ESCOLA NAVAL

talante bien faire

Maria Leonor de Sousa Mendes

Controlo de Divulgação Estatística na Marinha

Aplicação e comparação entre dados reais e dados protegidos

Dissertação para obtenção do Grau de Mestre em Ciências Militares Navais, na especialidade de Marinha

Alfeite

2023

Maria Leonor de Sousa Mendes

Controlo de Divulgação Estatística na Marinha Aplicação e comparação entre dados reais e dados protegidos

Dissertação para obtenção do Grau de Mestre em Ciências Militares Navais, na especialidade de Marinha

Orientação de: Professor Ricardo Pinto Moura
Co-orientação de: Professor Victor Sousa Lobo

O Aluno Mestrando,

Assinado por: **MARIA LEONOR DE SOUSA MENDES**
Num. de Identificação: 30285974
Data: 2023.09.25 14:47:13+01'00'

ASPOF M Sousa Mendes

O Orientador,

Assinado por: **RICARDO PINTO MOURA**
Num. de Identificação: BI11597965
Data: 2023.09.28 13.13.34 GMT Daylight time

CHAVE MÓVEL

Professor Ricardo Moura

Alfeite
2023

"I have lived a thousand lives and I've loved a thousand loves. I've walked on distant worlds and seen the end of time. Because I read."

- George R.R. Martin

Agradecimentos

Quem embarca na jornada de realizar ou já realizou uma dissertação de mestrado compreende a magnitude do trabalho, o esforço exigido, o empenho constante e a resiliência necessária para alcançar a sua conclusão. É impossível percorrer esta trajetória sem o valioso contributo de pessoas dedicadas e prestativas, assim, gostaria de expressar o meu sincero agradecimento a todos aqueles que me ajudaram e tornaram possível a conclusão desta dissertação, em especial:

Ao meu orientador, Professor Ricardo Moura pela orientação, apoio e mentoria inestimáveis ao longo de todo o processo de escrita desta dissertação de mestrado. A sua experiência, dedicação e constante incentivo foram fundamentais no rumo da minha pesquisa e no aprimoramento da qualidade deste trabalho.

À minha família pelo seu amor inabalável, compreensão e encorajamento constantes. O vosso apoio tem sido uma fonte de força e motivação durante esta jornada desafiadora. Estou eternamente grata por terem sempre acreditado em mim e pelos sacrifícios feitos para tornar esta empreitada possível.

Aos meus camaradas, que estiveram ao meu lado nesta luta partilhada de escrever uma dissertação, a camaradagem, apoio e esforço coletivo proporcionaram um senso de comunidade e encorajamento mútuo. Enfrentámos os mesmos desafios, celebrámos sucessos e perseverámos juntos em momentos difíceis.

E a todos aqueles que contribuíram, direta ou indiretamente, para a conclusão desta dissertação, sou profundamente grata. O vosso apoio, conselhos e incentivo foi inestimável, e tenho o privilégio de tê-los ao meu lado ao longo desta caminhada académica.

Resumo

No contexto de uma sociedade em constante evolução tecnológica, o reconhecimento da privacidade como um requisito essencial tem se intensificado de maneira significativa. À medida que a tecnologia avança e os dispositivos se tornam cada vez mais interconectados, a proteção dos dados pessoais e a segurança da informação têm assumido uma importância fundamental. A importância da proteção de dados torna-se ainda mais crucial no contexto das informações e operações militares, destacando a segurança nacional, as missões críticas e a proteção dos militares, militarizados e civis da organização.

Este trabalho tem como objetivo o estudo de métodos de Controlo de Divulgação Estatística (CDE) aplicados a bases de dados do foro naval, isto é, uma base de dados que engloba dados sobre embarcações de pesca e fiscalizações realizadas às mesmas por navios da Marinha Portuguesa. Para tal, foram utilizados dados reais, que foram transformados para garantir a sua proteção. Posteriormente, foi analisada a qualidade dos dados protegidos comparando-os com os dados originais. Nestes dados foram usadas as técnicas de adição de ruído, arredondamento, recodificação global, supressão local, imputação parcial e deslocamento geográfico.

