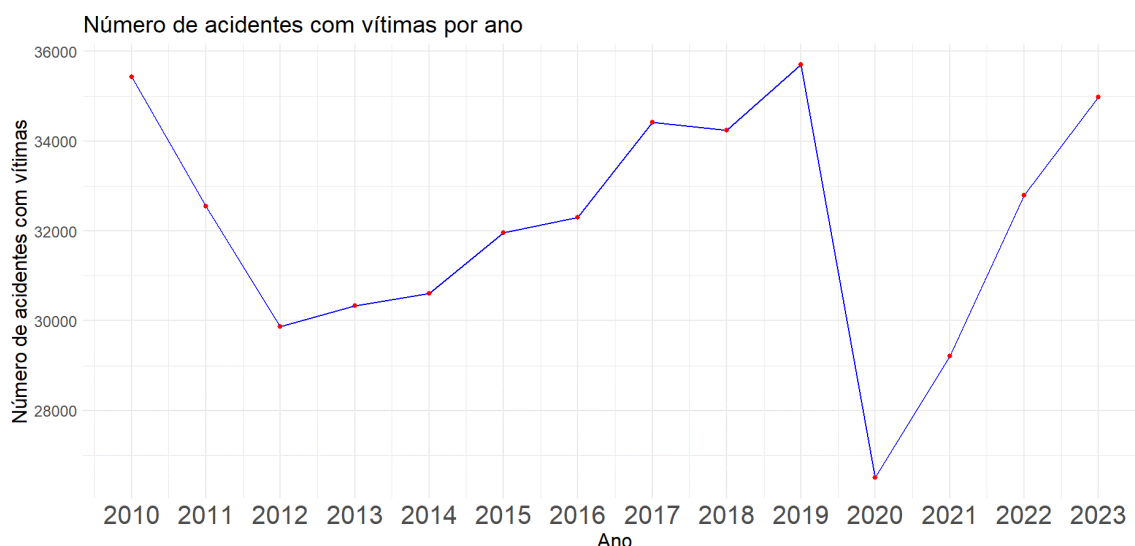


Figura 1: Evolução da Sinistralidade Rodoviária em Portugal Continental entre 2010 e 2023

Verifica-se uma variação considerável no número de acidentes com vítimas ao longo dos anos, destacando-se o ano de 2019 que registou o maior número no período analisado. Entre 2010 e 2012, observa-se uma diminuição no número de acidentes com vítimas, seguida de um aumento gradual até 2017.

Destacam-se os anos de 2020 e 2021 que coincidiram com um período de confinamento devido à pandemia de COVID-19, resultando numa redução da circulação rodoviária e, consequentemente, numa diminuição do número de acidentes.

Considerando que os anos de 2020 e 2021 registaram quebras na circulação rodoviária face a 2019 e, consequentemente na sinistralidade, a Comissão Europeia considerou este o ano como o ano de referência para monitorização das metas de redução do número de vítimas mortais e de feridos graves até 2030, critério que também foi adotado em Portugal na Estratégia Visão Zero, 2030 [2].

Na análise dos valores de sinistralidade em Portugal Continental, em 2023, a ANSR contabilizou 34 974 acidentes de viação com vítimas, das quais 467 vítimas mortais, 2 437 feridos graves e 41 058 feridos ligeiros [1].

Em Portugal Continental comparando com a média da última década (2010 a 2019), com 2023, houve uma tendência crescente no número de acidentes com vítimas (+6,8%), de feridos graves (+9,1%) e de feridos ligeiros (+3,2%). No entanto, o número de vítimas mortais diminuiu (-13,7%).

Face a 2019, em 2023 registaram-se menos acidentes com vítimas (-2,0%), menos vítimas mortais (-1,5%) e menos feridos ligeiros (-5%) . Contudo, houve mais feridos graves (+5,9%).

Salienta-se que, em comparação com o ano de 2022, houve um aumento da circulação rodoviária, em

2023, observando-se aumentos em todos os indicadores, registando-se mais 2186 acidentes com vítimas (+6,7%), mais vítimas mortais (+1,1%), mais feridos graves (+8,6%) e mais feridos ligeiros (+6,8%).

Relativamente ao número de acidentes de viação com vítimas por mês nos anos de 2019, 2022 e 2023 em Portugal Continental (Figura 2), observa-se que, no primeiro trimestre e no último quadrimestre, 2019 foi o ano que registou o maior número de acidentes com vítimas. O ano de 2022 regista o número mais baixo de acidentes com vítimas em todos os meses, à exceção de mês de julho cujo número de acidentes com vítimas é ligeiramente superior em relação a 2019. Nos anos observados, em 2023, houve um aumento do número de acidentes desde o mês de fevereiro atingindo o seu seu valor máximo em Agosto, seguido de uma diminuição até dezembro. Estas variações podem ser influenciadas por diferentes fatores, incluindo as condições meteorológicas ou variações do volume de tráfego.



Figura 2: Sinistralidade Rodoviária em Portugal Continental por mês

No estudo da sinistralidade, o ajuste de equações matemáticas, através de métodos estatísticos, permite relacionar o número de acidentes com diversos fatores, dando origem aos modelos de previsão de acidentes que constituem uma ferramenta fundamental na tomada de decisões.

1.1 Objetivos

A presente dissertação foi desenvolvida no âmbito da Unidade Curricular de Dissertação do Mestrado em Estatística para Ciência de Dados, da Universidade do Minho.

A problemática da sinistralidade rodoviária representa um desafio para a Saúde Pública. Apesar da redução no número de acidentes rodoviários nos países desenvolvidos nas últimas décadas, é primordial compreender os fatores que contribuem para esses acidentes, com o objetivo de reduzir a severidade dos

danos às vítimas.

Dado o contexto apresentado e a necessidade de abordagens que promovam a segurança rodoviária, o objetivo deste trabalho incide na construção de modelos, mais propriamente, em Modelos Lineares Generalizados Mistos (MLGM), envolvendo dados longitudinais, com recurso à regressão de *Poisson*, uma vez que se tratam de modelos para dados de contagem.

Pretende-se não apenas compreender as suas potencialidades, mas também estabelecer uma relação entre o número de acidentes rodoviários e outras variáveis, com o objetivo de por um lado, avaliar a influência dessas variáveis na frequência esperada do número de acidentes e, por outro, fornecer previsões de sinistralidade.

1.2 Estrutura da Tese

Esta dissertação encontra-se segmentada em diversos capítulos. Esta organização permite ter uma ideia geral da investigação, remetendo os detalhes do estudo para o corpo integral do capítulo.

O presente Capítulo apresenta uma introdução ao tema a ser estudado com uma concisa apresentação do mesmo, e os objetivos do estudo do trabalho.

O Capítulo 2 inclui uma breve revisão da literatura relacionada com os modelos lineares, em particular, os modelos mais comuns na modelação de dados de contagem.

São ainda apresentados alguns exemplos de aplicação de modelos estatísticos no estudo de Acidentes Rodoviários.

O Capítulo 3 apresenta os conteúdos teóricos relacionados com os Modelos Lineares Generalizados Mistos, e apresenta-se o modelo de regressão de *Poisson* com efeitos aleatórios para dados de contagem.

A aplicação das metodologias aos dados é apresentada no Capítulo 4. Inicialmente o capítulo refere-se à apresentação, descrição e análise exploratória do conjunto de dados. De seguida são aplicados os modelos de regressão para dados de contagem, efetua-se a comparação/seleção dos modelos ajustados, e por fim, o diagnóstico do modelo e a sua validação.

No último Capítulo, encontram-se as conclusões deste trabalho e as sugestões para temas a desenvolver no futuro.

Capítulo 2

Revisão da Literatura

Neste capítulo apresenta-se uma revisão da literatura sobre modelos estatísticos de regressão.

2.1 Modelos Estatísticos

Um modelo estatístico é um modelo que contém uma componente probabilística. No quotidiano, diversos problemas surgem com o objetivo de analisar a influência que uma ou mais variáveis, designadas por variáveis explicativas, exercem sobre uma variável, denominada de variável resposta, variável dependente ou de interesse. De um modo geral, este problema é abordado aplicando modelos de regressão, que relacionam a variável de interesse com as variáveis explicativas [3].

A seleção de modelos é de extrema importância em todas as investigações em que se deseja encontrar um modelo que seja o mais parcimonioso possível, e que consiga descrever com certo grau de precisão os dados observados [4].

O modelo estatístico clássico usado na análise de regressão, o Modelo Linear, surgiu no início do século XIX, através de Legendre e Gauss. De forma sucinta, este modelo exprime o valor esperado da variável resposta como uma combinação linear das variáveis explicativas, e é aplicado na situação em que a variável resposta segue uma distribuição normal.

Para dar resposta a situações que não se adequavam ao contexto dos Modelos Lineares, foram desenvolvidos os Modelos Lineares Generalizados (MLG), introduzidos por Nelder and Wedderburn Nelder e Wedderburn [5]. Segundo estes autores, a variável resposta deve seguir uma distribuição que pertença à família exponencial. Estes modelos apresentam uma estrutura linear nas variáveis explicativas e relacionam, de forma não linear, o valor esperado da variável resposta com a estrutura linear das variáveis explicativas, isto é, não modelam necessariamente o valor esperado da variável resposta, mas sim uma transformação dessa variável.

Os Modelos Lineares Mistos (MLM) para dados Gaussianos são uma extensão dos Modelos Lineares. Estes modelos incluem além da variação aleatória do acaso, pelo menos mais uma fonte de variação aleatória (efeitos aleatórios não observáveis), para modelar estruturas de dependência de dados hierár-

quicos ou longitudinais, ambas de forma linear. Estes modelos possibilitam uma abordagem mais realista das características de muitos conjuntos de dados, pela sua flexibilidade para modelar as estruturas de dependência e de variância não constante.

Em [6], os autores desenvolveram os Modelos Lineares Generalizados Mistos (MLGM), uma extensão dos MLG que permitem a adição de componentes de variabilidade devida a efeitos não observados, ou seja, a variabilidade não explicada pelos efeitos fixos. Os MLGM são, modelos que incluem efeitos aleatórios no preditor linear, além dos efeitos fixos, permitindo modelar a estrutura de correlação entre observações que pertencem ao mesmo indivíduo [7].

Estes modelos são particularmente relevantes na presença de heterogeneidade dos dados envolvidos e usados nas análises longitudinais, em que as mesmas unidades são medidas repetidamente ao longo do tempo (os MLGM podem levar em consideração a correlação entre as medidas repetidas de uma mesma unidade). São desenvolvidos para permitir que a variável resposta siga qualquer distribuição da família exponencial, e por isso são adequados para modelar dados de contagem.

Os MLGM têm como objetivo descrever as alterações do valor esperado da variável resposta de cada indivíduo e a relação destas alterações com as covariáveis [8]. Assim, estes modelos têm como finalidade fazer inferências para o indivíduo e não para a população.

2.2 Modelos Estatísticos para Acidentes Rodoviários

Os modelos de regressão para dados de contagem são amplamente utilizados em diversas áreas de estudo.

Para modelar o valor esperado do número de acidentes, a aplicação de modelos estatísticos tem sido objeto de estudo há muitos anos, com o objetivo de identificar os fatores que influenciam o risco de um acidente.

De seguida, pretende-se apresentar como abordagens metodológicas evoluíram ao longo do tempo no âmbito de estudo de ocorrências de acidentes rodoviários.

Em [9] os autores consideram o modelo de regressão de *Poisson*, dado que a variável dependente é um número inteiro não negativo, como um dos modelos adequados para modelar os dados dos acidentes rodoviários. Embora este modelo tenha servido como ponto de partida, os investigadores constataram que os dados de acidentes apresentam características que dificultam a aplicação do modelo de *Poisson*, uma vez que este modelo pressupõe que o valor médio e a variância da variável resposta, condicionados pelos valores das variáveis explicativas, sejam iguais. Na prática, a variância tende a exceder o valor

médio (sobredispersão), visto que há heterogeneidade entre indivíduos não captada pelas variáveis explicativas. Essa situação de sobredispersão limita a aplicação dos modelos de *Poisson* podendo-se obter resultados tendenciosos. Para superar esta situação, o modelo de regressão Binomial Negativa (BN) é uma alternativa ao modelo de regressão de *Poisson*. Uma outra característica nos dados de acidentes rodoviários, é a tendência de apresentarem muitos valores iguais a zero. Para lidar com esta situação, os autores [9] abordam o modelo Inflacionado de Zeros. Este modelo foi desenvolvido para lidar com dados que apresentam muitos valores iguais a zero (mais zeros do que seria de esperar num modelo de *Poisson* ou Binomial Negativa). Uma descrição teórica destes modelos pode ser encontrada em [10], e [11].

Em muitos casos, os dados de acidentes rodoviários são frequentemente agregados ao longo do tempo (por exemplo, acidentes por mês) e/ou espaço (acidentes ao longo de um segmento de estrada), o que pode gerar problemas de heterogeneidade não observada, que podem depender do tempo ou do espaço. Estudos têm controlado a heterogeneidade não observada, como correlação espacial e temporal, aplicando modelos de efeitos aleatórios [9].

A correlação entre observações também pode ser abordada utilizando Modelos hierárquicos/multiníveis e modelos de parâmetros aleatórios.

De acordo com [12], estes últimos são uma extensão dos modelos de efeitos aleatórios, que podem potencialmente capturar a heterogeneidade não observada permitindo que os parâmetros variem entre as observações (como um segmento de estrada) ou sejam fixos dentro de um grupo de observações mas variem entre grupos.

Em [13], os autores propõem a aplicação do modelo Binomial-Lindley Negativo de parâmetros aleatórios agrupados (G-RPNB-L) para explicar a heterogeneidade não observada em dados de acidentes com muitas observações iguais a zero. As diferenças em vários fatores temporais e espaciais resultam em variações dos coeficientes do modelo entre diferentes grupos de observações. Um modelo de parâmetros aleatórios agrupados é uma estratégia para explicar essa heterogeneidade não observada.

Em [14], os autores estudaram os fatores que influenciaram a frequência de acidentes nas estradas rurais da Colômbia.

Neste país, as condições orográficas são o principal critério para a concepção de estradas, não havendo estudos que avaliem a sua influência nos acidentes rodoviários. Dessa forma, este estudo avaliou diversas variáveis, tais como, o tipo de terreno e a sua influência no valor esperado do número de acidentes. Para tal, estimaram modelos de regressão Binomial Negativa Inflacionada de Zeros, Modelos de regressão Binomial Negativa, Modelos Lineares Generalizados Mistos, e realizaram uma análise comparativa dos resultados.

O MLGM mostrou-se ligeiramente melhor que o modelo de regressão Binomial Negativa Inflacionada de Zeros, mas o melhor dos três modelos apresentados foi o modelo de regressão Binomial Negativa, que apresentou maior função de log-verossimilhança.

Neste caso, a utilização de um modelo mais complexo, MLGM não proporcionou um melhor ajuste.

Diante do exposto na literatura, a escolha do Modelo Linear Generalizado Misto com distribuição de *Poisson* neste trabalho deve-se à natureza dos dados, à variável resposta que é uma variável de contagem, e à necessidade de modelar os dados que apresentam uma estrutura agrupada, com várias observações ao longo do tempo para os mesmos segmentos de estrada.

Neste contexto, os MLGM, envolvendo dados longitudinais, são capazes de capturar a complexidade e a heterogeneidade dos dados envolvidos. Estes modelos permitem incorporar tanto os efeitos fixos como aleatórios (captam a variabilidade não observada entre os segmentos e o tempo), que variam por grupo, permitindo modelar a relação entre diversas variáveis explicativas e o número de acidentes. A escolha da distribuição de *Poisson* está relacionada com variável de interesse uma vez que esta é de natureza discreta e sendo adequada para dados de contagem. Os modelos de regressão de *Poisson* tradicionais não conseguem lidar adequadamente com a estrutura agrupada dos dados, ignorando as correlações entre observações dentro dos mesmos segmentos ou anos, levando a resultados menos fidedignos. Por outro lado, o modelo de regressão Binomial Negativa é também muito aplicados em dados de contagem na presença de sobredispersão. Por fim, os Modelos Inflacionados de Zeros, que são apropriados em casos onde há muitos zeros nos dados, poderiam ser considerados. No entanto, as aplicações destes modelos será realizada em trabalhos futuros. A escolha do Modelo Linear Generalizado Misto com a distribuição de *Poisson* é corroborada pela sua capacidade de ajustar as dependências entre observações de contagens e, assim, fornecer uma melhor qualidade de ajustamento aos dados analisados, em comparação com outros modelos, como o de *Poisson* tradicional ou o modelo de regressão Binomial Negativa.

Capítulo 3

Metodologia

Em muitos estudos científicos é comum fazer-se pesquisas com o objetivo de estudar a relação entre a variável resposta e as variáveis explicativas medidas em unidades experimentais pertencentes a uma ou várias populações. Destacam-se duas principais estratégias de recolha de dados: a primeira, denominada estudo transversal, que envolve uma única observação (realizada num momento específico) da variável resposta para cada unidade experimental de uma amostra da população de interesse, a segunda, várias observações da variável resposta são realizadas sobre a mesma unidade experimental ao longo de uma condição de avaliação, ou seja, quando a mesma característica é medida de forma sequencial ao longo do tempo, resultando no estudo de dados longitudinais.

Existem técnicas para a modelação de experiências com dados longitudinais, sendo a metodologia aplicada neste trabalho, os Modelos Lineares Generalizados Mistos (MLGM), que modelam os dados longitudinais com qualquer variável resposta que pertença à família exponencial.

Quanto à implementação dos métodos estatísticos, recorre-se ao ambiente *R* (R Core Team) [15]. Esta é uma linguagem de programação e oferece uma série de técnicas para análise de dados, inferência estatística e permite a apresentação gráfica dos resultados obtidos.

Existem várias bibliotecas e funções específicas para aplicar os MLGM. A biblioteca *lme4* [16], permite a modelação de dados com estruturas de correlação e é muito utilizada na análise de dados longitudinais. A função para ajustar um MLGM da biblioteca *lme4* é a *glmer()*. Existem outras funções que podem ser utilizadas no ajustamento, como por exemplo a função *glmm()* da biblioteca *glmmML* [17] e a função *glmmPQL()* da biblioteca *MASS* [18]. A escolha da função utilizada recaiu sobre a função *glmer()* por ser a mais utilizada nos diversos estudos.

A biblioteca *Performance* [19] é usada para calcular os intervalos de confiança dos efeitos aleatórios.

3.1 Dados Longitudinais

Os estudos longitudinais são essenciais para avaliar as mudanças de uma determinada característica

ao longo do tempo em cada indivíduo, assim como, para analisar os efeitos dos fatores que influenciam essas mudanças. Se pensarmos que as observações se efetuam ao longo do tempo, obtém-se uma sequência temporal de duas ou mais observações para cada indivíduo [20].

Estudos com dados longitudinais ganharam grande destaque nas últimas décadas, especialmente devido ao rápido progresso da tecnologia da informação, que tornou possível armazenar e analisar de modo mais rápido e preciso grandes bases de dados [21].

Em [22], os autores consideram que este tipo de dados tem grande aplicação em modelos de regressão, onde a variável resposta pode ser contínua ou discreta.

Exemplos de dados longitudinais incluem dados relacionados com o número de acidentes em diversas regiões ao longo de vários anos [22]. Esta circunstância implica que as observações de um indivíduo ao longo do tempo tendem a estar correlacionadas e, por conseguinte, os dados longitudinais requerem métodos estatísticos específicos que tenham em conta essa correlação existente, para que as inferências realizadas sejam válidas [20].

Em [20], os autores consideraram três tipos de modelos para a análise de dados longitudinais: (i) o modelo marginal que usa Equações de Estimação Generalizadas (EEG), (ii) o modelo com efeitos aleatórios (MLGM) e o (iii) modelo de transição. A principal diferença entre estes modelos centra-se no objetivo da inferência que se pretende efetuar.

3.2 Modelo Linear Misto

A designação "modelo misto" advém do facto de que o modelo contém parâmetros de efeitos fixos (parâmetros associados a toda a população) e parâmetros de efeitos aleatórios (parâmetros associados aos indivíduos selecionados aleatoriamente da população) [23].

A necessidade de incluir efeitos aleatórios pode surgir por várias razões, entre elas:

- As unidades experimentais nas quais os dados são medidos podem estar agrupadas, e os dados do mesmo grupo são correlacionados.
- Medidas repetidas da mesma unidade experimental são correlacionadas.

A crescente popularidade dos modelos mistos surge, principalmente, pela flexibilidade que oferecem para analisar dados agrupados e dados com medidas repetidas. Pode-se modelar ambos os tipos de variação:

- Efeitos Fixos.

São as variações que se esperam em todas as observações. São variáveis que se podem medir diretamente.

- Efeitos Aleatórios.

São componentes de um modelo estatístico que capturam variações que não são explicadas pelos efeitos fixos. Essas variações podem ser devido a fatores não medidos em certos grupos dentro dos dados.