Com base nos resultados obtidos, neste trabalho foi possível demonstrar a eficácia das técnicas de proteção escolhidas através da comparação de histogramas de variáveis originais e protegidas, comparação de médias e variâncias, e ainda, através de matrizes de correlação de Pearson, Kendall e Point-Biserial.

A qualidade demonstrada dos dados neste trabalho contribuiu para o avanço do conhecimento sobre o CDE e das estratégias de proteção de dados na Marinha Portuguesa, estimulando investigações futuras no âmbito deste tema.

Palavras-chave: Controlo de Divulgação Estatística, Privacidade, Proteção de Dados, Divulgação, Análise de Padrões

Abstract

In a society constantly evolving technologically, the recognition of privacy as an essential requirement has significantly intensified. As technology advances and devices become increasingly interconnected, the protection of personal data and information security have taken on fundamental importance. The importance of data protection becomes even more crucial in the context of military information and operations, highlighting national security, critical missions and the protection of the organization's military, militarized and civilian personnel.

The objective of this work is the study of Statistical Disclosure Control (SDC) methods applied to naval databases, that is, a database that includes data on fishing vessels and inspections carried out by Portuguese Navy ships. For this, real data was used, which was transformed to guarantee its protection. Subsequently, the quality of the protected data was analyzed by comparing it with the original data. In these data, noise addition, rounding, global recoding, local suppression, partial imputation and geographic displacement techniques were used.

Based on the results obtained, in this work it was possible to demonstrate the effectiveness of the protection techniques chosen through the comparison of histograms of original and protected variables, comparison of means and variances, and also, through correlation matrices of Pearson, Kendall and Point-Biserial.

The demonstrated quality of the data in this work contributed to the advancement of knowledge about the SDC and data protection strategies in the Portuguese Navy, stimulating future investigations within the scope of this topic.

Keywords: Statistical Disclosure Control, Privacy, Data Protection, Disclosure, Pattern Analysis

Índice

Lista de Figuras

Lista de Tabelas

Lista de Equações

Lista de Abreviaturas, Siglas e Acrónimos

AIS	***A**utomatic **I**dentification **S**ystem*
CFR	***C**ommon **F**leet **R**egister*
CDE	**C**ontrolo (de) **D**ivulgação **E**statística
EPD	**E**ncarregado (da) **P**roteção (de) **D**ados
ERS	***E**lectronic **R**eporting **S**ystem*
FI	**F**icheiros (de) **I**nvestigação
FUP	**F**icheiros (de) **U**so **P**úblico
GDH	**G**rupo-**D**ata-**H**ora
IRCS	***I**nternational **R**adio **C**all **S**ign*
LBP	***L**ength **B**etween **P**erpendiculars*
LOA	***L**ength **O**verall*
MMSI	***M**aritime **M**obile **S**ervice **I**dentity*
RGPD	**R**egulamento **G**eral (de) **P**roteção (de) **D**ados
UE	**U**nião **E**uropeia
VMS	***V**essel **M**onitoring **S**ystem*

Introdução

O mundo está em constante adaptação e mudança, o ser humano procura sempre descobrir mais, inovar e melhorar as condições de vida, e apesar de descobrimos algo novo todos os dias, o ser humano não deixa de querer sempre mais. Nos últimos cento e cinquenta anos a tecnologia tem evoluído exponencialmente. Em 1876 foi realizada a primeira chamada telefónica entre Graham Bell e o seu assistente que se encontrava numa sala adjacente (Fitzgerald, 2001) e hoje podemos fazer chamadas para qualquer parte do mundo, partilhando fotos e vídeos.