O Modelo Linear Misto (MLM) é dado por :

$$\mathbf{Y}_i = \mathbf{X}_i\beta + \mathbf{Z}_i\mathbf{b}_i + \epsilon_i, \quad (3.1)$$

com $i = 1, \dots, n$, onde $\mathbf{Y}_i = (Y_{i1}, \dots, Y_{iT_i})^T$ é o vetor $T_i \times 1$ das variáveis resposta para o indivíduo i , β é o vetor $p \times 1$ dos efeitos fixos (parâmetros desconhecidos); $\mathbf{X}_i$ é a matriz $T_i \times p$ de covariáveis dos efeitos fixos; $\mathbf{b}_i$ é o vetor $q \times 1$ dos efeitos aleatórios; $\mathbf{Z}_i$ é a matriz $T_i \times q$ de covariáveis dos efeitos aleatórios e ϵ_i é o vetor $T_i \times 1$ dos erros aleatórios dentro do grupo. As condições subjacentes ao modelo são:

$$\mathbf{b}_i \sim N(0, \mathbf{D})$$

$$\epsilon_i \sim N(0, \Sigma_i)$$

$\mathbf{b}_i$ e ϵ_i independentes para diferentes grupos i e entre si

com $i = 1, \dots, n$, onde $\mathbf{D}$ é uma matriz de dimensão $q \times q$, e é a matriz de variância-covariância dos efeitos aleatórios presentes no vetor β , e Σ_i é uma matriz de dimensão $T_i \times T_i$, e é a matriz de variância-covariância dos erros aleatórios, sendo ambas definidas positivas.

Tem-se que:

$$E(\mathbf{Y}_i) = \mathbf{X}_i\beta \quad \text{e} \quad \mathbf{V}_i = V(\mathbf{Y}_i) = \mathbf{Z}_i\mathbf{D}\mathbf{Z}_i^T + \Sigma_i, \quad (3.2)$$

A matriz $\mathbf{V}_i$ de dimensão $T_i \times T_i$ e definida positiva, representa a variação entre-grupos (inter-indivíduos) [8].

É importante notar que, quando $\Sigma_i = \sigma^2 I$ e $\mathbf{Z}_i = 0$, o modelo misto é simplificado para um modelo linear.

3.3 Modelo Linear Generalizado Misto

Uma das extensões dos Modelos Lineares Mistos (MLM), são os Modelos Lineares Generalizados Mistos, para modelar variáveis resposta pertencentes à família exponencial.

Os MLGM são ainda uma extensão dos Modelos Lineares Generalizados (MLG) que incluem efeitos aleatórios no preditor linear, além dos efeitos fixos. A introdução destes efeitos aleatórios no preditor linear permite modelar a estrutura de correlação entre observações que pertencem ao mesmo indivíduo.

Na análise de acidentes rodoviários, este tipo de modelos é especialmente útil para avaliar como certos fatores influenciam a frequência ou a gravidade dos acidentes ao longo do tempo. Isto é importante, porque as observações dentro do mesmo grupo (por exemplo, acidentes no mesmo sublanço) podem estar correlacionadas. Neste caso, pode-se incluir efeitos aleatórios, por exemplo, sublanços, autoestradas ou segmentos da autoestrada, reconhecendo que há uma correlação intrínseca entre acidentes que ocorrem em locais próximos. Por exemplo, cada segmento da autoestrada pode ter características únicas que influenciam os acidentes. Essas características podem não estar completamente capturadas pelas variáveis fixas/preditoras que se medem.

A inclusão de efeitos aleatórios permite que o modelo reconheça e se ajuste à heterogeneidade inerente aos dados. Ao ajustar um Modelo Linear Generalizado Misto, o modelo estima os efeitos fixos, as variáveis conhecidas que afetam a frequência de acidentes e os efeitos aleatórios, que capturam as variações específicas. Por outras palavras, os Modelos Lineares Generalizados Mistos (MLGM) permitem uma compreensão mais profunda das inter-relações presentes nos dados de acidentes rodoviários, relações essas que frequentemente são perdidas em modelos mais simples que assumem independência entre as observações.

Ao nível computacional e na interpretação de parâmetros estimados, estes modelos são complexos devido à inclusão de efeitos aleatórios.

3.3.1 Estrutura do Modelo Linear Generalizado Misto

Seja y_{it} a resposta observada do indivíduo i para o tempo t realização da variável resposta Y_{it} , onde $t = 1, \dots, T_i$ e $i = 1, \dots, n$, sendo $\mathbf{x}_{it} = (x_{it1}, \dots, x_{itp})^T$ o vetor $p \times 1$ das covariáveis associado a cada vetor resposta, Y_{it} e $\boldsymbol{\beta} = (\beta_1, \beta_2, \dots, \beta_p)^T$ o vetor $p \times 1$ dos parâmetros de regressão desconhecidos. Considere $\mathbf{z}_{it}$ um vetor $q \times 1$ de covariáveis (usualmente um subconjunto de $\mathbf{x}_{it}$) associado a um vetor $q \times 1$ de efeitos aleatórios $\mathbf{b}_i$ [8].

Os MLGM podem ser especificados da seguinte forma :

- **1.** Dado $\mathbf{b}_i$, as variáveis resposta $Y_{i1}, \dots, Y_{iT_i}$ são mutuamente independentes e seguem uma distribuição pertencente à família exponencial com função densidade:

$$f(y_{it} | \mathbf{b}_i) = \exp \left[\frac{\omega_{it}}{\phi} (y_{it}\theta_{it} - c(\theta_{it})) + d(y_{it}, \phi) \right]. \quad (3.3)$$

O valor médio e a variância condicional são expressos, respetivamente, por:

$$E(Y_{it} | \mathbf{b}_i) = \mu_{it} = c'(\theta_{it}) \quad (3.4)$$

$$\text{Var}(Y_{it} | \mathbf{b}_i) = \sigma_{it} = c''(\theta_{it}) \frac{\phi}{\omega_{it}} \quad (3.5)$$

Tal como nos MLG, assume-se que:

$$g(\mu_{it}) = \mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \mathbf{b}_i \quad (3.6)$$

$$\sigma_{it} = V(\mu_{it}) \frac{\phi}{\omega_{it}} \quad (3.7)$$

onde $g(\cdot)$ é uma função de ligação pré-especificada e V é a função de variância conhecida, ϕ é um parâmetro de dispersão e ω_{it} é uma constante conhecida. Uma vez fixados os efeitos aleatórios $\mathbf{b}_i$, as variáveis resposta assumem-se mutuamente independentes.

- **2.** A média condicional da variável resposta Y_{it} , depende dos efeitos fixos e dos efeitos aleatórios. Deste modo, o Modelo Linear Generalizado Misto pode ser escrito por [8]:

$$g(E(Y_{it} | \mathbf{b}_i)) = \eta_{it} = \mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \mathbf{b}_i \quad (3.8)$$

- **3.** Qualquer distribuição multivariada poderia ser considerada para $\mathbf{b}_i$. Contudo, na prática, assume-se que a distribuição comum aos efeitos aleatórios $\mathbf{b}_i$ é a Gaussiana Multivariada com valor esperado zero e matriz de variância-covariância $\mathbf{D}$ [20]. Salienta-se o facto dos efeitos aleatórios $\mathbf{b}_i$ se assumirem independentes das covariáveis.

Modelo Linear Generalizado Misto para Resposta sob a forma de contagem

Ao seleccionar o MLGM mais adequado para os dados, é essencial considerar a distribuição da variável de interesse.

Para modelar o número de ocorrências de um determinado acontecimento pode-se aplicar o modelo de regressão com variável resposta seguindo a distribuição de *Poisson*.

Considere-se que Y_{it} , a variável resposta do indivíduo i no instante t , segue uma distribuição de *Poisson*. Nesta situação tem-se:

- **1.** A variável resposta Y_{it} , condicional ao vetor dos efeitos aleatórios $\mathbf{b}_i$, são independentes e seguem uma distribuição de *Poisson*, com

$$E(Y_{it}|\mathbf{b}_i) = Var(Y_{it}|\mathbf{b}_i) \quad (\phi = 1). \quad (3.9)$$

- **2.** Pretende-se modelar o valor médio como combinação linear das variáveis explicativas. Poderia escrever-se um modelo linear da forma:

$$\mu_{it} = \mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \mathbf{b}_i \quad (3.10)$$

Contudo, este modelo é desajustado uma vez que o preditor linear, do lado direito, pode assumir qualquer valor real enquanto que o termo do lado esquerdo (o valor médio de *Poisson*) só pode tomar valores não negativos. Uma solução para este problema passa por considerar a transformação logarítmica como função de ligação do modelo linear generalizado através da equação:

$$\log(\mu_{it}) = \mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \mathbf{b}_i \quad (3.11)$$

Na presença de dados de contagem *log* é a função de ligação canónica.

Aplicando a inversa da função logarítmica a ambos os lados da equação (3.11) obtém-se o número esperado de contagem, dado por:

$$\mu_{it} = \exp(\mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \mathbf{b}_i). \quad (3.12)$$

- **3.** Por fim, assume-se que o vetor dos efeitos aleatórios, $\mathbf{b}_i$, segue uma distribuição Gaussiana de valor médio zero e matriz de variância-covariância $\mathbf{D}$, de dimensão $q \times q$.

3.3.2 Interpretação dos Parâmetros do Modelo

A interpretação dos parâmetros fixos nos MLGM será assim afetada pelos efeitos aleatórios presentes no modelo, como explicam Cabral e Gonçalves [8].

Com efeito, nos MLGM,

$$E(Y_{it}) = E\{E(Y_{it}|\mathbf{b}_i)\} = E[g^{-1}(\mathbf{x}_{it}^T \boldsymbol{\beta} + \mathbf{z}_{it}^T \mathbf{b}_i)] \quad (3.13)$$

que, em geral, não pode ser simplificado devido à presença de funções não lineares em $g^{-1}(\cdot)$ e, portanto,

$$E[g(Y_{it}|\mathbf{b}_i)] \neq g[E(Y_{it}|\mathbf{b}_i)] \quad (3.14)$$

Note-se que no modelo linear a introdução de efeitos aleatórios não altera a interpretação dos parâmetros dos efeitos fixos, dado que a desigualdade (3.14) não é verificada. Existe apenas alteração na interpretação dos parâmetros dos efeitos fixos quando a função de ligação canônica dos MLGM não é linear [24]. De acordo com [8] nos MLGM podemos considerar as seguintes interpretações para os parâmetros fixos e para a componente aleatória do modelo:

- (i) **Parte Fixa:** As componentes β_j , $j = 0, 1, \dots, p$, do vetor β têm uma interpretação em termos específicos do indivíduo. Elas representam a influência das variáveis explicativas de interesse na alteração da resposta esperada de um determinado indivíduo. Estes coeficientes, β , são designados por *coeficientes de regressão específicos do indivíduo* dado que a interpretação de β_j depende dos efeitos aleatórios do i -ésimo indivíduo assumirem um valor fixo (não observado).
- (ii) **Parte Aleatória:** Uma forma de interpretar as estimativas das variâncias dos efeitos aleatórios passa por construir percentis dos efeitos aleatórios baseados na hipótese Gaussiana. Com base nos percentis obtidos e aplicando a inversa da função de ligação calcula-se os limites de variação dos valores esperados condicionados a μ_{it} . Outra maneira é considerar a representação gráfica dos perfis de cada indivíduo.

3.3.3 Estimação dos Parâmetros do Modelo

Assim como nos Modelos Lineares Generalizados, a estimação dos parâmetros nos Modelos Lineares Generalizados Mistos também recorre ao Método da Máxima Verossimilhança. Nesta situação, a estimação de β e dos parâmetros da matriz $\mathbf{D}$ é realizada através da maximização da função de verossimilhança, obtida através da integração da função de verossimilhança sobre os efeitos aleatórios $\mathbf{b}_i$. Como explicitado em [8], a contribuição do i -ésimo indivíduo para a função de verossimilhança é dada por:

$$L^A(\beta, \phi, \mathbf{D}) = \int \prod_{t=1}^{T_i} f(y_{it}|\mathbf{b}_i, \beta, \phi) f(\mathbf{b}_i|\mathbf{D}) d\mathbf{b}_i, \quad (3.15)$$

onde $\prod_{t=1}^{T_i} f(y_{it} | \mathbf{b}_i, \beta, \phi)$ diz respeito à função verossimilhança do modelo com efeitos aleatórios e $f(\mathbf{b}_i|\mathbf{D})$ representa a distribuição dos efeitos aleatórios, assumindo como sendo uma distribuição normal multivariada. O expoente A na função de verossimilhança indica que se trata da função de verossimilhança do modelo com efeitos aleatórios [8]. A função de verossimilhança para a amostra, dos n indivíduos é

expressa por:

$$L^A(\beta, \phi, \mathbf{D}) = \prod_{i=1}^n L_i(\beta, \phi, \mathbf{D}) = \prod_{i=1}^n \int L_i^F(\beta, \phi, \mathbf{D}) f(\mathbf{b}_i | \mathbf{D}) d\mathbf{b}_i, \quad (3.16)$$

onde o expoente F na função de verosimilhança indica que se trata da função de verosimilhança do modelo com efeitos fixos. Fazendo o logaritmo da função de verosimilhança, tem-se, naturalmente,

$$l^A(\beta, \phi, \mathbf{D}) = \sum_{i=1}^n \log L_i^A(\beta, \phi, \mathbf{D}) \quad (3.17)$$

Maximizando a função (3.17), obtém-se os estimadores de máxima verosimilhança dos parâmetros β e da matriz variância-covariância dos efeitos aleatórios $\mathbf{b}_i$. Contudo, ao contrário do que acontece nos MLG, onde estes parâmetros são facilmente estimados através do método iterativo de mínimos quadrados ponderados, nos Modelos Lineares Generalizados Mistos este processo não permite calcular estas estimativas [11]. Neste tipo de modelos, a maximização da função log-verosimilhança é bastante complexa e o problema, geralmente, não tem uma solução analítica. Consequentemente, para que se possa inferir com base na verosimilhança, tem de se recorrer ao cálculo numérico.

Como qualquer método, o cálculo numérico possui limitações, podendo tornar a obtenção das estimativas um processo computacionalmente intensivo, aumentando o seu tempo de resposta quando se está na presença de grandes amostras e/ou quando a estrutura do modelo é complexa.

Como se refere em [8], de forma a resolver este problema têm sido propostas várias aproximações numéricas:

- **1. Método de Aproximação dos Dados**

Consiste na decomposição dos dados em valor médio e uma componente apropriada, com um desenvolvimento em série de *Taylor* em torno do valor médio que é uma função não linear do preditor. O objetivo é linearizar a função de ligação de forma a realizar inferências para os parâmetros de regressão. Inserem-se nesta última metodologia o método de quase-verosimilhança penalizada (PQL), proposto por [6] e o método de quase-verosimilhança marginal apresentado por [25].

- **2. Método de Aproximação da Função a Integrar**

Tem como base a aproximação de *Laplace*. O método de *Laplace* consiste em aproximar integrais na forma:

$$I = \int e^{Q(\mathbf{b})} d\mathbf{b} \quad (3.18)$$

em que $Q(\mathbf{b})$ é uma função limitada conhecida, de variável $\mathbf{b}$, com dimensão q .

- **3. Método de Aproximação do Integral**

Dos vários métodos de integração, salientam-se, o método numérico da quadratura de Gauss-Hermite e o método adaptativo da quadratura de Gauss-Hermite. Estes métodos baseiam-se na aproximação de integrais na forma:

$$\int f(z)\phi(z)dz \quad (3.19)$$

em que $f(z)$ é uma função densidade conhecida e $\phi(z)$ é a função densidade de uma distribuição Gaussiana multivariada.

3.3.4 Inferência-Testes de Hipóteses e Seleção do Modelo

Para aplicar a metodologia dos MLGM a um conjunto de dados, é necessário, após a formulação do modelo considerado adequado, realizar inferências sobre esse modelo. A inferência nos MLGM é baseada na Função de Verossimilhança e nas distribuições assintóticas dos estimadores de máxima verossimilhança. Com efeito, não só o método da máxima verossimilhança é o método de eleição para estimar os parâmetros β e os parâmetros em $\mathbf{D}$, como também os testes de hipóteses sobre os parâmetros do modelo são baseados na verossimilhança. Assumindo que o modelo é apropriado, os estimadores obtidos têm assintoticamente uma distribuição de Gauss com valor médio igual ao parâmetro a estimar e variância dada pelo inverso da matriz de informação de Fisher.

Nos MLGM, o teste de Wald pode ser utilizado para comparar as estimativas padronizadas com distribuição $N(0, 1)$. Os testes de Razão de Verossimilhança são geralmente usados para comparar a estrutura fixa de dois modelos encaixados que tenham os mesmos efeitos aleatórios. Assim, para testar a nulidade de um subvetor de r componentes de β

$$H_0 : \beta_r = 0 \quad \text{vs} \quad H_1 : \beta_r \neq 0$$

utiliza-se a estatística de teste

$$2(\log L_1 - \log L_0) \sim \chi^2_{k_1 - k_0},$$

em que L_1 corresponde à verossimilhança do modelo mais geral, com k_1 parâmetros e L_0 corresponde à verossimilhança do modelo encaixado, com k_0 parâmetros. Sob a hipótese nula de que o modelo restrito é mais adequado (ou seja de que os $r = k_1 - k_0$ parâmetros adicionais são iguais a zero) a estatística de teste seguem uma distribuição assintótica Qui-quadrado com $k_1 - k_0$ graus de liberdade.

Há ainda o interesse em se testar a necessidade de quais os efeitos aleatórios a incluir no modelo. Para um dado β , a comparação de modelos encaixados com diferentes estruturas de variância-covariância dos efeitos aleatórios, corresponde a testar a hipótese nula de haver q efeitos aleatórios *versus* a hipótese alternativa de haver $q + k$ efeitos aleatórios, ou testar, por exemplo, a hipótese nula de os q efeitos aleatórios serem independentes *versus* a hipótese alternativa de não o serem.

Quando dois modelos não são encaixados, o teste de Razão de Verossimilhança não pode ser utilizado. Nestes casos, é aconselhável usar um critério que avalie a quantidade de informação captada pelo modelo a partir dos dados, em contraste com o ruído que o modelo não consegue explicar. Em [26], o autor afirma que o melhor critério de seleção de modelos é o Critério de Informação Akaike (AIC), dado por:

$$AIC = -2 \log L + 2k,$$

sendo L o valor máximo da função de verossimilhança, e k o número de parâmetros a estimar.

Dado um conjunto de modelos ajustados aos dados, o modelo escolhido é aquele que apresenta o menor valor de AIC. O AIC é uma medida de qualidade do ajuste que penaliza modelos com muitos parâmetros. No entanto, é importante destacar que o AIC não indica, em termos absolutos, se um modelo se ajusta bem aos dados, mas apenas se é preferível em relação a outro modelo comparado. Um outro critério de informação utilizado é Critério de Informação Bayesiano (BIC) dado por :

$$BIC = -2 \log L + 2k \log(N)$$

sendo L o valor máximo da função de verossimilhança, k o número de parâmetros estimados e N o número de observações.

Vários autores preferem utilizar o BIC em vez do AIC uma vez que este, para além de penalizar o número de parâmetros do modelo, penaliza ainda o número de observações. Contudo, nesta dissertação utilizar-se-á o critério AIC para a seleção de modelos.

A escolha de modelos é fundamental em todas as investigações em que se pretende encontrar uma equação que seja mais parcimoniosa possível e que consiga descrever com certo grau de precisão os dados observados [4].

3.3.5 Análise dos Resíduos

Os resíduos são utilizados para explorar a adequação do ajustamento do modelo aos dados no que diz respeito à escolha da função de variância, da função de ligação e dos termos no preditor linear. Para além

disso, os resíduos indicam também a presença de valores discrepantes nos dados, isto é, observações mal ajustadas que não são bem explicadas pelo modelo [27].

Os resíduos são nada mais do que a diferença entre valores observados e os ajustados pelo modelo. O i -ésimo resíduo é dado por:

$$\hat{e}_i = y_i - \hat{y}_i, \quad i = 1, 2, \dots, n$$

Por exemplo, nos modelos lineares clássicos, os resíduos padronizados são utilizados para verificar a homocedasticidade, a presença de pontos discrepantes, a normalidade e a independência dos erros.

Nos MLGM, a análise de resíduos é realizada recorrendo essencialmente a métodos gráficos.

Em [28], o autor apresenta um método baseado em simulações para criar resíduos escalonados (também chamados de resíduos quantílicos) facilmente interpretáveis para modelos de efeitos mistos. Os resíduos resultantes são padronizados para valores entre 0 e 1 e podem ser interpretados de forma tão intuitiva quanto aos resíduos de uma regressão linear.

Esta metodologia encontra-se implementada no *package* DHARMA do *software* R que fornece uma série de funções gráficas e testes para problemas de especificação incorreta do modelo, como sobre-dispersão, subdispersão, autocorrelação espacial e temporal. Os detalhes estatísticos e teóricos para realizar estas simulações e criar resíduos escalonados podem ser encontrados nas referências do *package* DHARMA [28].

Capítulo 4

Aplicações nos Acidentes Rodoviários

Neste capítulo, pretende-se estudar o efeito das diferentes variáveis que definem o estado do pavimento, a geometria da estrada e o volume de tráfego sobre a sinistralidade rodoviária, nos segmentos das duas autoestradas em estudo, aplicando a metodologia dos Modelos Lineares Generalizados Mistos.

Realizar-se-á uma descrição detalhada do trabalho que foi realizado na organização dos dados, na descrição das variáveis, na análise exploratória, na estimação dos modelos, e por fim, na seleção do modelo que melhor se ajusta aos dados.

4.1 Base de Dados

A base de dados em estudo, disponibilizada pela Concessionária Ascendi, refere-se a acidentes rodoviários em sublanços de duas autoestrada da região norte de Portugal, especificamente a Autoestrada de Trás-os-Montes e Alto Douro (A4) e a Circular Regional Exterior do Porto (A41), concessionados pela mesma. Esta base de dados é composta por variáveis como o estado do pavimento, a geometria da estrada, o volume de tráfego, entre outras, abrangendo o período de 2014 a 2021.

Neste estudo, a extensão analisada totaliza 21,44 km (treze sublanços) correspondente a 8,58 km da A4 (seis sublanços) e 12,86 km da A41 (sete sublanços).

Na base de dados inicial com um total de 97 132 observações, cada observação representa os dados de um segmento de estrada de 10 metros de extensão de uma determinada via, sentido e ano.

Na análise preliminar dos dados, constatou-se que, em muitos dos segmentos de estrada, a frequência de acidentes rodoviários era zero. Uma vez que a presença de um elevado número de zeros, pode originar problemas de sobredispersão, optou-se por agrupar os dados, de modo que cada observação passasse a representar um segmento de 500 metros de extensão de uma determinada via, sentido e ano. Para implementar essa alteração, foi preciso ajustar todas as variáveis numéricas, originalmente definidas nos segmentos de 10 metros, para os segmentos de 500 metros, calculando-se os valores médios, valores máximos e valores mínimos nos novos segmentos. É importante realçar que, embora o compri-

mento padrão dos segmentos seja geralmente de 500 metros, ele pode ser menor em alguns casos. Por exemplo, no sublanço "Matosinhos-Sendim" com comprimento total de 660 metros, reagruparam-se os dados em dois segmentos: construiu-se um segmento de 500 metros e outro segmento de 160 metros.

Inicialmente a base de dados era constituída por 19 variáveis.

A Tabela 1 apresenta uma descrição das variáveis presentes na base de dados inicial.

Tabela 1: Descrição das variáveis da base de dados inicial

	Variáveis	Descrição
Estado do Pavimento	<i>IRI</i>	Índice Internacional de irregularidade médio (m/km)
	<i>RD</i>	Profundidade das Rodeiras (mm)
	<i>MPD</i>	Profundidade Média de Perfil (mm)
	<i>Atrito</i>	Coefficiente de Atrito (mm)
Características Geométricas em Planta	<i>C1</i>	Indica a característica geométrica em planta: Alinhamento Reto (AR), alinhamento em Clotóide (CL) e em Curva Circular (CC) (m).
Características Geométricas em Perfil	<i>C2</i>	Indica a característica geométrica em perfil: Trainel (T), Curva Convexa (CCX) e em Curva Côncava (CCV).
Tráfego	<i>TMDAlig</i>	Tráfego Médio Diário Anual de veículos ligeiros
	<i>TMDApes</i>	Tráfego Médio Diário Anual de veículos pesados
	<i>TMDAt</i>	Tráfego Médio Diário Anual Total de veículos ligeiros e pesados
Outras variáveis complementares	<i>Sub</i>	Identificação do sublanço
	<i>Objeto</i>	Identificação da autoestrada (A4,A41)
	<i>Ano</i>	Identificação do ano de observação
	<i>Pk_inir</i>	Identificação do valor do quilómetro inicial do segmento de estrada
	<i>Pk_fnir</i>	Identificação do valor do quilómetro final do segmento de estrada
	<i>N_Vias</i>	Identificação do número de vias do segmento de estrada
	<i>Via</i>	Identificação da via do segmento de estrada (VD - Via Direita, VC - Via Central e VE - Via Esquerda)
	<i>Sentido</i>	Identificação do sentido do segmento de estrada (C - crescente e D - decrescente)
	<i>ENT_S</i>	Identificação de entrada ou saída do segmento de estrada (E- entrada, S-saída e O- sem entrada e saída)
	<i>Acd</i>	Número de acidentes do segmento de estrada

Variáveis referentes ao estado do pavimento

O estado do pavimento refere-se às condições da superfície de uma estrada definidos pelos indicadores de conservação. Estas variáveis foram medidas apenas em 2014 e em 2018.

Para cada uma das variáveis analisadas nos segmentos iniciais de 10 metros de extensão, foram construídas novas variáveis agrupadas para segmentos de 500 metros, como já descrito. A seguir, descrevem-se as variáveis originais e as respectivas novas variáveis associadas:

- Índice Internacional de Irregularidade

O Índice Internacional de Irregularidade (*IRI*), quantifica o grau de irregularidade de uma estrada ao longo de um determinado segmento. À medida que o valor do *IRI* aumenta, o pavimento torna-se menos regular e suave, isto é, mais desconfortável.

Construíram-se as seguintes variáveis:

IRI_md: média do índice de internacional irregularidade no segmento de 500 m.

IRI_max: máximo do índice de internacional irregularidade no segmento de 500 m.

- Profundidade das Rodeiras

A Profundidade das Rodeiras (*RD*) é uma medida que caracteriza a profundidade das irregularidades transversais na superfície do pavimento. À medida que o valor da *RD* aumenta, a superfície do pavimento torna-se mais irregular.

Construíram-se as seguintes variáveis:

RD_md: média da profundidade das rodeiras no segmento de 500 m.

RD_max: máximo da profundidade das rodeiras no segmento de 500 m.

- Profundidade Média do Perfil

A Profundidade Média do Perfil (*MPD*) é uma medida que avalia a textura superficial do pavimento, e está diretamente ligado com propriedades de drenagem e atrito.

Construíram-se as seguintes variáveis:

MPD_md: média da profundidade média do perfil no segmento de 500 m.

MPD_min: mínimo da profundidade média do perfil no segmento de 500 m.

- Coeficiente de Atrito

O Coeficiente de Atrito (*Atrito*) é avaliado através da medição da resistência ao deslizamento de um pneu padronizado sobre a superfície do pavimento. Este coeficiente varia de 0 a 1, onde valores mais elevados indicam uma superfície com melhor aderência.

Construíram-se as seguintes variáveis:

ATRITO_md: média do coeficiente de atrito no segmento de 500 m.

ATRITO_min: mínimo do coeficiente de atrito no segmento de 500 m.

Variáveis referentes às características geométricas em planta

O desenho em planta de uma estrada é constituído por curvas circulares de raio constante, curvas de transição com raio variável e alinhamentos retos. As curvas são inseridas no desenho para conectar alinhamentos retos consecutivos e incluem dois tipos principais: curvas circulares e curvas de transição.

Em relação a esta variável, *C1*, categorizada com *AR* - Alinhamento Reto, *CC* - Clotóide e *CC* - Curva Circular na base de dados inicial que indica a característica geométrica em planta, no segmento de estrada de 10 metros, foram criadas três novas variáveis: *AR*, *CC* e *CL*, que representam a extensão em Alinhamento Reto, Clotóide e Curva Circular, respetivamente, no segmento de 500 metros.

- Alinhamentos Retos (em metros)

Trata-se de um segmento reto de uma estrada, sem curvatura, projetado para permitir uma condução sem desvios. Na base de dados inicial estava presente o valor de *AR* na variável *C1*, no segmento de 10 metros de extensão e construiu-se uma nova variável, *AR*, que representa a extensão do alinhamento reto no segmento de 500 m.

- Curva Circular (em metros)

Trata-se de uma curva com raio de curvatura constante ao longo de todo o seu comprimento. Na base de dados inicial estava presente o valor de *CC*, na variável *C1*, no segmento de 10 metros de extensão e construiu-se uma nova variável, *CC*, que representa a extensão da Curva Circular no segmento de estrada de 500 m.

- Clotóide (em metros)

A Clotóide é uma curva de transição com raio gradualmente variável, utilizada para suavizar a passagem entre um segmento reto e um curvo. Na base de dados inicial estava presente o valor de *CL*, na variável *C1*, no segmento de 10 metros de extensão e construiu-se uma nova variável, *CL*, que representa a extensão da Clotóide no segmento de estrada de 500 m.

Variáveis referentes às características geométricas em perfil

A configuração vertical de uma infraestrutura rodoviária é realizada a partir de uma linha contínua ao longo da plataforma, que corresponde ao perfil longitudinal da via. Esse perfil é composto por trainéis, que são segmentos retos, com inclinação positiva ou com inclinação negativa, e por concordâncias verticais, divididas em curvas côncavas (com raio negativo) ou convexas (com raio positivo).

De igual modo em relação a esta variável *C2*, categorizada em *T*- Trainel, *CCV*- Curva Côncava e *CCX* - Curva Convexa na base de dados inicial que indica a característica geométrica em perfil, no segmento de 10 metros, foram criadas três variáveis: *T*, *CCV* e *CCX* respectivamente, no segmento de estrada de 500 metros.

- Trainel (em metros)

Os trainéis são os elementos mais simples do perfil longitudinal, caracterizados por uma inclinação constante. Quando a inclinação é positiva, são chamados de rampa, e, quando negativa, são denominados declive. A variável *T* representa a extensão em trainel no segmento de estrada de 500 m.

- Curva Côncava (em metros)

As curvas côncavas são usualmente designadas por depressões. A variável *CCV* representa a extensão em curva côncava no segmento de estrada de 500 m.

- Curva Convexa (em metros)

As curvas convexas são usualmente designadas por lombas. A variável *CCX* representa a extensão em curva convexa no segmento de estrada de 500 m.

Variáveis Complementares

Relativamente à variável, *ENT_S*, na base de dados inicial que identifica a existência de entradas ou saídas no segmento de estrada numa extensão de 10 m, esta foi categorizada da seguinte forma: E – quando há vias apenas de entrada no segmento de 500 metros, S - quando há vias apenas de saída no segmento de 500 metros, ES – quando há vias de entrada e saída no segmento de 500 metros e 0 – os segmentos sem entradas e saídas.

Quanto à variável inicial, *N_Vias*, que indica o número de vias num segmento de estrada numa extensão de 10 m, foi construída a variável *NVIASMAX*, que regista o número máximo de vias num segmento de estrada de 500 m.

As restantes variáveis são :

- *Ano*: Identifica o ano em que foram observados os acidentes.
- *PK_inir*: Identifica o valor do quilómetro inicial do segmento de estrada.
- *PK_finr*: Identifica o valor do quilómetro final do segmento de estrada.
- *Vias*: Identifica a via do segmento de estrada. Está codificada em VD - via direita, VC - via central, VE - via esquerda.
- *Sentido*: Identifica o sentido do segmento de estrada. Está codificada como C - crescente, D - decrescente.
- *TMDAligMEDIA*: Tráfego médio diário anual de veículos ligeiros.
- *TMDApesMEDIA*: Tráfego médio diário anual de veículos pesados.
- *TMDAtMEDIA*: Tráfego médio diário anual total de veículos ligeiros e pesados.
- *Auto*: Identifica a autoestrada. Variável com as categorias: A4 e A41.
- *Sub*: Identifica o sublanço da estrada.
- *Acd_t*: Número total de acidentes no segmento de estrada de 500 m de extensão.

Após agrupar os dados em segmentos de 500 metros, a base de dados final é constituída por 2188 observações e 27 variáveis.

Destas observações, 876 referem-se à autoestrada A4 e 1312 dizem respeito à autoestrada A41.

Na Tabela 2 encontram-se as variáveis e a sua descrição da base de dados final.

Tabela 2: Descrição das variáveis da base de dados final

		Descrição
Estado do Pavimento	<i>IRI_md</i>	Índice de Rugosidade Internacional da Irregularidade médio (m/km)
	<i>IRI_max</i>	Índice de Rugosidade Internacional da Irregularidade máximo (m/km)
	<i>RD_md</i>	Profundidade das Rodeiras - valor médio (mm)
	<i>RD_max</i>	Profundidade das Rodeiras - valor máximo (mm)
	<i>MPD_md</i>	Profundidade Média de Perfil - valor médio (mm)
	<i>MPD_min</i>	Profundidade Média de Perfil - valor mínimo (mm)
	<i>ATRITO_md</i>	Coeficiente de Atrito - valor médio (mm)
	<i>ATRITO_min</i>	Coeficiente de Atrito - valor mínimo (mm)
Características Geométricas em Planta	<i>AR</i>	Extensão em alinhamento reto (m)
	<i>CL</i>	Extensão em clotóide (m)
	<i>CC</i>	Extensão em curva circular (m)
Características Geométricas em Perfil	<i>T</i>	Extensão em trainel (m)
	<i>CCV</i>	Extensão em curva côncava (m)
	<i>CCX</i>	Extensão em curva convexa (m)
Tráfego	<i>TMDAligMEDIA</i>	Tráfego médio diário anual de veículos ligeiros
	<i>TMDApesMEDIA</i>	Tráfego médio diário anual de veículos pesados
	<i>TMDAtMEDIA</i>	Tráfego médio diário anual total de veículos ligeiros e pesados
Outras variáveis complementares	<i>Sub</i>	Identificação do sublanço
	<i>Ano</i>	Identificação do ano das observações
	<i>Pk_inir</i>	Identificação do valor do quilómetro inicial do segmento
	<i>Pk_fnir</i>	Identificação do valor do quilómetro final do segmento
	<i>NVIASMAX</i>	Identificação do número máximo de vias
	<i>Vias</i>	Identificação da via (VD- Via Direita, VC- Via Central e VE- Via Esquerda)
	<i>Sentido</i>	Identificação do sentido (C- crescente e D- decrescente)
	<i>Auto</i>	Identificação da autoestrada
	<i>ENT_S</i>	Identificação de entrada e saída do segmento de estrada (E-Entrada, S-Saída, ES-Entrada e Saída, O-sem Entrada e Saída)
	<i>Acd_t</i>	Número total de acidentes no segmento de 500 metros

Na Tabela 3 encontram-se identificados os sublanços, a sua extensão e a respetiva autoestrada a que pertencem.

Tabela 3: Identificação dos sublanços, extensão e autoestrada

Sublanços	Extensão (km)	Autoestrada
Matosinhos - Sendim (A28/A4)	0,66	A4
Sendim (A28/A4) - Guifões	0,50	A4
Guifões - Custoias	1,89	A4
Custoias - Via Norte	2,63	A4
Via Norte - Ponte da Pedra	1,06	A4
Ponte da Pedra - Águas Santas	1,84	A4
Freixeiro - Aeroporto	2,73	A41
Aeroporto - Lipor	1,00	A41
Lipor - EN13	1,59	A41
Z.I. Maia - EN14	0,50	A41
EN14 - EN107	1,97	A41
EN107 - Maia	2,94	A41
Maia - Alfena	2,13	A41

4.2 Análise Exploratória dos Dados

Nesta secção, realizar-se-á uma análise exploratória da base de dados. Pretende-se compreender as características e comportamentos das variáveis do estudo, identificar padrões, e verificar as possíveis relações entre elas.

4.2.1 Variável Resposta

Neste estudo, a variável resposta é o número total de acidentes, Acd_t , num segmento, ano, sentido e via (designado por segmento).

Na Figura 3 observa-se que na maioria dos segmentos não ocorrem acidentes. O número de segmentos com muitos acidentes (4 ou mais) é baixo.

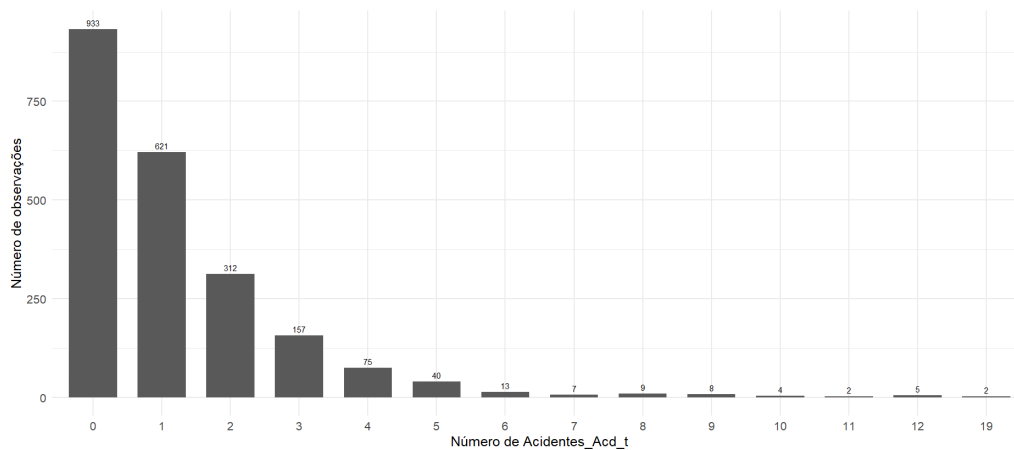


Figura 3: Frequência do número de acidentes por segmento

As estatísticas descritivas do número de acidentes num segmento de estrada, encontram-se registradas na Tabela 4.

Tabela 4: Estatísticas descritivas da variável resposta: *Acd_t*

Mínimo	0
1º quartil	0
Média	1.21
Mediana	1
3º quartil	2
Máximo	19
Desvio padrão	1.71

Em cerca de 50% dos segmentos de estrada, observa-se que o número de acidentes é inferior ou igual a 1 acidente e 25% dos segmentos têm 2 ou mais acidentes. Regista-se que 50% dos segmentos tem uma ocorrência entre 0 e 2 acidentes. O número de acidentes observados por segmento, varia entre 0 e 19.

Na Figura 4 representa-se o número de acidentes por ano usando diagramas em caixa de bigodes. A amplitude interquartil indica que a variabilidade do número de acidentes é semelhante ao longo dos anos com exceção do ano de 2015 e 2020, nas quais se observaram um menor número de acidentes. Em 2020, este resultado era esperado uma vez que ocorreu o confinamento.

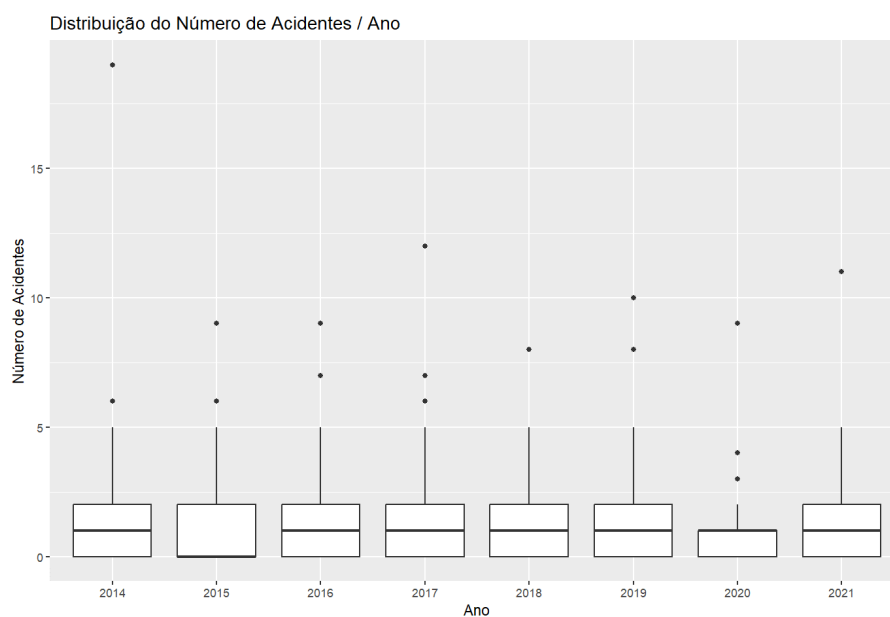


Figura 4: Distribuição do número de acidentes por ano

Existem vários *outliers* em todos os anos, indicando que houve segmentos com um número de acidentes mais elevado do que a maioria dos outros segmentos.

Em 2014, há um *outlier*, indicando um segmento de estrada com um número de acidentes excepcionalmente alto, valor máximo (19 acidentes), que corresponde ao sublanço "Matosinhos-Sendim".

A dispersão do número de acidentes varia consideravelmente entre os diferentes sublanços (Figura 5).

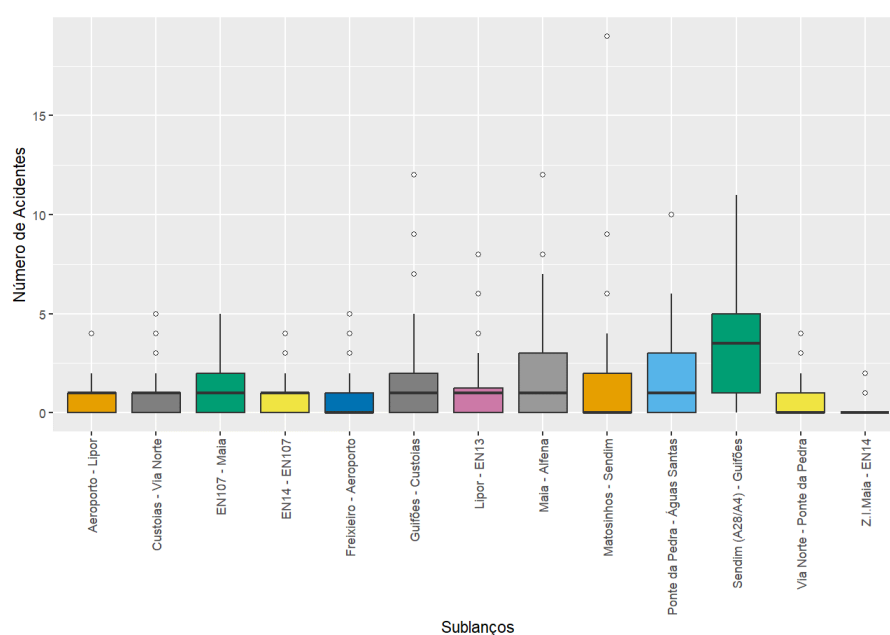


Figura 5: Número de acidentes nos sublanços

O sublanço “Sendim (A28/A4) – Guifões” apresenta uma grande dispersão no número de acidentes entre os segmentos. Muitos sublanços têm observações *outliers*, indicando que há segmentos com um número de acidentes muito diferente dos restantes.