No entanto, com o avanço da tecnologia e a crescente interconectividade, surge o desafio da partilha segura de dados. À medida que a quantidade e a variedade de dados disponíveis aumentam exponencialmente, é crucial estabelecer mecanismos eficazes para garantir a privacidade e a proteção das informações pessoais. A partilha de dados tornou-se essencial em diversos setores, desde a investigação científica até à colaboração empresarial, mas é necessário encontrar um equilíbrio entre a utilização dos dados para fins legítimos e a salvaguarda dos direitos individuais (Alén-Savikko & Pitkänen, 2016). Neste contexto, é fundamental desenvolver políticas e regulamentações adequadas que promovam a partilha de dados de forma ética e segura, assegurando a confiança dos utilizadores e o respeito pela privacidade[1] e pela confidencialidade[2] das informações.

A nível nacional existe a Comissão Nacional de Proteção de Dados, autoridade responsável pelo controlo e fiscalização do cumprimento do Regulamento Geral de Proteção de Dados (RGPD) da União Europeia (UE), onde se considera um direito fundamental a proteção do tratamento de dados das pessoas singulares (Lei nº58/2019, 2019).

A natureza sensível das informações e operações militares torna a proteção de dados ainda mais crucial no meio em que vivemos, salientando não só a segurança

[1]Privacidade, no contexto da proteção de dados, é definida como o acordo entre a entidade que fornece os dados e a entidade que os recebe, delineando o nível de proteção a ser concedido a esses dados. Considera-se assim, que a privacidade aplica-se aos titulares dos dados e a confidencialidade aplica-se aos dados. Fonte: Hundepool et al. (2010)

[2]A confidencialidade dos dados é uma propriedade dos dados, geralmente resultante de medidas legislativas, que impede a sua divulgação não autorizada. Fonte: Benschop et al. (2022)

nacional, mas também as missões críticas nas quais nos envolvemos em território nacional e internacional, e ainda a proteção dos militares, militarizados e civis da instituição. Como mencionado no artigo 37.°do RGPD, deve ser designado um encarregado da proteção de dados (EPD) sempre que se tratar de um organismo público, como tal, por forma a cumprir com o regulamento a Marinha Portuguesa nomeou o EPD, referido no Despacho do Almirante Chefe do Estado-Maior da Armada, de 23 de dezembro de 2021.

O tema "Controlo de Divulgação Estatística" (CDE), ao longo dos anos tem vindo a evoluir, pois cada vez mais as pessoas se apercebem da sua importância. Este controlo consiste em limitar a capacidade de qualquer pessoa através de dados públicos inferir informações confidenciais sobre indivíduos ou organizações, por meios lógicos ou estatísticos (Mendes, 2010). Considerando uma base de dados pública que contém informações médicas de uma determinada população. Embora essa base de dados não inclua nomes individuais, é possível identificar o registo médico de um indivíduo específico se houver apenas uma pessoa com 70 anos nessa população. Assim, essa associação entre idade e registo médico pode comprometer a privacidade do indivíduo, revelando informações confidenciais sem o seu consentimento.

Contudo, este tema não pretende levar as organizações a reduzirem a sua partilha de dados, mas sim perceber quais são os dados que devem realmente ser protegidos por forma a não pôr em risco a confidencialidade dos mesmos, e com isto, atingir um certo nível de confiança por parte das organizações e dos indivíduos na partilha dos dados. Visto que a evolução tecnológica tem aumentado exponencialmente, observou-se o acesso e divulgação dos dados a tornar-se uma preocupação para os países a nível mundial. E apesar da importância do tema, são poucos os estudos e obras desenvolvidas nesta área, especialmente a nível nacional.

Todavia, existem soluções para este problema, por exemplo, a utilização de métodos de CDE, isto é, métodos que minimizam o risco de divulgação a um nível aceitável divulgando o máximo de informação possível (Hundepool et al., 2010).

Considerando o que antecede, estudar o tema proposto, torna-se de inegável interesse, pelo facto de que estudando e analisando os diversos métodos estatísticos de controlo da divulgação e comparando os dados reais da base de dados FISCREP[3](*Fiscalization Report*) 2015-2023 e os dados protegidos, será possível inferir

[3]O FISCREP é um comunicado operacional, que tem como finalidade relatar o resultado de uma inspeção realizada a embarcações de pesca ou de recreio, tanto nacionais quanto estrangeiras, em águas sob soberania ou jurisdição nacional. Fonte: Comando Naval (2020)

se realmente conseguimos continuar a partilhar dados sem pôr em risco a confidencialidade dos mesmos.