Verifica-se que no sublanço “Z.I.Maia-EN14” o número de acidentes é muito baixo.

O número total de acidentes por ano apresenta uma tendência crescente entre 2015 e 2017, sendo que a partir deste, se verifica uma diminuição até ao ano de 2020, ano com o menor número de acidentes, como seria esperado (Figura 6).

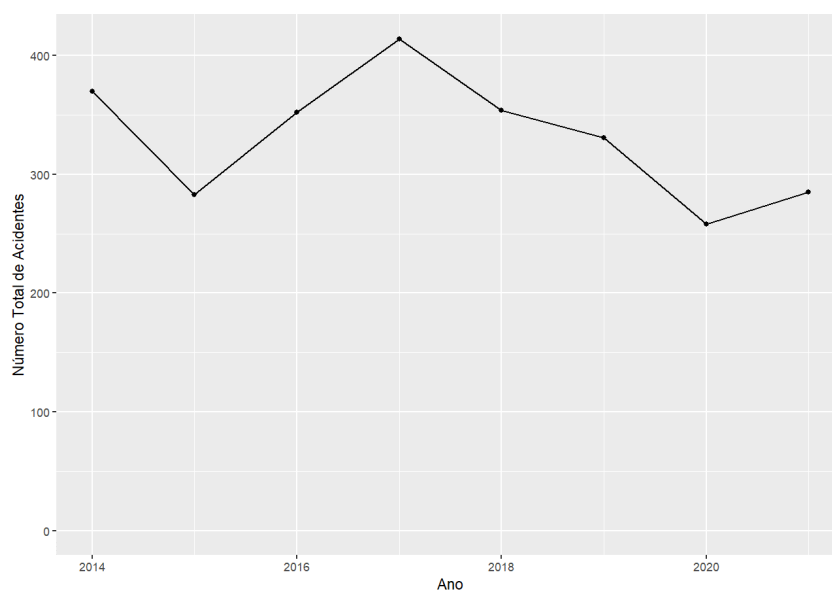


Figura 6: Evolução do número total de acidentes ao longo dos anos

Em termos percentuais, o ano de 2017 foi o ano com o maior número de acidentes, representando quase 16% do número total de acidentes, enquanto o ano de 2020 teve a menor percentagem, correspondente a cerca de 10% do número total de acidentes nos oito anos. (Figura 7).

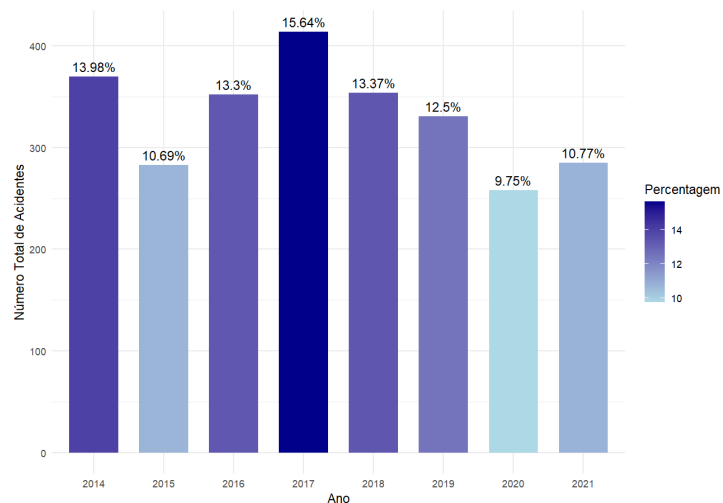


Figura 7: Percentagem do número total de acidentes por ano

A Tabela 5 apresenta o número total de acidentes por ano e sublanço, contabilizando 2647 acidentes.

Tabela 5: Distribuição do número total de acidentes por ano e sublanço nos dois sentidos e em todas as vias

Sublanços	Número total de acidentes	2014	2015	2016	2017	2018	2019	2020	2021
Matosinhos - Sendim	143	55	17	21	16	10	14	6	4
Sendim (A28/A4) - Guifões	187	21	18	32	30	12	24	22	28
Guifões - Custoias	289	35	33	61	57	24	39	21	19
Custoias - Via Norte	223	30	39	9	43	21	21	22	38
Via Norte - Ponte da Pedra	88	3	18	3	12	10	16	8	18
Ponte da Pedra - Águas Santas	293	30	40	21	35	37	45	47	38
Freixeiro - Aeroporto	167	17	14	33	29	27	6	21	20
Aeroporto - Lipor	72	6	0	9	3	24	12	6	12
Lipor - EN13	228	54	15	27	45	45	15	15	12
Z.I. Maia - EN14	12	0	3	0	0	3	0	0	6
EN14 - EN107	153	15	12	15	30	15	27	21	18
EN107 - Maia	391	78	39	81	42	54	34	30	33
Maia - Alfena	401	26	35	40	72	72	78	39	39
Total	2647	370	283	352	414	354	331	258	285

Verifica-se que no sublanço "Z.I.Maia-EN14" não se registam acidentes em alguns anos, embora a sua extensão seja menor.

O sublanço "Sendim (A28/A4)-Guifões" tem a mesma extensão que o sublanço "Z.I.Maia-EN14", no entanto, o número de acidentes é mais elevado.

Observa-se que o sublanço "Freixeiro-Aeroporto" é o segundo sublanço com maior extensão, contudo o número de acidentes não é muito elevado comparando com os outros sublanços.

Conclui-se que o sublanço com o maior número de acidentes é "Maia - Alfena" (401 acidentes), não correspondendo ao sublanço com maior extensão. O sublanço com maior extensão é "EN107 – Maia" com um número total de acidentes de 391.

Estes dois sublanços são os mais críticos, pois apresentam as maiores percentagens do número de acidentes (Figura 8).

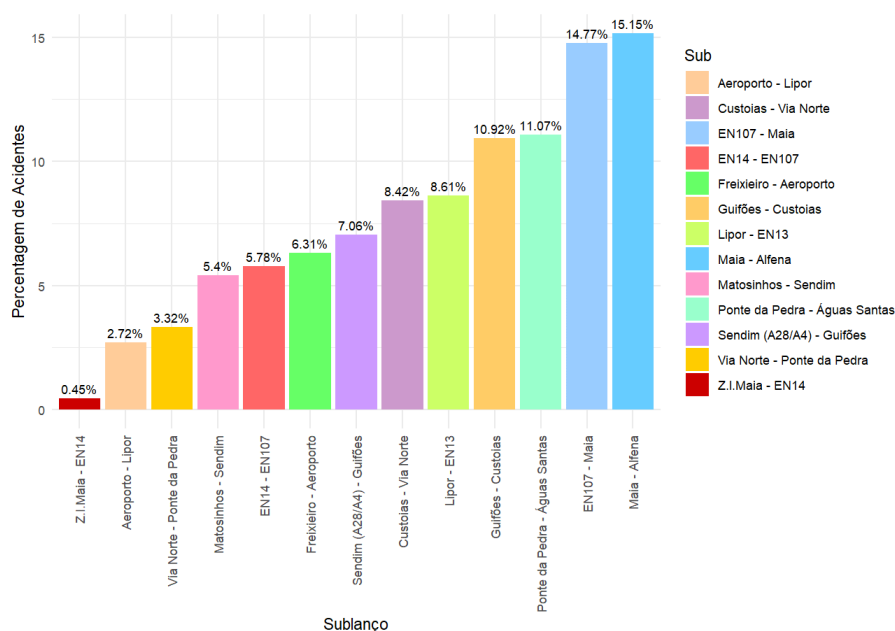


Figura 8: Percentagem do número de acidentes por sublanços em todos os sentidos e vias

A evolução do número de acidentes ao longo dos anos nos diferentes sublanços pode ser observada na Figura 9.

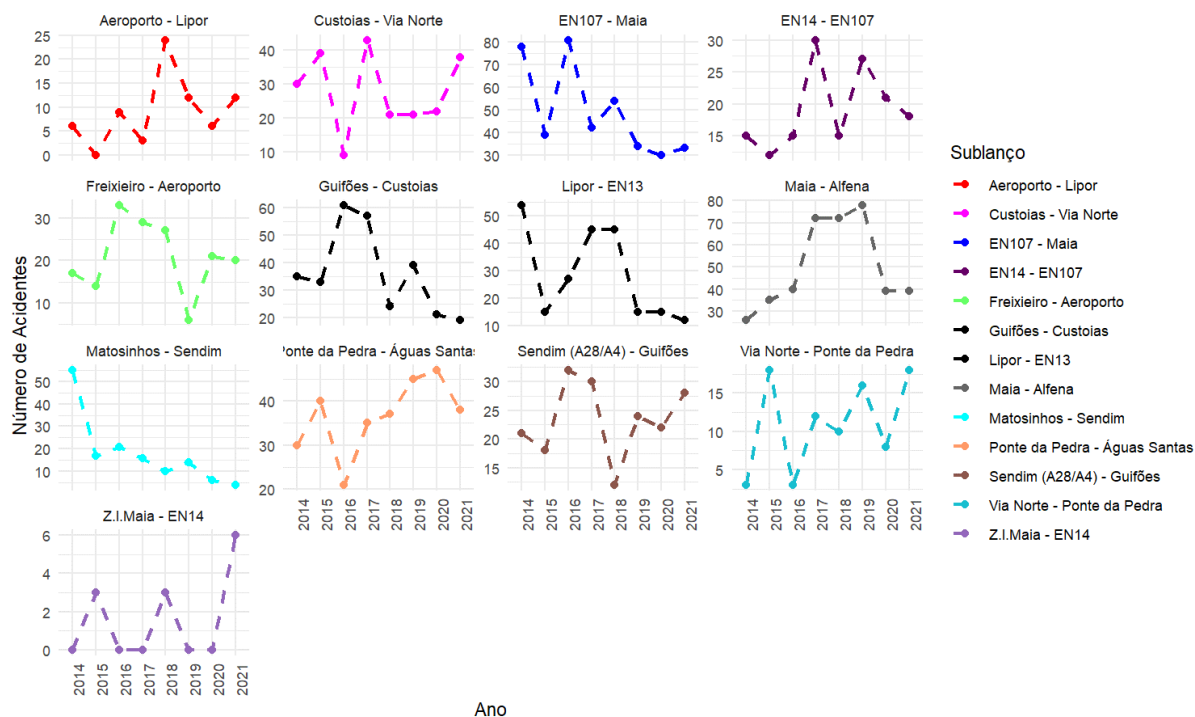


Figura 9: Evolução do número de acidentes por Ano/Sublanço

Observam-se que alguns sublanços apresentam uma tendência mais estável enquanto outros evidenciam variações relevantes, por exemplo, o sublanço "Z.I.Maia - EN14", mostra uma tendência mais estável ao longo dos anos.

Os sublanços "EN107 - Maia", "Guifões - Custóias", e "Maia - Alfena" apresentam o maior número de acidentes em vários anos, no entanto, observam-se valores extremos nos sublanços "EN107 - Maia" e "Maia - Alfena". Estes valores podem ser devidos a possíveis mudanças na estrada ou acontecimentos específicos que aumentaram o número de acidentes nesses anos.

Destacam-se os sublanços "EN107 - Maia" e "Custóias - Via Norte", que apresentam comportamentos até 2019 opostos, enquanto o número de acidentes aumenta no primeiro sublanço, no segundo sublanço o número de acidentes diminui, apesar do aumento constante do volume de tráfego nesses sublanços. Um aspeto a referir é que esses sublanços pertencem a autoestradas diferentes.

Na Figura 10 observa-se que o sublanço "Maia-Alfena" apresenta o maior número de acidentes no sentido crescente e "EN107 - Maia" no sentido decrescente.

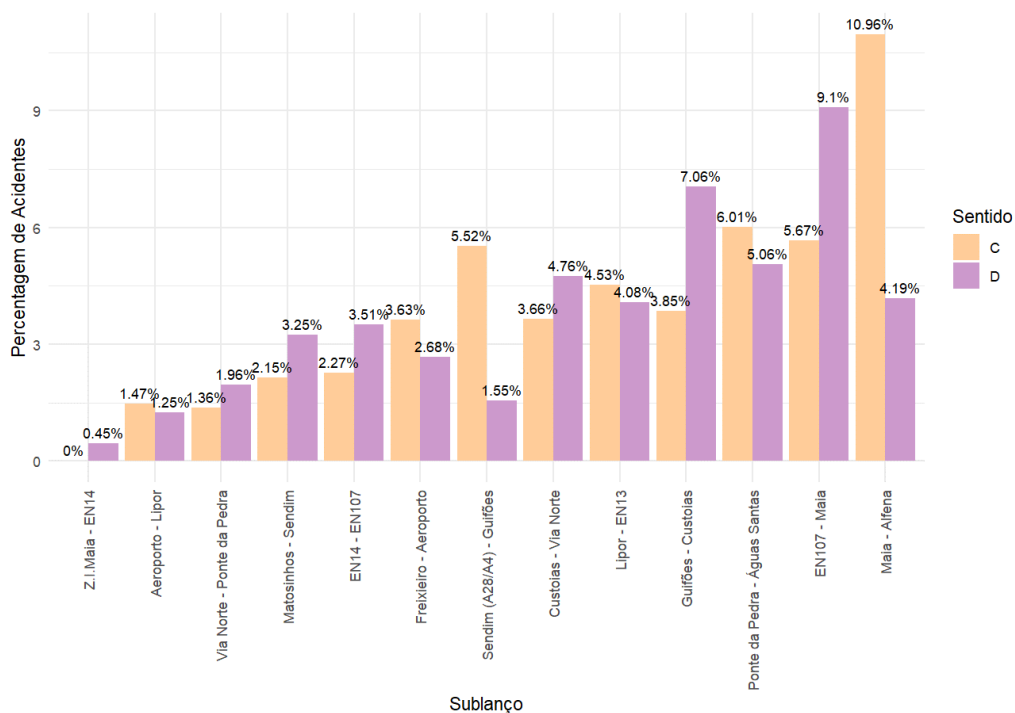


Figura 10: Percentagem do número de acidentes por sublanoço e por sentido

A distribuição do número de acidentes por sublanoços e nos dois sentidos é diferente (Figura 11).

Realça-se o sublanoço “Sendim (A28/A4) – Guifões” que apresenta uma grande dispersão do número de acidentes, no sentido crescente.

O valor mediano mais alto no sentido crescente é indicativo de maior risco de acidente rodoviário, relativamente ao sentido decrescente.

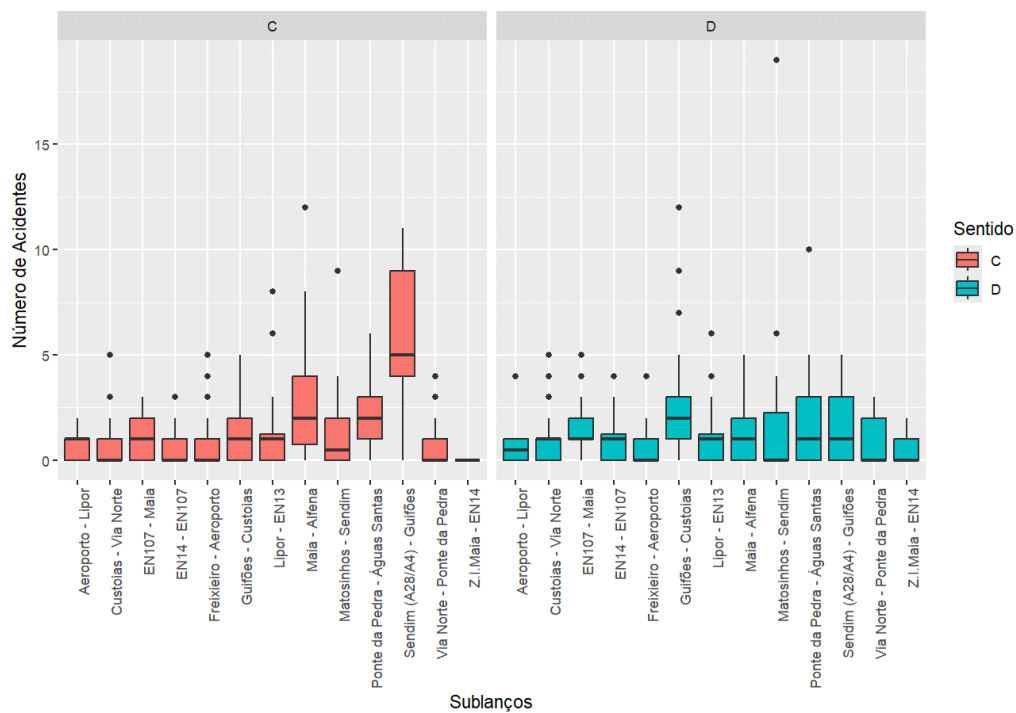


Figura 11: Número de acidentes por sublance e sentido

Quanto ao número de vias, observa-se que o maior número de acidentes ocorre em segmentos com 4 vias (Figura 12).

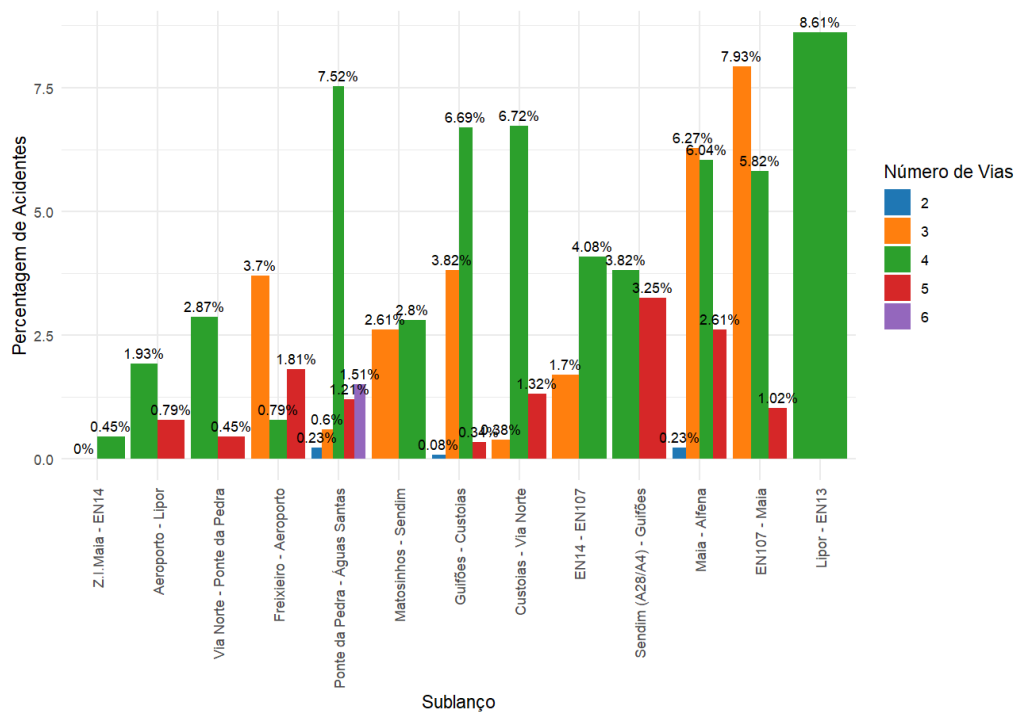


Figura 12: Distribuição percentual do número de acidentes por sublance e número de vias

Em relação ao sentido, observa-se que, em ambos os sentidos, o maior número de acidentes ocorre também em segmentos com 4 vias (Figura 13).

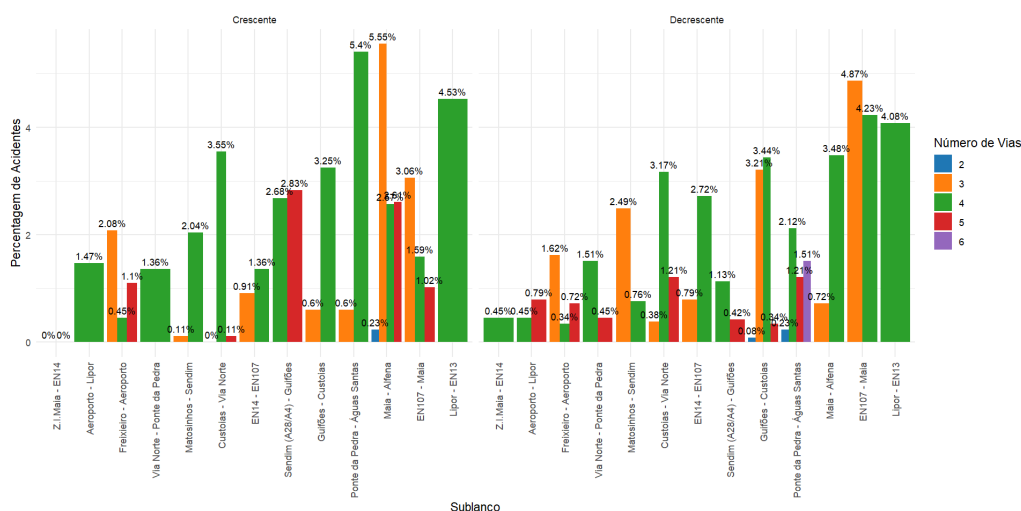


Figura 13: Distribuição do número de acidentes por sublãnço, sentido e número de vias

Relativamente ao número de acidentes relacionados com o tipo de via (Figura 14), constata-se que em muitos sublãnços, as vias do tipo VE e VD têm o mesmo número de acidentes, embora na VE esquerda ocorra, ligeiramente, um maior número de acidentes.

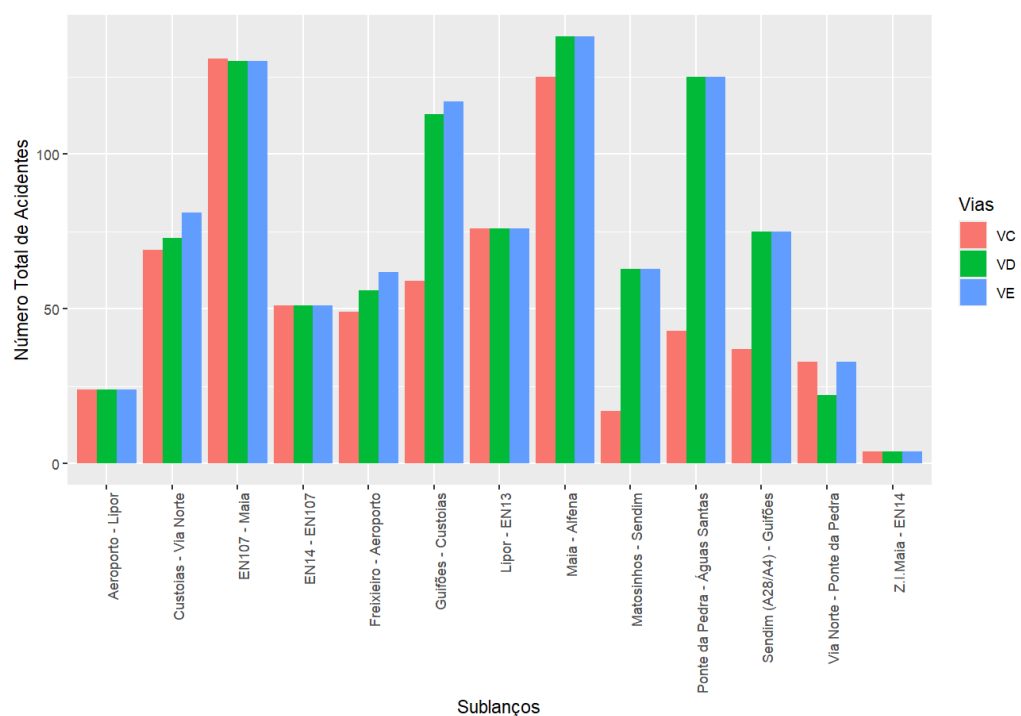


Figura 14: Número de acidentes por sublãnço e tipo de via

Da Tabela 6, conclui-se que os segmentos com entradas e saídas têm um número maior de acidentes em comparação com os outros segmentos.

Tabela 6: Distribuição do número de acidentes pela variável *ENT_S*

ENT_S	Número total de acidentes
O	688
E	631
ES	757
S	571

A distribuição de acidentes varia consideravelmente entre os sublanços e nos diferentes segmentos de estrada com entradas e saídas (Figura 15).

Como já mencionado registaram-se mais acidentes onde existem entradas e saídas de veículos, destacando-se o sublanço "Lipor - EN13".

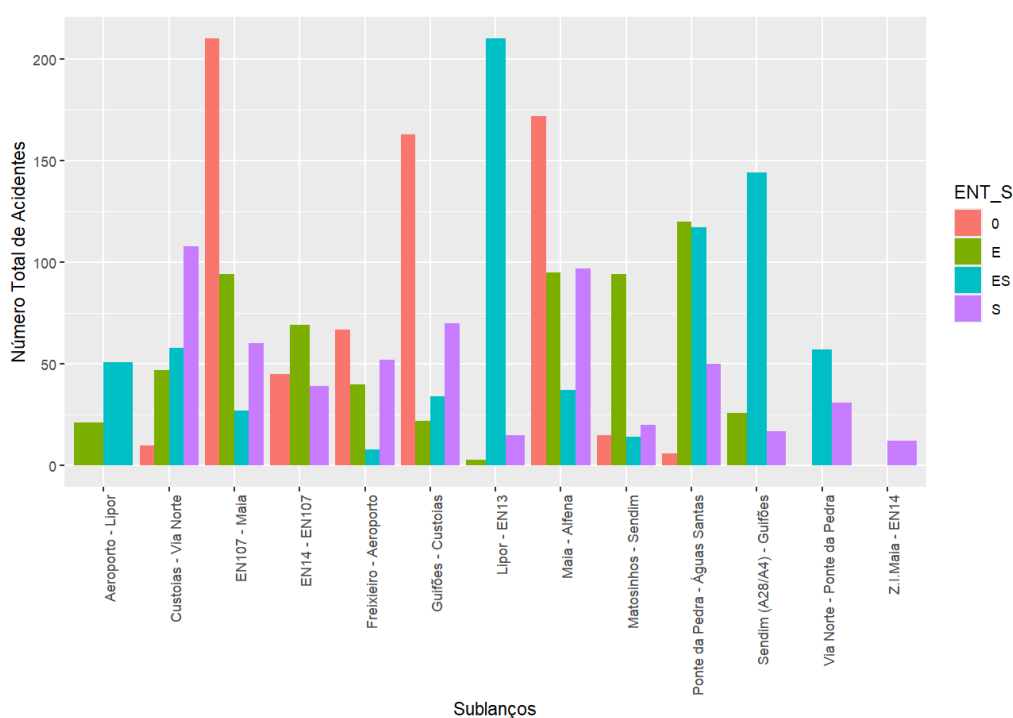


Figura 15: Número de acidentes por sublanço e entradas e saídas de veículos

4.2.2 Análise dos Acidentes nas Autoestradas A4 e A41: Uma Abordagem Comparativa

Este estudo pretende analisar as diferenças no comportamento dos acidentes nas autoestradas A4 e A41.

Realça-se que a extensão da A41 é superior à da A4.

A Tabela 7 apresenta o número de acidentes por ano nas diferentes autoestradas de 2014 a 2021.

Tabela 7: Número de acidentes por autoestrada e por ano

AUTOESTRADA	2014	2015	2016	2017	2018	2019	2020	2021	TOTAL
A4	174	165	147	193	114	159	126	145	1223
A41	196	118	205	221	240	172	132	140	1424
TOTAL	370	283	352	414	354	331	258	285	2647



Figura 16: Número de acidentes por ano e autoestrada

A autoestrada A41 registou um número mais elevado de acidentes em determinados anos atingindo o seu valor máximo em 2018, enquanto que a A4 mostrou-se mais estável, com alguns aumentos e quedas acentuadas.

Um aspeto interessante a destacar é que, no ano de 2018, as autoestradas A4 e A41 apresentaram comportamentos opostos em termos de acidentes. Enquanto a A4 registou o menor número de acidentes em todo o período analisado, a A41 alcançou o seu valor máximo (Figura 16).

Em termos percentuais, a autoestrada A41 regista mais acidentes, representando cerca de 54% do total de acidentes em comparação com a A4, que representa pouco mais de 46%.

Neste estudo analisou-se ainda o número de acidentes nas autoestradas nos dois sentidos (crescente

e decrescente) (Figura 17). Na autoestrada A41 observou-se um número ligeiramente maior de acidentes no sentido crescente em comparação com a A4.

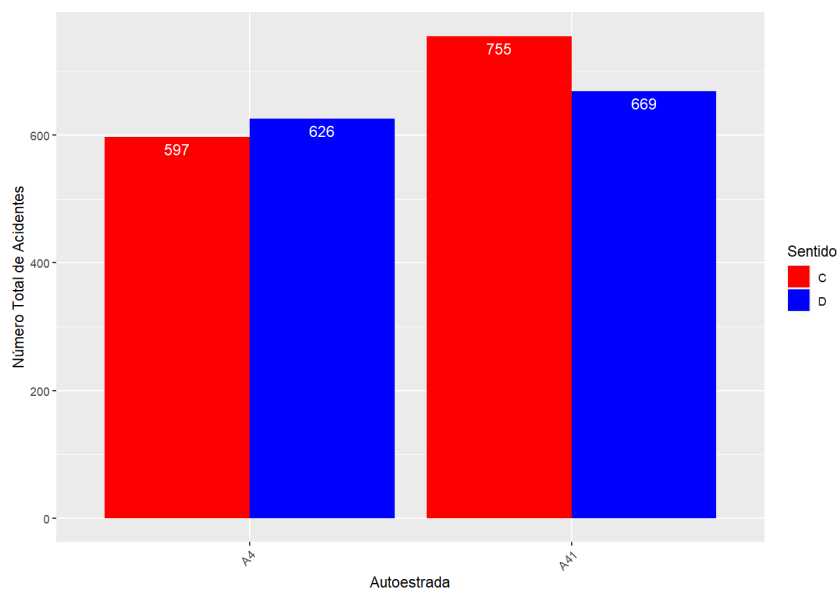


Figura 17: Número de acidentes por autoestrada e sentido

Quanto ao tipo de via, para ambas as autoestradas (A4 e A41), a via esquerda (VE) registou o maior número de acidentes (Figura 18).

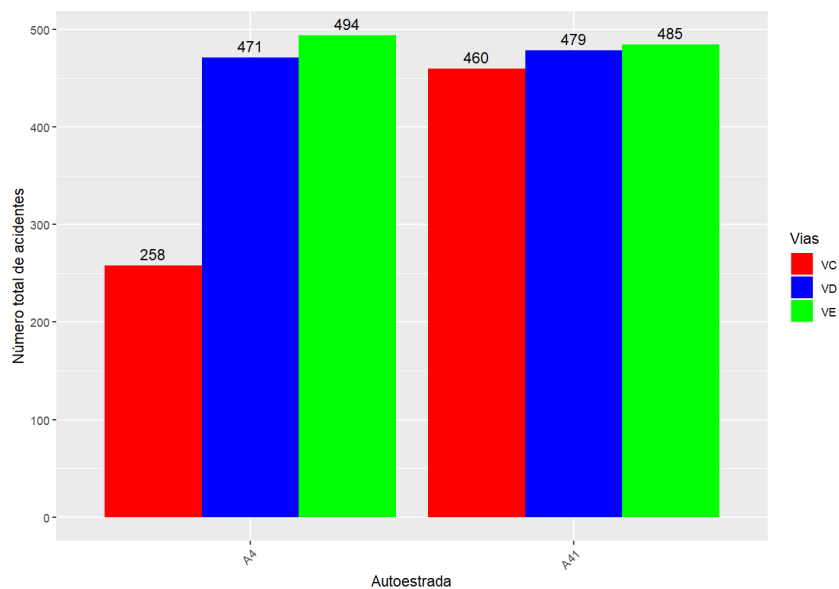


Figura 18: Número de acidentes por autoestrada e via

A autoestrada A4 regista mais acidentes neste tipo de vias no sentido decrescente e a autoestrada A41, no sentido decrescente (Figura 19).

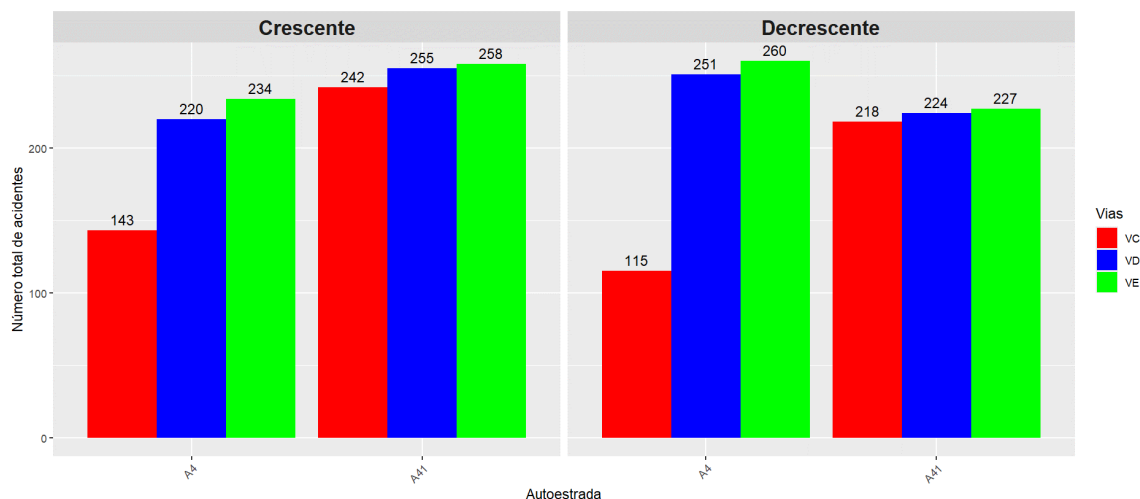


Figura 19: Número de acidentes por autoestrada, sentido e via

Tendo em conta a autoestrada, sentido e ano, (Figura 20), a maioria dos segmentos registam valores de acidentes entre 0 e 5 acidentes, existindo algumas observações *outliers* que excedem esses valores, especialmente na A4.

A A4 apresenta uma variabilidade maior no número de acidentes em comparação com a A41, em ambos os sentidos e por ano.

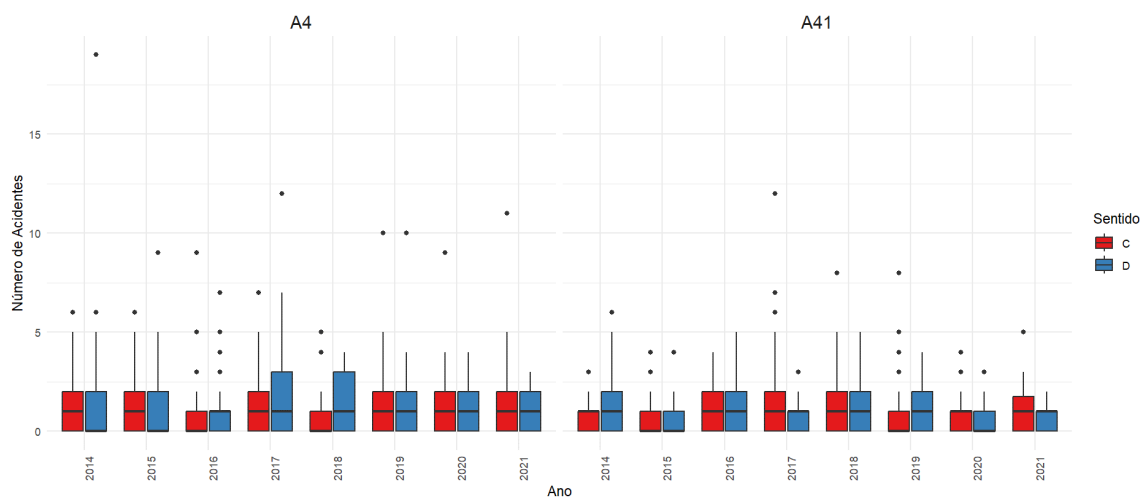


Figura 20: Distribuição do número de acidentes por autoestrada, *sentido* e *Ano*

4.2.3 Análise Exploratória do Número de Acidentes no Sublanço “Maia-Alfena”

De seguida realiza-se uma análise exploratória dos acidentes ocorridos no sublanço “Maia-Alfena”, identificado como o sublanço com maior número de acidentes (não com maior volume de tráfego).

Analisando este sublanço por sentido, observa-se que a maior ocorrência do número de acidentes se verifica no sentido crescente (Figura 21)

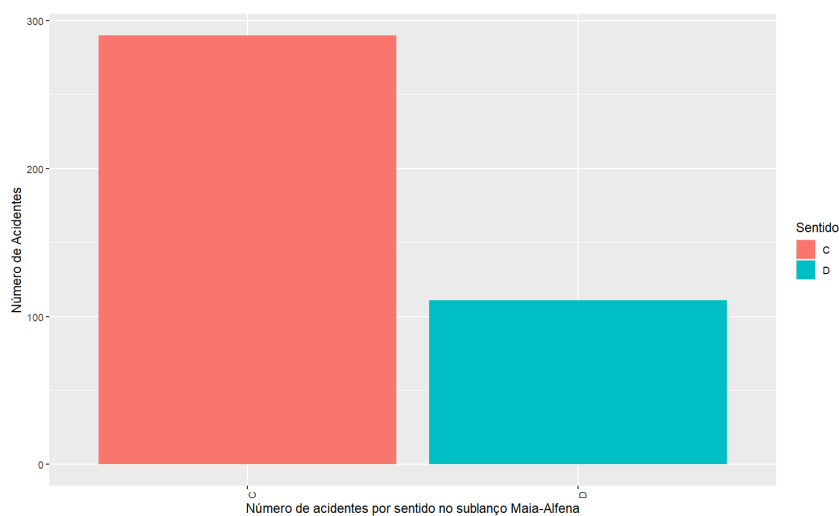


Figura 21: Número total de acidentes por sentido no sublanço Maia-Alfena

A Figura 22 mostra a percentagem do número de acidentes por segmento e sentido no sublanço “Maia-Alfena”. Cada segmento é identificado com o valor de *PK_inir* e *PK_fnir*.

Conclui-se que, em geral, o sentido crescente (C) tem uma maior percentagem de acidentes em comparação com o sentido decrescente (D). Isso pode indicar a necessidade de se procurar para entender as razões desta tendência.

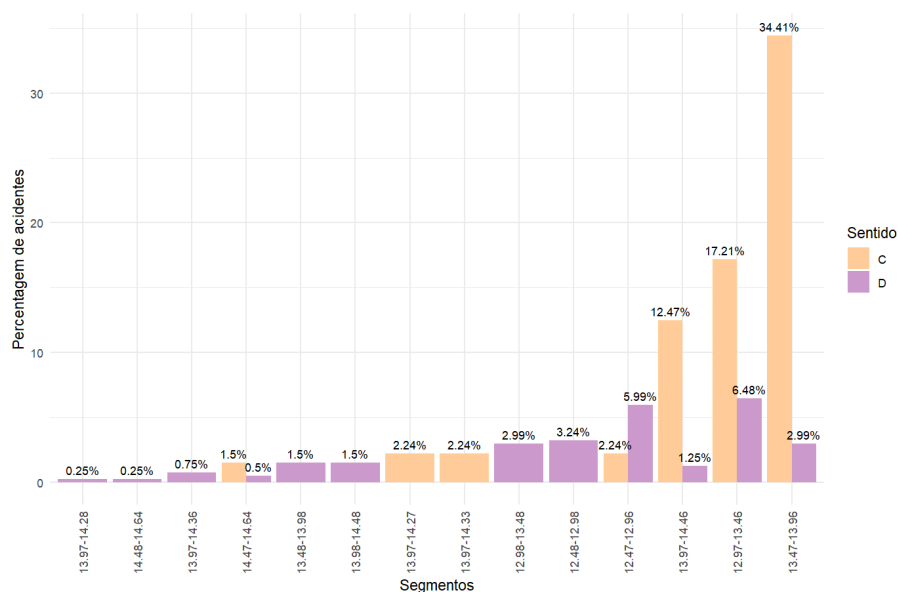


Figura 22: Percentagem de acidentes por segmento de estrada e sentido no sublanço

Observando-se a Figura 22 identifica-se o segmento 13.47-13.96 como o segmento com percentagem mais elevada de acidentes (34.41%). Por esta razão, este segmento é o mais crítico devendo ser uma prioridade para intervenções de segurança.

Quanto ao sentido decrescente, o segmento de estrada 12.97-13.46, deve ser a principal prioridade para intervenções de segurança, dado que apresenta maior percentagem de acidentes.

Realça-se ainda que no segmento de estrada 13.47-13.96, observa-se uma grande discrepância no número de acidentes entre os sentidos crescente e decrescente.

Uma vez que se observa uma grande ocorrência de acidentes no sentido crescente, de seguida analisam-se os dados deste sublanço, neste sentido.