Tendo em conta a pertinência e o âmbito da investigação, torna-se importante estabelecer quais os objetivos propostos que se pretendem alcançar com a realização da mesma, destacando-se os seguintes:

- Compreender e comparar os diversos métodos de CDE;
- Analisar os dados reais pré-processados, por forma a compreender quais são as variáveis que necessitam de ser transformadas;
- Transformar os dados reais em dados protegidos, através de métodos já existentes, para garantir a confidencialidade dos mesmos;
- Estudar as propriedades estatísticas dos dados reais e protegidos, apurando a eficácia das transformações efetuadas;
- Medir o nível de proteção obtida pelos métodos de CDE, mostrando que se pode proteger os dados para além da simples anonimização dos dados, preservando ao máximo a qualidade[4] dos dados.

Posto isto, pretende-se efetuar um estudo sobre os diversos métodos de CDE, transformar os dados sensíveis em dados protegidos e, por fim, comparar as propriedade estatísticas entre a base de dados original e a base de dados protegida.

Em termos de estrutura da dissertação, a mesma será dividida em quatro capítulos, o Capítulo 1 apresenta um enquadramento teórico sobre a importância da confidencialidade, definindo os diversos tipos de ficheiros, tipos de divulgação, as classificações das variáveis, mencionando ainda as etapas essenciais para o processo de divulgação de dados.

Posteriormente, o Capítulo 2, aborda os métodos de CDE existentes, focando-se mais nos métodos que serão posteriormente utilizados na proteção e análise dos dados.

De seguida, o Capítulo 3, apresentada a base de dados em estudo, mencionando a sua obtenção, a descrição detalhada das variáveis presentes na base de dados em estudo, passando pela descrição de todas as ações de processamento inerentes à criação da base de dados protegida, finalizando com uma avaliação de risco.

[4] A qualidade dos dados é medida pela sua capacidade de serem utilizados de forma eficiente, económica e rápida para informar e avaliar a tomada de decisões. Fonte: Karr et al. (2005)

Por último, o Capítulo 4 é dedicado à apresentação e validação dos resultados obtidos após comparação entre a base de dados original e a base de dados protegida, analisando as diversas variáveis transformadas. Por fim, serão apresentadas as conclusões do estudo, respondendo às questões do problema e apresentando perspetivas de futuros estudos nesta temática.

Capítulo 1

Confidencialidade de Dados

A informação estatística é um bem essencial nas sociedades modernas, pois contribui de forma inquestionável para o desenvolvimento económico e social, por exemplo, na tomada de decisão das empresas, na alocação de recursos, na identificação de problemas sociais, entre outros. A proteção da privacidade dos dados é necessária por motivos legais, ligados à salvaguarda da confidencialidade individual, mas também por uma obrigação moral assumida por muitas entidades responsáveis pela recolha de informação estatística, por exemplo, institutos de estatística ou outras entidades responsáveis pela recolha e divulgação de dados estatísticos (Mendes, 2010).

A discussão em torno da privacidade dos dados surge das questões sobre qualidade da informação, isto é, a crescente exigência dos utilizadores de informação leva a um consequente aumento da qualidade dos dados estatísticos e com isto um aumento do risco da divulgação. Ainda, a aquisição de computadores e de programas sofisticados por qualquer indivíduo é cada vez mais fácil, tornando assim a capacidade de identificação maior. Além disso, a partilha de bases de dados abrangentes contendo informações detalhadas sobre a sociedade, englobando não apenas indivíduos, mas também empresas, pode facilitar a identificação de registos individuais por meio da combinação de múltiplas fontes de dados (Mendes, 2010).