Relativamente ao número de acidentes distribuídos por ano, destaca-se o ano de 2017 com o valor mais elevado de ocorrências (Figura 23)

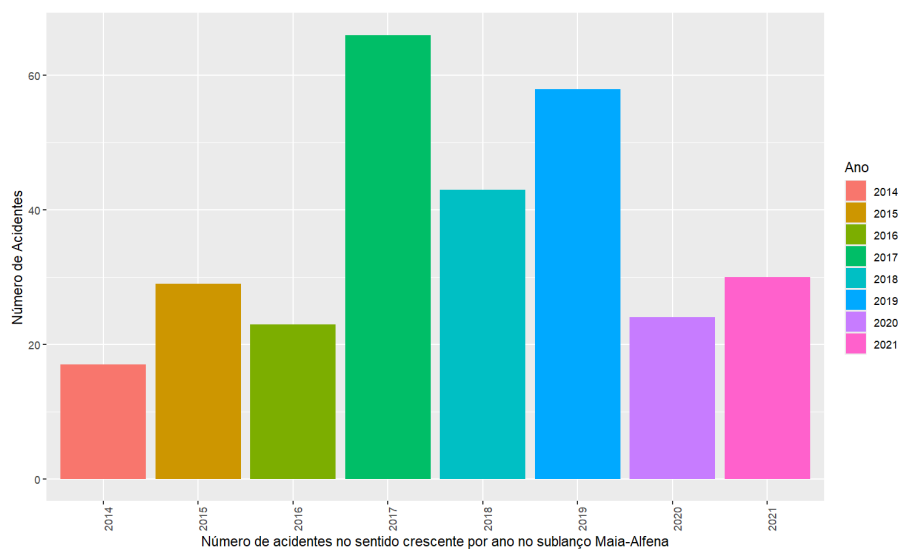


Figura 23: Número de acidentes no sentido crescente por ano no sublanço Maia-Alfena

Relativamente à distribuição do número de acidentes por segmento de estrada no sublanço "Maia-Alfena", o segmento identificado por 13.47-13.96 apresenta o valor mediano maior e destacando-se como o segmento com maior variabilidade de acidentes (Figura 24).

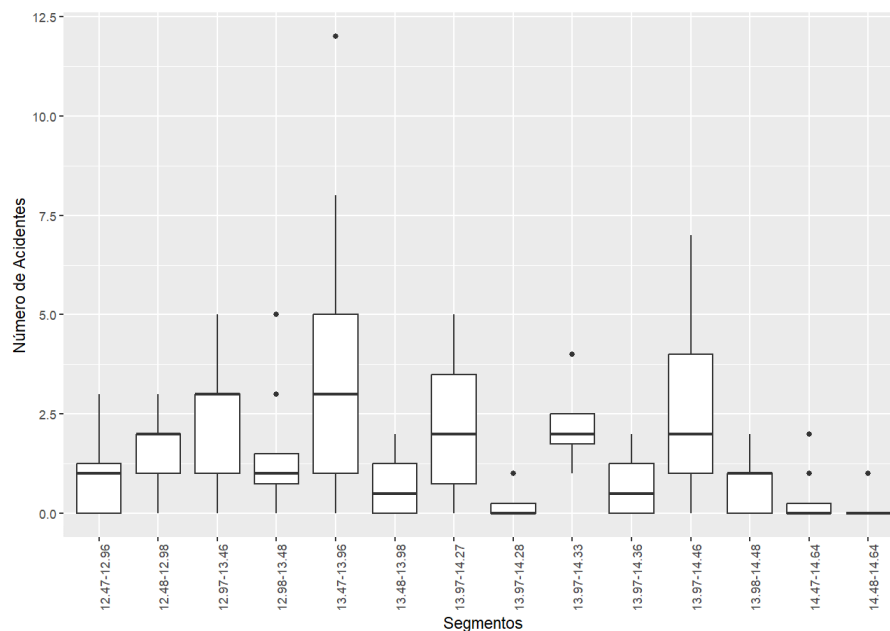


Figura 24: Distribuição do número de acidente por segmentos

Como referido anteriormente, o ano de 2017 foi o ano em que se registou o maior número de acidentes no sentido crescente. Dessa forma, optou-se por analisar o valor médio das variáveis referentes ao estado do pavimento, comparando-as com a média global, para avaliar uma potencial relação, que se

encontram registrados na Tabela 8.

Tabela 8: Média das variáveis referentes ao estado do pavimento

Variáveis	Média no sublanço no sentido crescente	Média global
IRI_md	1.190	1.270
IRI_max	1,500	1,60
RD_md	1,080	1,150
RD_max	2,290	2,370
MPD_md	1,080	1,420
MPD_min	0,218	1,060
ATRITO_md	0,376	0,411
ATRITO_min	0,332	0,348

Observa-se que todas as variáveis têm uma média menor do que a média global. Verifica-se que o pavimento é mais suave, menos irregular, no entanto, a aderência do veículo à superfície piorou.

4.2.4 Estado do Pavimento

Na Tabela 9 estão registradas as estatísticas descritivas das variáveis *IRI* e *IRI_max* nos anos 2014 e 2018 e sentido (crescente e decrescente).

Observa-se que há um aumento no valor médio das variáveis *IRI_md* e *IRI_max* de 2014 para 2018, o que indica um aumento da irregularidade na superfície do pavimento, tornando-o menos suave e mais desconfortável. Verifica-se ainda maior irregularidade no sentido decrescente.

Tabela 9: Estatísticas descritivas das variáveis IRI_md e IRI_max por ano e sentido

ANO	SENTIDO	MÉDIA	MEDIANA	DESVIO PADRÃO	MÍNIMO	MÁXIMO
IRI_md (m km)						
2014	Crescente	1,25	1,21	0,246	0,882	2,41
	Decrescente	1,28	1,20	0,324	0,817	2,69
2018	Crescente	1,31	1,26	0,261	0,886	2,35
	Decrescente	1,35	1,29	0,319	0,857	2,60
IRI_max (m km)						
2014	Crescente	1,60	1,51	0,443	1,010	3,47
	Decrescente	1,61	1,54	0,541	0,940	4,28
2018	Crescente	1,68	1,62	0,418	1,000	3,50
	Decrescente	1,73	1,66	0,490	0,925	3,06

Estes resultados sugerem a necessidade de intervenções de manutenção na estrada, especialmente no sentido decrescente, para melhorar a regularidade da superfície embora os valores observados no período em estudo estejam ainda abaixo do máximo referido no plano de controlo de qualidade (4,5 m|km).

Relativamente à distribuição das variáveis *IRI_md* e *IRI_max* nos anos 2014 e 2018 (Figura 25), e considerando os sentidos (crescente e decrescente), observa-se que, em 2014, o sentido decrescente apresenta maior variabilidade dos dados em comparação com o sentido crescente. O mesmo se observa em 2018.

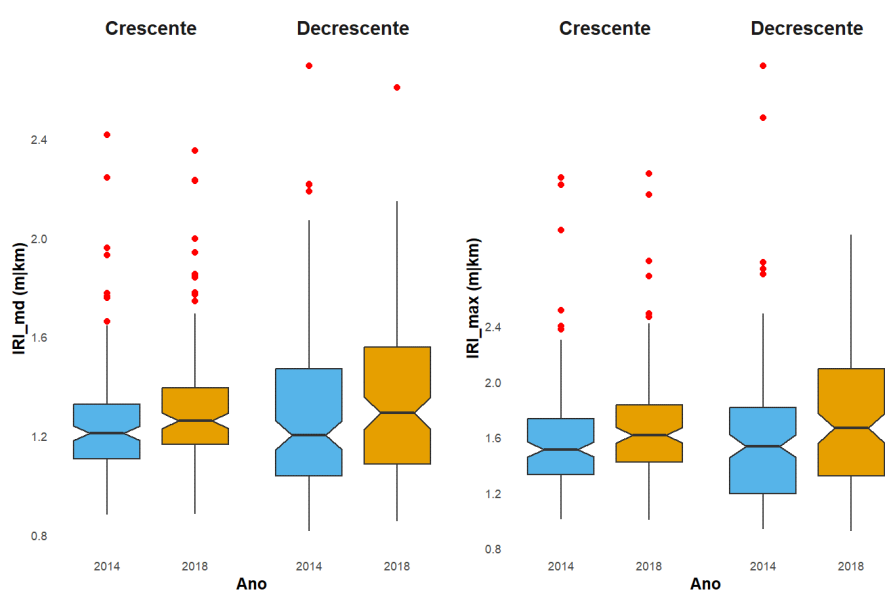


Figura 25: Distribuição das variáveis *IRI_md* e *IRI_max* por ano e sentido

A Tabela 10 fornece as estatísticas descritivas das variáveis *RD_md* e *RD_max* nos anos 2014 e 2018 e sentido (crescente e decrescente).

Em relação às variáveis *RD_md* e *RD_max*, entre 2014 e 2018, houve um aumento do seu valor médio em ambos os sentidos, o que sugere uma maior irregularidade na superfície do pavimento.

Tabela 10: Estatísticas descritivas das variáveis RD_{md} e RD_{max} por ano e sentido

ANO	SENTIDO	MÉDIA	MEDIANA	DESVIO PADRÃO	MÍNIMO	MÁXIMO
RD_{md} (mm)						
2014	Crescente	1,15	1,04	0,623	0,117	3,13
	Decrescente	1,14	1,06	0,629	0,098	3,28
2018	Crescente	1,23	1,06	0,647	0,097	3,13
	Decrescente	1,25	1,15	0,676	0,103	3,63
RD_{max} (mm)						
2014	Crescente	2,38	2,20	1,04	0,25	6,35
	Decrescente	2,37	2,05	1,27	0,35	10,5
2018	Crescente	2,62	2,45	1,19	0,50	6,35
	Decrescente	2,71	2,45	1,26	0,60	7,75

Relativamente à distribuição das variáveis RD_{md} e RD_{max} observa-se que não existem grandes diferenças na variabilidade dos dados entre os anos e sentidos com valores medianos próximos (Figura 26).

Observa-se que na distribuição da variável RD_{max} existem mais observações *outliers* no ano de 2018, no sentido crescente, indicando que existem segmentos de estrada com o pavimento mais irregular.

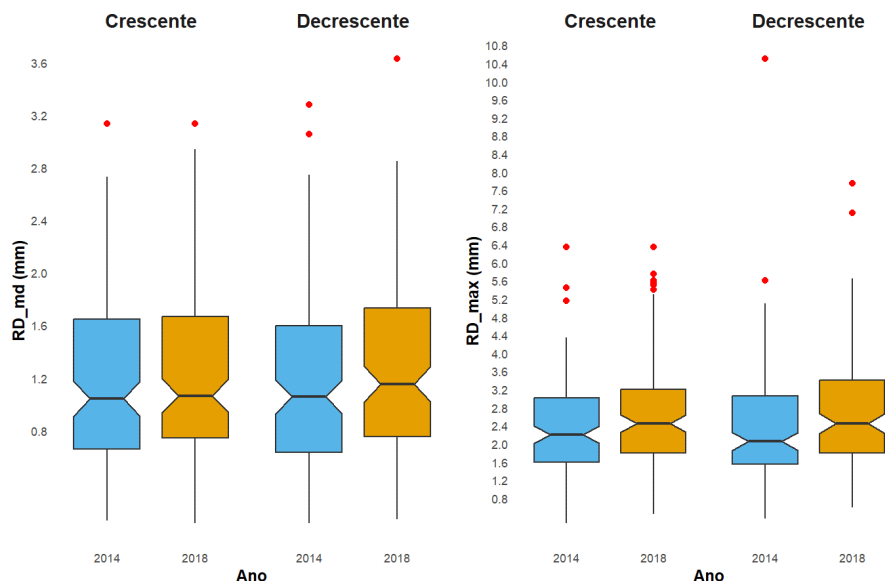


Figura 26: Distribuição das variáveis RD_{md} e RD_{max} por ano e sentido

A Tabela 11 fornece as estatísticas descritivas das variáveis MPD_{md} e MPD_{min} dos anos 2014 e 2018 por sentido (crescente e decrescente).

Observa-se um aumento nos valores médios e valores mínimos de MPD de 2014 para 2018, o que sugere um agravamento natural no estado do pavimento ao longo dos anos. Em 2014, os valores médios

nos dois sentidos são muito semelhantes. O desvio padrão da variável *MPD_md* diminuiu de 2014 para 2018, sugerindo uma menor variabilidade nos dados mais recentes.

Tabela 11: Estatísticas descritivas das variáveis *MPD_md* e *MPD_min* por ano e sentido

ANO	SENTIDO	MÉDIA	MEDIANA	DESVIO PADRÃO	MÍNIMO	MÁXIMO
MPD_md (mm)						
2014	Crescente	1,43	1,47	0,277	0,335	2,29
	Decrescente	1,42	1,45	0,282	0,098	2,23
2018	Crescente	1,55	1,54	0,140	1,19	2,20
	Decrescente	1,53	1,53	0,157	1,13	2,16
MPD_min (mm)						
2014	Crescente	1,05	1,16	0,402	0	1,80
	Decrescente	1,07	1,13	0,368	0	1,50
2018	Crescente	1,24	1,29	0,224	0,541	1,80
	Decrescente	1,25	1,27	0,212	0,570	1,87

Na Figura 27, observa-se um ligeiro aumento na mediana das variáveis *MPD_md* e *MPD_min* de 2014 para 2018. A variabilidade de *MPD_min* diminuiu de 2014 para 2018.

A presença de observações *outliers* é frequente em ambos os anos e para ambas as variáveis.

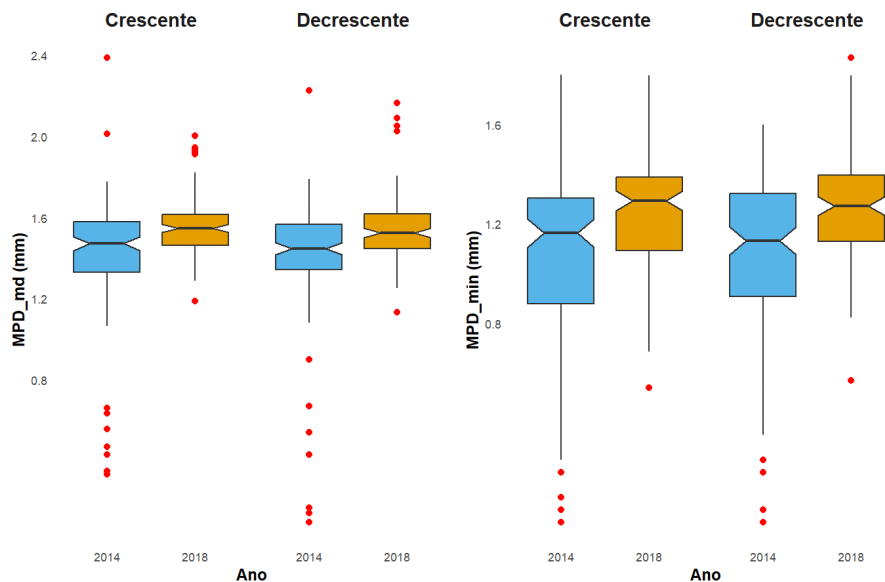


Figura 27: Distribuição das variáveis *MPD_md* e *MPD_min* por ano e sentido

A Tabela 12 fornece as estatísticas descritivas das variáveis *ATRITO_md* e *ATRITO_min* nos anos 2014 e 2018 e sentido (crescente e decrescente).

Verifica-se um aumento nos valores médios e nos valores mínimos das variáveis *ATRITO_md* e *ATRITO_min* de 2014 para 2018, o que indicia uma melhor aderência.

No sentido crescente, a superfície apresenta melhor atrito, uma vez que os valores medianos são maiores em comparação com o sentido oposto.

O desvio padrão diminuiu em 2018, sugerindo uma maior consistência nas medidas de atrito nesse ano em comparação com 2014.

Tabela 12: Estatísticas descritivas das variáveis *ATRITO_md* e *ATRITO_min* por ano e sentido

ANO	SENTIDO	MÉDIA	MEDIANA	DESVIO PADRÃO	MÍNIMO	MÁXIMO
ATRITO_md						
2014	Crescente	0,420	0,417	0,0619	0,301	0,640
	Decrescente	0,401	0,392	0,0567	0,290	0,615
2018	Crescente	0,496	0,497	0,048	0,378	0,628
	Decrescente	0,482	0,484	0,047	0,353	0,590
ATRITO_min						
2014	Crescente	0,354	0,342	0,059	0,215	0,535
	Decrescente	0,341	0,330	0,055	0,225	0,545
2018	Crescente	0,442	0,445	0,051	0,315	0,615
	Decrescente	0,424	0,420	0,048	0,310	0,530

O atrito, tanto o valor médio como o valor mínimo, melhorou de 2014 para 2018 em ambos os sentidos, o que sugere melhores condições de travagem.

Relativamente à variabilidade dos dados (Figura 28), em ambos os sentidos, existe pouca diferença nos anos de 2014 e 2018.

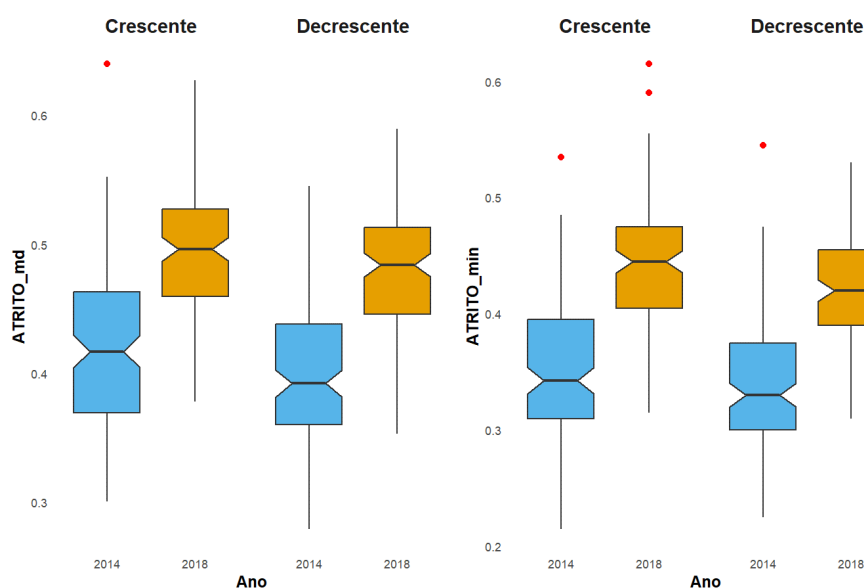


Figura 28: Distribuição das variáveis *ATRITO_md* e *ATRITO_min* por ano e sentido

4.2.5 Características Geométricas de Planta e de Perfil

Na Tabela 13 apresentam-se as estatísticas descritivas das variáveis referentes às características geométricas da estrada.

Tabela 13: Estatísticas descritivas de características geométricas em planta e em perfil

Variáveis	Sentido	Mínimo	Máximo
Características geométricas em planta			
AR	C	0,0	170,0
	D	0,0	170,0
CC	C	0,0	200,0
	D	0,0	200,0
CL	C	0,0	240,0
	D	0,0	240,0
Características geométricas em perfil			
T	C	0,0	180,0
	D	0,0	180,0
CCX	C	0,0	200,0
	D	0,0	200,0
CCV	C	0,0	240,0
	D	0,0	240,0

Relativamente às características geométricas em planta, em ambos os sentidos, observa-se uma extensão máxima de alinhamentos reta de 170 metros, em curva circular de 200 metros e em clotoide, 240 metros. Nas características geométricas em perfil, também em ambos os sentidos, observa-se uma extensão máxima em trainel de 180 metros, em curva convexa de 200 metros e em curva côncava, 240 metros.

4.2.6 Volume de Tráfego

Relativamente ao número de veículos em circulação (Figura 29), observa-se um crescimento contínuo do volume de tráfego ao longo dos anos, à exceção do ano de 2020 correspondente ao período de confinamento imposto pela pandemia da Covid-19.

Entre 2014 a 2015 e entre 2017 a 2019, verifica-se um aumento de veículos em circulação (Figura 29) e uma diminuição do número de acidentes (Figura 6), contrariando a dedução de, maior volume de tráfego, mais interações entre veículos e um aumento do número de acidentes. Pode-se inferir que, apesar do aumento no volume de tráfego, eventuais medidas de segurança ou melhoria na sua gestão parecem ter contribuído para uma redução no número de acidentes, mesmo com o volume de tráfego

em crescimento.

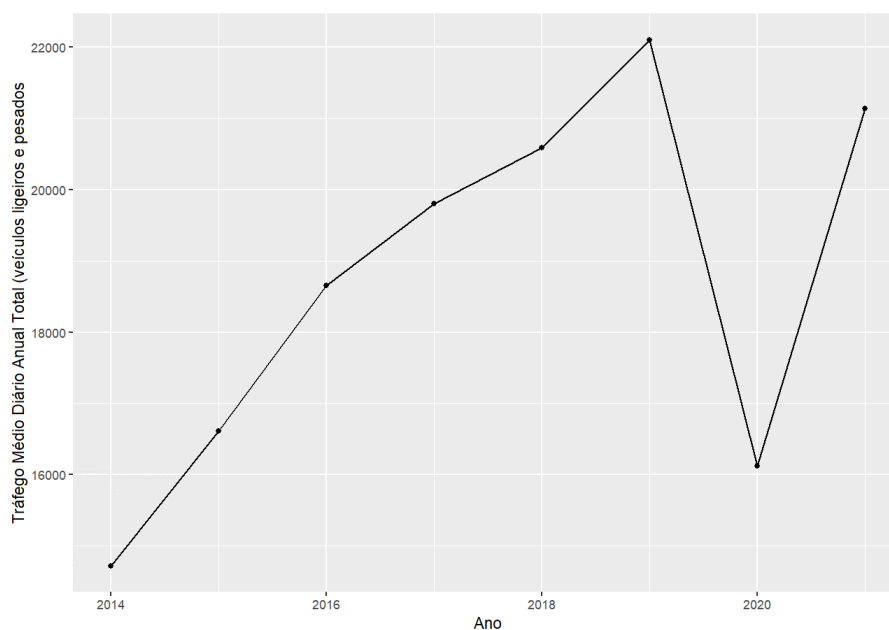


Figura 29: Distribuição do volume de tráfego médio diário anual total

Todos os sublanços apresentam um crescimento semelhante no volume de tráfego médio diário anual de veículos ligeiros ao longo dos anos (Figura 30).

O sublanço “Sendim (A28/A4) - Guifões “ destaca-se como o sublanço com o maior volume de tráfego médio diário anual de veículos ligeiros. Este sublanço mantém um volume de tráfego elevado ao longo dos anos, sugerindo que é uma rota muito relevante.

O volume de tráfego médio diário anual de veículos pesados apresenta um crescimento mais irregular em comparação com os veículos ligeiros (Figura 31).

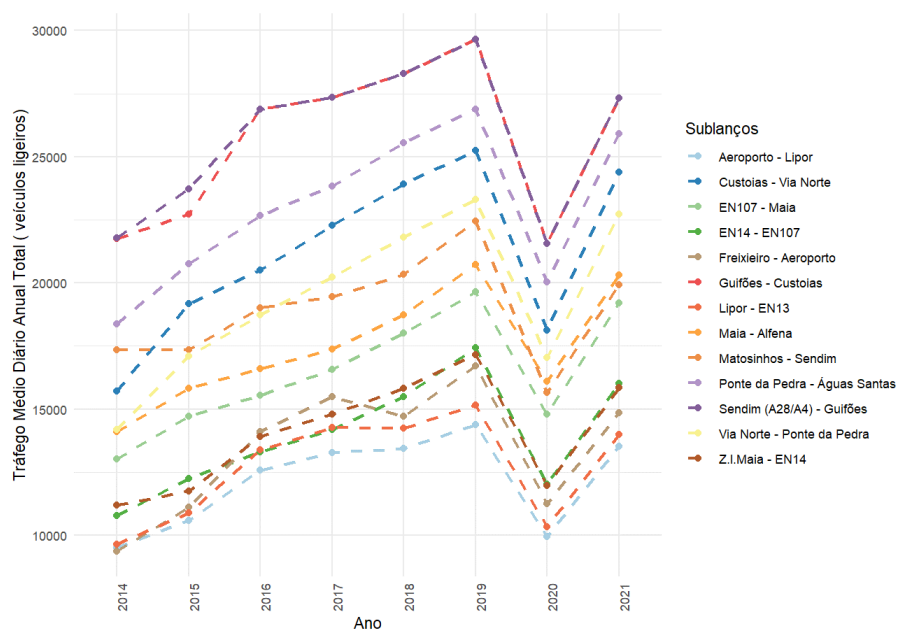


Figura 30: Volume de Tráfego Médio Diário Anual de veículos ligeiros por ano e sublânço

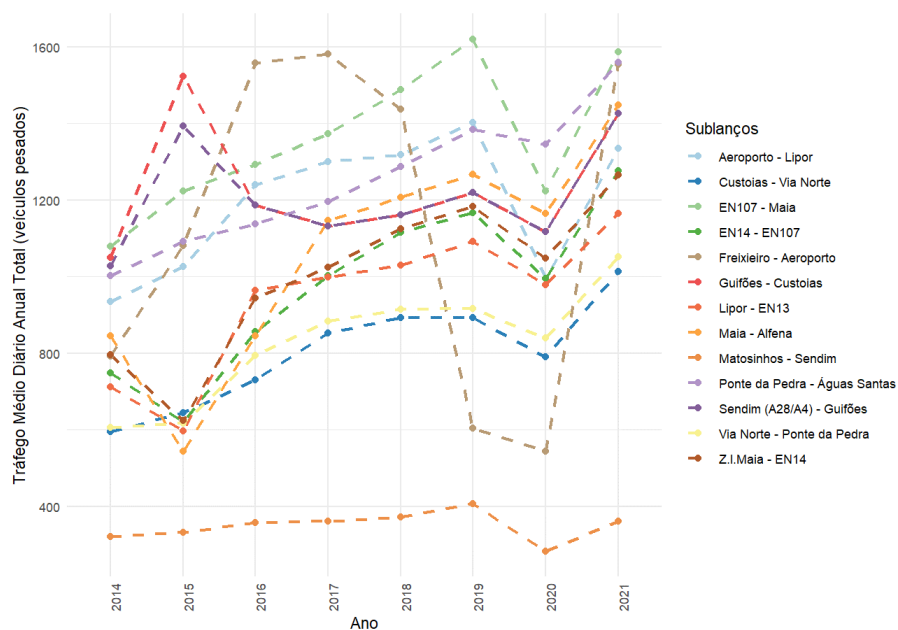


Figura 31: Volume de Tráfego Médio Diário Anual de veículos pesados por ano e sublânço

A Tabela 14 refere-se aos valores do volume de tráfego médio diário anual de veículos ligeiros e pesados, nos treze sublânços, ao longo dos anos.

Tabela 14: Volume de tráfego médio anual de veículos ligeiros e pesados

Sublanços	Tipo de Veículo/Ano	2014	2015	2016	2017	2018	2019	2020	2021
Matosinhos-Sendim	Ligeiros	17353	17345	19020	19458	20338	22254	15692	19918
	Pesados	320	332	358	361	372	406	284	360
Sendim (A28/A4) - Guifões	Ligeiros	21768	23707	26878	27348	28292	29655	21569	27329
	Pesados	1028	1294	1187	1133	1162	1220	1118	1426
Guifões-Custoias	Ligeiros	21745	23702	26878	27348	28292	29655	21569	27329
	Pesados	1051	1523	1187	1133	1162	1220	1118	1426
Custoias-Via Norte	Ligeiros	15704	19168	20513	22288	23904	25249	18126	24377
	Pesados	595	644	730	852	892	893	791	1014
Via Norte-Ponte Pedra	Ligeiros	14186	17092	18729	20236	21798	23295	17046	22723
	Pesados	606	616	793	884	914	917	841	1052
Ponte Pedra-Águas Santas	Ligeiros	18374	20765	22668	23839	25559	26874	20030	25920
	Pesados	1003	1039	1159	1279	1384	1366	1340	1580
Freixeiro - Aeroporto	Ligeiros	9373	11107	14089	15494	14741	16700	11258	14844
	Pesados	793	1082	1558	1581	1437	1605	544	1554
Aeroporto - Lipor	Ligeiros	9502	10577	12583	13264	13444	14365	9935	13511
	Pesados	835	1025	1362	1340	1360	1458	380	1560
Lipor - EN13	Ligeiros	9626	10985	13387	14276	12431	15156	10337	13980
	Pesados	713	597	996	1030	1093	1092	979	1165
Z.I.Maia - EN14	Ligeiros	11196	11777	13489	14798	15828	17154	11962	16588
	Pesados	796	835	1044	1144	1250	1309	847	1266
EN14 - EN107	Ligeiros	10775	12238	14905	15479	17415	17415	12002	16712
	Pesados	749	620	857	1002	1115	1167	996	1197
EN107 - Maia	Ligeiros	14005	14251	16572	18004	19658	17997	11971	17719
	Pesados	1079	1294	1294	1505	1607	1620	1224	1588
Maia - Alfena	Ligeiros	14092	16530	17368	18727	20719	16101	12045	18106
	Pesados	845	544	847	1207	1268	1166	1146	1448

Observa-se que:

- Em todos os sublanços, à execução dos sublanços "EN107-Maia" e "Maia-Alfena" o ano de 2019 registou o maior volume de tráfego médio diário anual de veículos ligeiros. Por outro lado, o ano de 2014, foi o ano com o menor volume de tráfego médio diário anual de veículos pesados em oito dos treze sublanços;
- O ano de 2021 foi o ano de maior volume de tráfego médio diário anual de veículos pesados com execução dos sublanços "Matosinhos - Sendim", "Guifões-Custóias", "Freixeiro-Aeroporto", "Z.I.Maia-EN14" e "EN107-Maia";
- O sublanço "Matosinhos-Sendim", registou o menor volume de tráfego médio diário anual de veículos pesados, e o sublanço "Aeroporto-Lipor" na categoria de veículos ligeiros;

- O sublanço "Sendim-Guifões" foi identificado como o sublanço com maior volume de tráfego médio diário anual de veículos ligeiros;
- No ano de 2017, o volume de tráfego médio diário anual de veículos pesados, atingiu o valor máximo no sublanço "EN107-Maia";
- Nos sublanços "Custóias - Via norte", "Via Norte - Ponte da Pedra", "Ponte da Pedra - Águas Santas" e "EN14 - EN107" houve uma tendência crescente do volume de tráfego médio diário anual de veículos ligeiros e pesados durante os seis primeiros anos (2014 a 2019) e no ano 2020 ocorreu um decréscimo do volume de tráfego médio diário anual;
- É de realçar que no ano 2020, foram aplicadas medidas de confinamento a nível nacional, resultando numa redução do volume de tráfego médio diário anual de veículos ligeiros e pesados em todos os sublanços.

A Tabela 15 mostra as estatísticas descritivas do volume de tráfego médio diário anual de veículos ligeiros (*TMDA_{lig}MEDIA*), de veículos pesados (*TMDA_{pes}MEDIA*), e total (*TMDA_tMEDIA*) por sentido (crescente, decrescente).

Tabela 15: Estatísticas descritivas das variáveis *TMDA_{lig}MEDIA*, *TMDA_{pes}MEDIA*, *TMDA_tMEDIA* por sentido

SENTIDO	MÉDIA	MEDIANA	DESVIO PADRÃO	MÍNIMO	MÁXIMO
TMDA_{lig}MEDIA					
C	17696	16700	4911	9373	29655
D	17583	16603	4822	9373	29655
TMDA_{pes}MEDIA					
C	1052	1082	321	284	1621
D	1048	1080	322	284	1621
TMDA_tMEDIA					
C	18748	17677	4987	101	30875
D	18632	17642	4894	10165	30875

Relativamente ao volume de tráfego médio diário anual de veículos ligeiros (*TMDA_{lig}MEDIA*), os valores médios e da mediana são semelhantes entre os dois sentidos, no entanto, o desvio padrão é ligeiramente menor no sentido decrescente.

Quanto ao volume de tráfego médio diário anual de veículos pesados (*TMDA_{pes}MEDIA*), os valores das estatísticas descritivas são semelhantes entre os dois sentidos, indicando uma distribuição praticamente igual para o volume de tráfego médio diário anual de veículos pesados em ambos os sentidos.

Relativamente ao volume de tráfego médio diário anual total (*TMDAtMEDIA*), a média e a mediana são muito próximas entre os dois sentidos. O desvio padrão é ligeiramente maior no sentido crescente indicando que há uma maior variação do volume de tráfego.

4.3 Modelos de Regressão para Dados de Contagem

Finalizada a análise exploratória, inicia-se a estimação do MLGM, utilizando as técnicas adequadas para tratar dos dados longitudinais com estruturas de dependência.

Na análise exploratória, considerou-se a base de dados referente aos anos de 2014 a 2021. No entanto, esta foi separada em dois conjuntos: um abrangendo o período de 2014 a 2019 para o ajustamento do modelo e outro contendo apenas o ano de 2021 para a validação do modelo.

Inicialmente foi necessário proceder-se a uma reescala das variáveis geométricas (em planta e em perfil), de forma a facilitar a estimação do modelo. Assim, as variáveis finais resultam da divisão por 10 das variáveis iniciais com o mesmo nome.

Na construção de um modelo MLGM, é essencial identificar quais os efeitos aleatórios que devem ser incluídos. O efeito aleatório foi escolhido através do *Critério de Informação de Akaike* (AIC), do *Critério de Informação Bayesiano* (BIC) e do teste de Razão de Verosimilhança.

Foram estudados dois modelos: o primeiro com um efeito aleatório na interseção e o segundo com dois efeitos aleatórios, um na interseção e outro no declive da variável *Ano*.

Na construção dos modelos, a variável referente ao volume de tráfego médio diário anual total de veículos ligeiros e pesados, *TMDAtMEDIA*, foi incluída como um *offset*, para padronizar o número de acidentes em relação ao volume de tráfego.

4.3.1 Modelo Linear Generalizado Misto com Distribuição de Poisson

Considere a variável resposta, Acd_t_{ij} , número total de acidentes no segmento i e ano j , e a variável $TMDAtMEDIA_{ij}$, o volume de tráfego médio diário anual total de veículos ligeiros e pesados no segmento i no ano j , $i = 1, \dots, 281$ e $j = 1, \dots, 8$, tem-se:

$$Acd_t_{ij} \sim \text{Poisson}(\mu_{ij})$$

O modelo nulo com um efeito aleatório na interseção é dado por:

$$\log(\mu_{ij}) = \beta_0 + b_{0i} + \log(TMDAtMEDIA_{ij})$$

sendo

$$\mathbf{b}_i = b_{0i}$$

o efeito aleatório com distribuição normal de valor esperado nulo e matriz de variância-covariância **D**.

O modelo nulo com dois efeitos aleatórios, na interseção e no declive da variável *Ano* é dado por:

$$\log(\mu_{ij}) = \beta_0 + b_{0i} + b_{1i} \times \text{Ano}_{ij} + \log(\text{TMDAtMEDIA}_{ij})$$

com

$$\mathbf{b}_i = (b_{0i}, b_{1i})^T$$

o vetor de efeitos aleatórios com distribuição normal de valor esperado nulo e matriz de variâncias-covariâncias **D**.

Na construção dos modelos introduziu-se o efeito aleatório referente à autoestrada, ao sublanço e ao valor do quilómetro inicial e foram ajustados os vários modelos, conforme apresentado na Tabela 16.

Tabela 16: Modelos nulos com efeitos aleatórios na interseção e no declive da variável *Ano*

	MODELO NULO	Efeito Aleatório	Métricas	Valores
Efeito aleatório na interseção	Modelo I	Auto	AIC	5679.3
			BIC	5690.1
			logLik	-2837.6
	Modelo II	Sub	AIC	5354.4
			BIC	5365.2
			logLik	-2675.2
	Modelo III	PK_inir	AIC	4873.0
			BIC	4883.9
			logLik	-2434.5
Efeito aleatório na interseção e no declive da variável Ano	Modelo IV	Auto	AIC	5661.2
			BIC	5682.9
			logLik	-2826.6
	Modelo V	Sub	AIC	5262.0
			BIC	5283.7
			logLik	-2627.0
	Modelo VI	PK_inir	AIC	4805.9
			BIC	4827.5
			logLik	-2398.9

A comparação dos modelos nulos não encaixados com diferentes efeitos aleatórios, foi realizada usando dos critérios de informação AIC e BIC. Dessa forma, os modelos escolhidos foram os **Modelo III** e **Modelo VI** (menor valor de AIC e BIC).

De seguida compararam-se estes modelos, utilizando-se o teste de Razão de Verossimilhança para avaliar se a inclusão do efeito aleatório no declive da variável *Ano* melhora significativamente o ajuste do

modelo.

Os resultados desta análise encontram-se na Tabela 17.

Tabela 17: Teste de Razão de Verosimilhança entre os Modelos III e VI

Modelos	Valor de prova	Estatística de teste	Graus de liberdade
Modelo III vs Modelo VI	<0.001 ***	71,157	2

Com base no valor de prova, conclui-se que a inclusão do efeito aleatório da variável *Ano* no **Modelo VI** resulta numa melhoria estatisticamente significativa no ajuste do modelo.

Foi assim selecionado o modelo **Modelo VI**, com dois efeitos aleatórios, um na interseção e o outro no declive da variável *Ano*.

Como já foi referido anteriormente, na formação da base de dados foram construídas duas variáveis para cada variável correspondente ao estado de pavimento, *IRI_md* e *IRI_max*, *RD_md* e *RD_max*, *MPD_md* e *MPD_min*, e, *ATRITO_md* e *ATRITO_min*. De seguida, procedeu-se à análise de qual a variável a utilizar no modelo, ajustando-se um Modelo Linear Generalizado Misto Simples com cada uma destas variáveis.

Na Tabela 18 apresentam-se os valores de prova obtidos pelo teste de Razão de Verosimilhança quando se compara cada Modelo Linear Generalizado Simples com o Modelo Nulo.

Tabela 18: MLGM simples das variáveis referentes ao estado de pavimento

Variável	AIC	Valor de prova
IRI_md	4806.1	0.1808
<i>IRI_max</i>	4807.8	0.8114
RD_md	4807.3	0.4379
<i>RD_max</i>	4807.5	0.5618
MPD_md	4805.2	0.1005
<i>MPD_min</i>	4807.9	0.9766
<i>ATRITO_md</i>	4804.2	0.0563
ATRITO_min	4782.9	<0.001 ***

Observando o valor do AIC e o valor de prova obtidos para cada uma das variáveis, optou-se por escolher as seguintes variáveis: *IRI_md*, *RD_md*, *MPD_md* e *ATRITO_min* (modelos com menor AIC e menor valor de prova).

De seguida foi necessário escolher duas variáveis entre três variáveis associadas às características geométricas em planta (AR, CC e CL) e associadas às características geométricas em perfil (T, CCX e

CCV), uma vez que as três variáveis estavam fortemente correlacionadas entre si. Nesta situação existem dois grupos de variáveis que estão relacionadas entre si.

Dessa forma contruíram-se os modelos com duas das três variáveis possíveis.

Na Tabela 19 apresentam-se os valores de prova obtidos pelo teste de Razão de Verossimilhança quando se compara cada Modelo Linear Generalizado Misto com o Modelo Nulo.

Tabela 19: MLGM para as variáveis referentes às características geométricas

Combinação das Variáveis	AIC	Valor de prova
AR+CC	4755.3	<0.001 ***
AR+CL	4766.4	<0.001 ***
CC+CL	4783.1	<0.001 ***
CCV+CCX	4804.2	0.05863
CCV+T	4762.2	<0.001 ***
CCX+T	4766.5	<0.001 ***

Observando o valor do AIC e o valor de prova obtidos para cada uma das variáveis, optou-se por escolher as seguintes variáveis: combinação de *AR* e *CC* e combinação de *CCV* e *T*.

De seguida, ajustou-se o modelo completo que se apresenta na Tabela 20.

Tabela 20: Estimativas dos parâmetros para os efeitos fixos, erro padrão, estatística de teste e valores de prova do modelo completo

Variáveis	Estimativas dos coeficientes (β)	Erro padrão	Estatística de teste	Valor de prova
Constante	-12.4195	0.5568	-22.304	<0.001 ***
SentidoC (Referência)				
SentidoD	-0.0083	0.0535	-0.155	0.8766
ViasVC (Referência)				
ViasVD	0.0298	0.0889	0.335	0.7375
ViasVE	0.1415	0.0683	2.073	0.0382 *
NVIASMAX	0.1867	0.0802	2.329	0.0199 *
ENT_SO (Referência)				
ENT_SE	-0.4280	0.1525	-2.807	0.0050 **
ENT_SES	-0.2957	0.1792	-1.651	0.0988
ENT_SS	-0.4178	0.1445	-2.892	0.0038 **
AutoA4 (Referência)				
AutoA41	1.5577	0.3118	4.995	<0.001 ***
IRI_md	0.1592	0.1497	1.063	0.2877
RD_md	0.0407	0.0712	0.572	0.5675
MPD_md	0.1359	0.1157	1.175	0.2398
ATRITO_min	-1.8826	0.5052	-3.726	0.0001 ***
AR	-0.0406	0.0207	-1.963	0.0496 *
CC	-0.0708	0.0266	-2.665	0.0077 **
CCV	0.0973	0.0234	4.164	<0.001 ***
T	0.2058	0.0428	4.806	<0.001 ***
Ano	-0.0296	0.0278	-1.067	0.2860

Retirando-se sucessivamente as variáveis estatisticamente não significativas, obtêm-se as estimativas dos parâmetros do modelo final, **Modelo VII** (Tabela 21) e as estimativas dos parâmetros dos efeitos aleatórios (Tabela 22).

Tabela 21: Estimativas dos parâmetros para os efeitos fixos, erro padrão, estatística de teste e valores de prova do **Modelo VII**

Variáveis	Estimativas dos coeficientes (β)	Exp(β)	Erro padrão	Estatística de teste	Valor de prova
Constante	-11.6623	0.0001	0.4239	-27.508	<0.001 ***
NVIASMAX	0.1986	1.2197	0.0802	2.477	0.0132 *
AutoA4 (Referência)					
AutoA41	1.0959	2.99175	0.1740	6.297	<0.001 ***
ATRITO_min	-1.5986	0.1782	0.4583	-3.706	0.0005 ***
CC	-0.0251	0.9752	0.0111	-2.266	0.0234 *
CCV	0.0619	1.0638	0.0115	5.403	<0.001 ***
T	0.1349	1.1445	0.0169	7.985	<0.001 ***

Tabela 22: Estimativas dos parâmetros dos efeitos aleatórios do **Modelo VII**

Efeitos aleatórios	Desvio Padrão	Intervalo de Confiança
b_{0i}	0.7708	(0.7304; 0.9069)
b_{1i}	0.1704	(0.1856; 0.1976)
A $\text{Corr}(b_{0i}, b_{1i}) = -0.57$ (-0.7209; -0.4897)		

Assim o Modelo Linear Generalizado Misto (**MLGM VII**) final é dado por :

$$Acd_{t_{ij}}|b_i \sim \text{Poisson}(\mu_{ij}) \quad (4.1)$$

$$\begin{aligned} \log(\mu_{ij}) = & \log(\text{TMDAtMEDIA}_{ij}) + \beta_0 + b_{0i} + \beta_1 \text{NVIASMAX}_{ij} \\ & + \beta_2 \text{AutoA41}_{ij} + \beta_3 \text{ATRITO_min}_{ij} + \beta_4 \text{CC}_{ij} \\ & + \beta_5 \text{CCV}_{ij} + \beta_6 T_{ij} + b_{1i} \text{Ano}_{ij} \end{aligned} \quad (4.2)$$

Mantendo as restantes variáveis constantes, observa-se que:

- Com o aumento do número de vias, o risco de ocorrência do número de acidentes aumenta.
- O risco de ocorrência do número de acidentes na autoestrada A41 é maior do que na autoestrada A4.
- Com o aumento do valor mínimo do atrito (ATRITO_min), o risco de ocorrência do número de acidentes diminui.
- Com o aumento da extensão em curva circular (CC), o risco de ocorrência do número de acidentes diminui.
- No entanto, com o aumento da extensão em curva côncava (CCV) e extensão em trainel (T), o risco de ocorrência do número de acidentes aumenta.