Quando se processam e analisam bases de dados é necessário considerar quais os dados que podem ou não ser divulgados. Por um lado, deseja-se fornecer a maior quantidade com a melhor qualidade de informação possível, por outro lado, quanto mais detalhados e precisos os dados forem, maior é o risco de identificação dos indivíduos. Assim, o direito da sociedade à informação entra em conflito com o direito à privacidade dos seus dados. Podemos observar, na Figura 1.1, uma análise gráfica da relação entre o risco de divulgação e a utilidade dos dados, dividido por uma linha reta horizontal que indica o risco máximo tolerável, sendo que num

extremo desta linha encontra-se sempre os dados originais (*Original Data*) onde a confidencialidade foi comprometida, no outro extremo os dados sem utilidade (*No Data*) e, por fim, os dados que são realmente divulgados (*Released Data*) devem encontrar-se o mais à direita no eixo da utilidade de dados. Através desta imagem é demonstrando o problema que se enfrenta quando se pretende divulgar dados o mais consistentes aos originais, mas sem colocar em causa a identificação dos indivíduos.

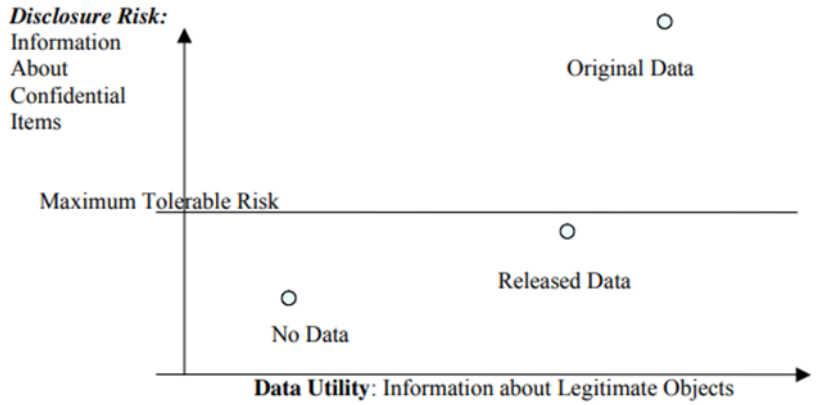

FIGURA 1.1: O problema da limitação da divulgação estatística.
Fonte: Duncan et al. (2001)

Segundo o Regulamento da Comissão Europeia n.°223/2009, considera-se segredo estatístico como a proteção de dados confidenciais de unidades estatísticas individuais obtidos diretamente para fins estatísticos ou indiretamente de fontes administrativas ou outras. Mais ainda, a divulgação é a ação de disponibilizar aos utilizadores as informações e análises estatísticas produzidas. Ainda, definiu dados confidenciais como informações que permitem a identificação direta ou indireta de unidades estatísticas, revelando assim informações pessoais.

Existem várias formas de gerir o risco de divulgação dos dados, entre elas os métodos de CDE que serão mais aprofundados no Capítulo 2. Ao aplicar estes métodos, ocorre uma perda de informação nos dados afetando a perceção que os utilizadores têm sobre os dados, causando assim incerteza. Para alcançar um equilíbrio entre a confidencialidade e a utilidade dos dados, é necessário encontrar o ponto ideal, que maximize a utilidade dos dados e minimize o risco de divulgação (Hundepool et al., 2010), no entanto, é difícil definir esse limite.

Adicionalmente, a compensação entre risco e utilidade no processo de proteção varia com os utilizadores e com as condições em que um arquivo de microdados[5]

[5]Arquivos de Microdados são conjuntos de dados que contêm informações sobre cada indivíduo em relação a várias variáveis, geralmente gerados a partir de pesquisas sociais, de negócios ou fontes administrativas. Fonte: Hundepool et al. (2010).