Uma vez que a empresa pretendia utilizar este modelo numa outra estrada com as mesmas variáveis e com as mesmas condições, procedeu-se à construção de um outro modelo, sem a variável *Auto*, para o mesmo conjunto de dados.

Apresenta-se na Tabela 23 as estimativas dos parâmetros para os efeitos fixos, erro padrão, estatística de teste e valores de prova do modelo completo.

Tabela 23: Estimativas dos parâmetros para os efeitos fixos, erro padrão, estatística de teste e valores de prova do modelo completo

Variáveis	Estimativas dos coeficientes (β)	Erro padrão	Estatística de teste	Valor de prova
Constante	-10.4269	0.3932	-26.518	<0.001 ***
SentidoC (Referência)				
SentidoD	-0.0204	0.0533	-0.382	0.7021
ViasVC (Referência)				
ViasVD	0.0993	0.0873	1.137	0.2554
ViasVE	0.1774	0.0679	2.613	0.0089 **
NVIASMAX	0.1969	0.0807	2.440	0.0147 *
ENT_S (Referência)				
ENT_SE	-0.3143	0.1499	-2.098	0.0359 *
ENT_SES	-0.2436	0.1768	-1.378	0.1682
ENT_SS	-0.3382	0.1423	-2.376	0.0175 *
IRI_md	-0.0499	0.1383	-0.361	0.7182
RD_md	0.0367	0.0699	0.525	0.5999
MPD_md	0.0942	0.1149	0.820	0.412
ATRITO_min	-1.8892	0.5065	-3.730	0.0001 ***
AR	0.0398	0.0118	3.372	0.0007 ***
CC	0.0320	0.0151	2.109	0.0349 *
CCV	-0.0023	0.0099	-0.226	0.8212
T	0.0151	0.0151	1.000	0.3171
Ano	-0.0296	0.0279	-1.060	0.2890

Retirando-se sucessivamente as variáveis estatisticamente não significativas, obtêm-se assim as estimativas parâmetros do modelo final, **Modelo VIII** (Tabela 24) e as estimativas dos parâmetros dos efeitos aleatórios (Tabela 25).

Tabela 24: Estimativas dos parâmetros para os efeitos fixos, erro padrão, estatística de teste e valores de prova do **Modelo VIII**

	Estimativas dos coeficientes (β)	Exp(β)	Erro padrão	Estatística de teste	Valor de prova
Variáveis					
Constante	-10.435	0.00002	0.3365	-31.008	<0.001 ***
ViasVC (Referência)					
ViasVD	0.1243	1.1324	0.0641	1.942	0.0522
ViasVE	0.1553	1.1680	0.0622	2.494	0.0127 *
NVIASMAX	0.2065	1.2294	0.0096 **	2.592	0.0132 *
ATRITO_min	-1.9954	0.1359	0.4642	-4.298	<0.001 ***
AR	0.0503	1.0516	0.0082	6.100	<0.001 ***
CC	0.0365	1.0371	0.0084	4.321	<0.001 ***

Tabela 25: Estimativas dos parâmetros dos efeitos aleatórios do Modelo VIII

Efeitos aleatórios	Desvio Padrão	Intervalo de Confiança
b_{0i}	0.7740	(0.7970; 0.8547)
b_{1i}	0.1731	(0.1659; 0.1787)
A $\text{Corr}(b_{0i}, b_{1i}) = -0.59$ (-0.6759; -0.4887)		

Assim o Modelo Linear Generalizado Misto (**MLGM VIII**) final é dado por :

$$Acd_t_{ij}|b_i \sim \text{Poisson}(\mu_{ij}) \quad (4.3)$$

$$\begin{aligned} \log(\mu_{ij}) = & \log(\text{TMDAtMEDIA}_{ij}) + \beta_0 + b_{0i} + \beta_1 \text{ViasVD}_{ij} \\ & + \beta_2 \text{ViasVE}_{ij} + \beta_3 \text{NVIASMAX}_{ij} + \beta_4 \text{ATRITO_min}_{ij} \\ & + \beta_5 \text{AR}_{ij} + \beta_6 \text{CC}_{ij} + b_{1i} \text{Ano}_{ij} \end{aligned} \quad (4.4)$$

Mantendo as restantes variáveis constantes, observa-se que:

- O risco de ocorrência de acidentes na via direita (VD) e na via esquerda (VE) é maior do que na via central (VC).
- Com o aumento do número de vias, o risco de ocorrência do número de acidentes aumenta.
- Com o aumento do valor mínimo do atrito (ATRITO_min), o risco de ocorrência do número de acidentes diminui.
- Com o aumento da extensão em alinhamento reto (AR), e da extensão em curva côncava (CC) o risco de ocorrência do número de acidentes aumenta.

Relativamente aos modelos ajustados, conclui-se que no **Modelo VIII**, as variáveis referentes às características geométricas em planta (alinhamentos retos) e a variável referente ao tipo de via, têm uma influência na frequência do número de acidentes, o que não se verifica no **Modelo VII**. Verifica-se também que a variável T referente à extensão em trainel e a variável CCV referente à extensão em curva côncava não têm qualquer influência na ocorrência dos acidentes, contrariamente ao modelo com a variável Auto . As restantes variáveis referentes ao atrito, número de vias e à extensão em curva circular influenciam a frequência do número de acidentes em qualquer estrada.

4.4 Análise de Diagnóstico do Modelo

Uma vez construído o modelo, procedeu-se à análise do diagnóstico do mesmo. Neste estudo aplicaram-se os métodos de diagnóstico proposto por [28] usando a *package* 'DHARMA' do *software* estatístico R baseado nos resíduos quantílicos. Esta abordagem é baseada em simulações de novos conjuntos de dados baseados no modelo final estimado e nos parâmetros ajustados.

Na Figura 32 está representado o gráfico de quantil-quantil dos resíduos quantílicos, que permite avaliar os pressupostos do modelo.

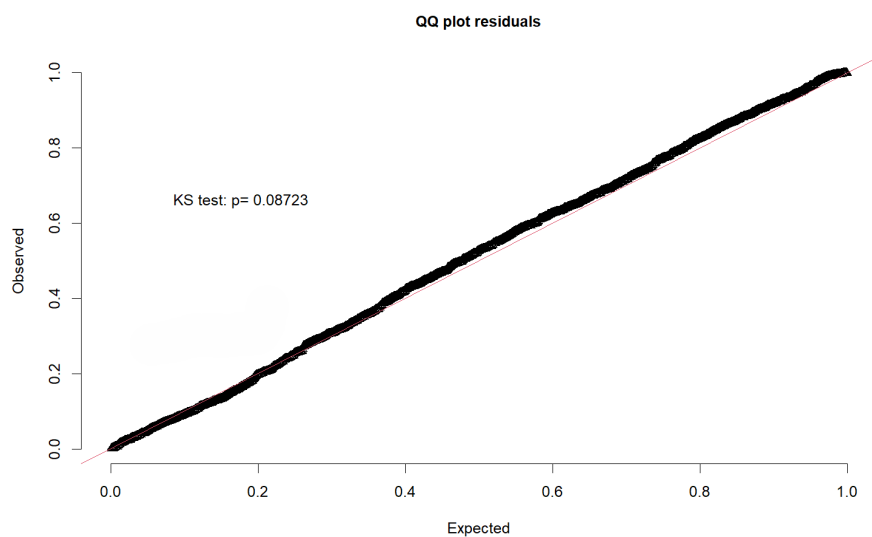


Figura 32: Gráfico QQ dos resíduos Quantílicos

No gráfico da Figura 32 os pontos devem estar todos alinhados com a linha vermelha, sugerindo que os resíduos estão distribuídos de maneira próxima ao esperado pelo modelo (caso contrário, concluíamos que existiria problemas de especificação do modelo).

No caso do modelo em estudo, parece não haver desvios relevantes dos valores esperados.

Na aplicação do Teste de *Kolmogorov-Smirnov*, para avaliar a normalidade dos erros, obteve-se um valor de prova igual a 0.0872, logo há evidência estatística para não rejeitar a hipótese nula de que os erros seguem uma distribuição normal.

Na Figura 33 está representado um gráfico dos resíduos quantílicos versus os valores preditos pelo modelo. As linhas presentes no gráfico são as regressões quantílicas que mostram os quantis 0,25, 0,50 e 0,75. Numa situação ideal estas linhas deveriam ser retas horizontais para cada quantil.

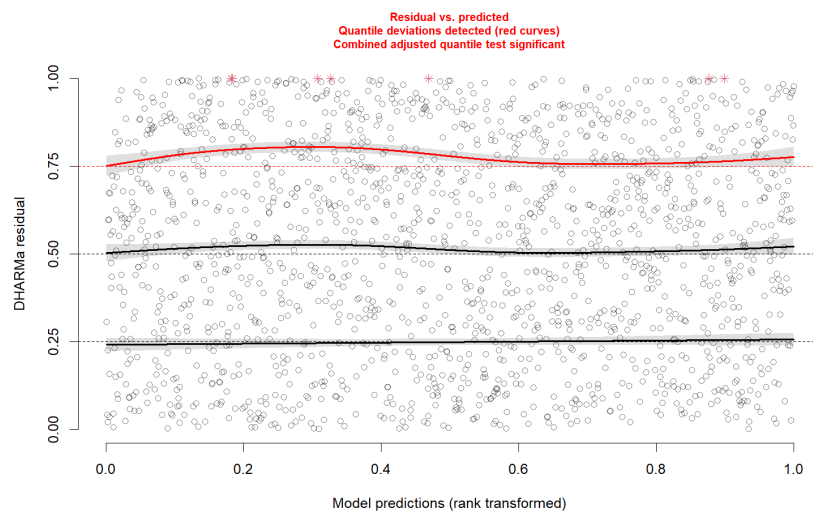


Figura 33: Gráfico dos resíduos quantílicos vs valores preditos

Os resíduos estão distribuídos uniformemente (em torno de 0.5), indicando que não há um padrão de tendência ou desvio dos resíduos em relação às previsões do modelo.

4.5 Validação do Modelo

A validação dos modelos estatísticos é uma etapa importante para avaliar o desempenho do modelo em relação a dados que não foram utilizados na construção do modelo.

Para validar o modelo ajustado com os dados de acidentes rodoviários, utilizou-se o conjunto de dados referente ao ano de 2021 (enquanto que o modelo de ajuste foi construído com os dados de 2014 a 2019).

Nesta análise, pretende-se avaliar a capacidade preditiva do modelo.

O gráfico da Figura 34 compara os valores observados e os valores previstos.

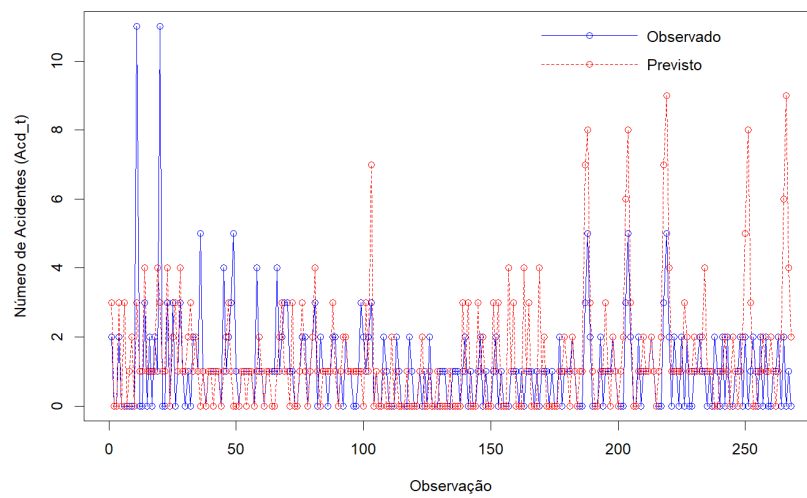


Figura 34: Gráfico dos valores observados e valores previstos

As barras vermelhas representam os valores previstos pelo modelo, enquanto as barras azuis mostram os valores observados no conjunto de validação.

De um modo geral, o modelo apresenta uma razoável capacidade preditiva.

Capítulo 5

Conclusão e Trabalho Futuro

Este estudo centrou-se na aplicação de Modelos Lineares Generalizados Mistos para analisar a sinistralidade rodoviária nas autoestradas portuguesas, utilizando dados longitudinais.

A análise exploratória dos dados indicou que a autoestrada com maior volume de tráfego e número de acidentes rodoviários é a A41, no sentido crescente. O maior número de acidentes foi registado em 2017, embora o maior volume de tráfego tenha ocorrido em 2019. Em 2020, houve uma queda significativa no número de acidentes rodoviários devido à pandemia e ao confinamento.

Observou-se que sublanços como "Maia-Alfena" no sentido crescente e "EN107 - Maia" no sentido decrescente são os mais críticos em termos do número de acidentes. A dispersão do número de acidentes varia consideravelmente entre os sublanços, com alguns *outliers*, sugerindo a existência de segmentos de estrada com um número de acidentes muito diferente dos restantes. Sublanços com maior número de vias (como os de 3 e 4 vias) apresentaram um maior número de acidentes, especialmente na via esquerda. Verificou-se ainda, que a maioria dos acidentes ocorre nos segmentos com entradas e saídas.

Em relação ao estado do pavimento, observou-se um aumento da irregularidade da superfície entre 2014 e 2018, embora o atrito do pavimento tenha melhorado ao longo dos anos eventualmente devido ao aumento da macrotextura.

Foi construído um Modelo Linear Generalizado Misto para dados de contagens com dois efeitos aleatórios na intersecção e no declive da variável *Ano*, considerando o efeito aleatório da variável *PK_inir*.

Os resultados obtidos através dos MLGM, aplicando a regressão de *Poisson*, demonstraram que os modelos capturaram bem a heterogeneidade dos dados de acidentes.

Concluiu-se ainda que as variáveis relativas ao número de vias, autoestrada, atrito, características geométricas em planta (*CC*) e características geométricas em perfil (*CCV* e *T*) influenciam o número de acidentes.

Foi construído um outro Modelo Linear Generalizado Misto com as mesmas variáveis e com as mesmas condições, mas sem a variável *Auto*, para o mesmo conjunto de dados.

Concluiu-se que as variáveis relativas ao tipo de via, número de vias, atrito e características geomé-

tricas em planta (*CC* e *AR*) influenciam o número de acidentes.

Para trabalho futuro seria interessante explorar diferentes metodologias de modelação estatística, tais como: a aplicação de outro tipo de modelos estatísticos como o modelo Binomial Negativa, modelo de Poisson Inflacionado de Zeros e modelo Binomial Negativa Inflacionado de Zeros, técnicas de modelação espacial para analisar a distribuição geográfica dos acidentes e identificar as zonas de maior risco, permitindo a implementação de medidas de segurança específicas para essas áreas.

Seria também interessante estudar o efeito das condições climatéricas tais como a temperatura e a precipitação, na ocorrência de acidentes.

Bibliografia

- [1] Autoridade Nacional de Segurança Rodoviária (ANSR). *Relatório de Sinistralidade e Fiscalização Rodoviária*. 2024. URL: <http://www.ansr.pt/Estatisticas/RelatoriosDeSinistralidade/Pages/default.aspx>.
- [2] *Visão Zero 2030*. 2024. URL: <https://visaozero2030.pt/#visao>.
- [3] A. Turkman e G. Silva. *Modelos Lineares Generalizados: Da Teoria à Prática*. Lisboa, Portugal: Edições SPE, 2000.
- [4] G. M. Cordeiro e C.G.B. Demétrio. *Modelos Lineares Generalizados e Extensões*. Universidade de São Paulo, Escola Superior de Agricultura, 2004.
- [5] J. A. Nelder e R. W. M. Wedderburn. «Generalized linear models». Em: *Journal of the Royal Statistical Society. Series A* (1972).
- [6] N. E. Breslow e D. G. Clayton. «Approximate inference in generalized linear mixed models». Em: *Journal of the American Statistical Association* (1993).
- [7] D. Hedeker. «Generalized Linear Mixed Models». Em: *Encyclopedia of Statistics in Behavioral Science*. John Wiley & Sons, Inc., 2005.
- [8] M. S. Cabral e M. H. Gonçalves. *Análise de Dados Longitudinais*. Sociedade Portuguesa de Estatística, 2011.
- [9] D. Lord e F. Mannering. «The statistical analysis of crash-frequency data: A review and assessment of methodological alternatives». Em: *Transportation Research Part A: Policy and Practice* 44.5 (2010).
- [10] A. Cameron e P. Trivedi. *Regression Analysis of Count Data*. 1st ed. New York, NY, USA: Cambridge University Press, 1998.
- [11] A. Zuur, E. Ieno, N. Walker, A. Saveliev e G. Smith. *Mixed Effects Models and Extensions in Ecology with R*. Springer Science+Business Media, 2009.
- [12] F. L. Mannering e C. R. Bhat. «Analytic methods in accident research: Methodological frontier and future directions». Em: *Analytic Methods in Accident Research* (2014). URL: <https://www.sciencedirect.com/science/article/pii/S2213665713000031>.

- [13] A.S.M. Mohaiminul Islam, Mohammadali Shirazi e Dominique Lord. «Grouped Random Parameters Negative Binomial-Lindley for accounting unobserved heterogeneity in crash data with preponderant zero observations». Em: *Analytic Methods in Accident Research* (2023).
- [14] A. Arévalo-Támara, M. Orozco-Fontalvo e V. Cantillo. «Factors Influencing Crash Frequency on Colombian Rural Roads». Em: *Safety and Security in Traffic* (2020). URL: <https://traffic.fpz.hr/index.php/PROMTT/article/view/3385/561561836>.
- [15] R Development Core Team. *R: A language and environment for statistical computing*. R Foundation for Statistical Computing. Vienna, Austria, 2022. URL: <https://www.R-project.org/>.
- [16] D. Bates, M. Mächler, B. Bolker e S. Walker. «Fitting Linear Mixed-Effects Models Using lme4». Em: *Journal of Statistical Software* 67.1 (2015), pp. 1–48.
- [17] G. Broström. *glmmML: Generalized Linear Models with Clustering*. R package version 1.1.5. 2023. URL: <https://CRAN.R-project.org/package=glmmML>.
- [18] W. N. Venables e B. D. Ripley. *MASS: Modern Applied Statistics with S*. R package version 7.3-60. 2002. URL: <https://CRAN.R-project.org/package=MASS>.
- [19] D. Lüdtke, M. S. Ben-Shachar, I. Patil e D. Makowski. *performance: Assessment of Regression Models Performance*. R package version 0.11.2. 2021. URL: <https://CRAN.R-project.org/package=performance>.
- [20] P.J. Diggle. *Analysis of Longitudinal Data*. 2nd ed. Oxford, UK: Oxford University Press, 2002.
- [21] Marcel de Toledo Vieira. «A Consideração da Amostragem Complexa na Análise de Dados Longitudinais». Em: *57ª Reunião Anual da RBras*. Universidade de São Paulo, 2012.
- [22] A. C. Cameron e P. K. Trivedi. *Regression Analysis of Count Data*. 7th. New York: Cambridge University Press, 2008.
- [23] G. M. Fitzmaurice, N. M. Laird e J. H. Ware. *Applied Longitudinal Analysis*. 2nd. Wiley Series in Probability and Statistics. John Wiley & Sons, 2011.
- [24] N. M. Larid G. Fitzmaurice e J. H. Ware. *Applied Longitudinal Analysis*. Boston, Massachusetts, USA: Wiley, 2004.
- [25] H. Goldstein. «Nonlinear Multilevel Models, with an Application to Discrete Response Data». Em: *Biometrika* (1991).
- [26] A. Agresti. *An Introduction to Categorical Data Analysis*. 2nd ed. New Jersey, EUA: Wiley, 2007.

- [27] P. McCullagh e J. A. Nelder. *Generalized Linear Models*. London: Chapman & Hall, 1989.
- [28] F. Harting. *DHARMA: Residual Diagnostics for Hierarchical (Multi-Level/Mixed) Regression Models*. Theoretical Ecology, University of Regensburg. 2022. URL: <https://cran.r-project.org/web/packages/DHARMA/vignettes/DHARMA.html>.