é disponibilizado. Podemos, assim, identificar três tipos de métodos de divulgação de dados que se aplicam a diferentes grupos-alvo (Dupriez & Boyko, 2010):

- Ficheiros de Uso Público (FUP): são conjuntos de dados gerados a partir de arquivos de microdados que podem ser acedidos por qualquer pessoa que concorde em cumprir um conjunto básico de condições, geralmente disponibilizados online. Os FUP são disseminados sem condições ou com condições mínimas, relacionadas ao uso dos dados, e visam minimizar o risco de identificar informações individuais através da eliminação de todo o conteúdo que possa identificá-los, como nomes, endereços e números de telefone. Contudo, em vez de excluir registos inteiros, existem métodos de CDE alternativos que podem minimizar o risco de divulgação e maximizar a utilidade da informação, explicados posteriormente no Capítulo 2.

- Ficheiros de Investigação (FI): Os FI, não são disponibilizados ao público geral, sendo restritos a utilizadores autorizados após a apresentação de uma solicitação documentada e assinatura de um contrato. Embora os FI sejam anonimizados[6] para minimizar o risco de identificação de indivíduos, eles podem conter dados potencialmente identificáveis se vinculados a outros ficheiros. Identificadores diretos, como nomes, são removidos dos conjuntos de dados, mas variáveis indiretas ainda podem identificar os indivíduos.

- Microdados disponíveis num centro de dados de pesquisa controlada: Os microdados podem ser disponibilizados em centros de dados de pesquisa controlada para assegurar a proteção máxima dos dados considerados sensíveis, por exemplo, conjuntos de dados completos sobre população, pesquisas empresariais e conjuntos de dados relacionados à saúde contendo informações altamente confidenciais. Estas instalações são equipadas com computadores que não estão conectados à internet ou redes externas e proíbem o download de informações por meio de drives USB, CD-DVDs e outros dispositivos. O acesso é restrito a utilizadores que tenham uma necessidade legítima de aceder a um subconjunto específico de dados para cumprir uma finalidade estatística ou de pesquisa declarada. No entanto, a operação de um centro de pesquisa controlada é dispendiosa, pois exige instalações especializadas, equipamentos de informática e pessoal qualificado para gerir o processo de solicitação, revisão de resultados e manutenção dos servidores de arquivos.

Outras formas de acesso aos dados estão disponíveis, como ficheiros para fins educacionais, ficheiros para outras finalidades específicas, acesso remoto ou execução remota. É importante realçar que o nível de proteção necessário varia de acordo com

[6] A Anonimização consiste no uso de técnicas que convertem dados confidenciais em dados anonimizados modificando informações da base de dados. Fonte: Benschop et al. (2022)

o tipo de ficheiro divulgado: um FUP requer uma proteção muito mais rigorosa do que um FI (tipo de ficheiro presente neste trabalho), que por sua vez deve ser mais protegido do que um ficheiro que só está disponível numa instalação local. Posteriormente, será possível especificar em que condições os microdados devem ser fornecidos consoante o tipo de divulgação.

1.1 Tipos de Divulgação

Avaliar o risco de divulgação é uma etapa crítica do processo de CDE. As medidas de risco são utilizadas para determinar se um conjunto de dados pode ser divulgado em segurança, contudo, antes de avaliar o risco de divulgação, é necessário definir que tipo de divulgação é relevante para os dados em questão, segundo Lambert (1993) e Hundepool et al. (2010) podemos classificar a divulgação em três tipos:

- Divulgação de Identidade: ocorre quando um intruso[7] pode associar um registo de dados divulgado a um indivíduo conhecido, associando-o a informações externas. Isso pode ser alcançado utilizando apenas um pequeno subconjunto de variáveis e, uma vez estabelecida a ligação, o intruso obtém acesso a todas as outras informações nos dados divulgados relacionados a esse indivíduo.

- Divulgação de Atributos: ocorre quando um intruso pode inferir novas informações sobre um indivíduo com base nos dados divulgados. Isto acontece quando o intruso identifica com sucesso um indivíduo e aprende novas características sobre ele a partir das informações do conjunto de dados, sendo que pode ocorrer mesmo sem divulgação de identidade. Por exemplo, se um conjunto de dados revela que todos os indivíduos dentro de uma determinada faixa etária têm uma condição médica específica, um intruso pode inferir a condição médica de qualquer indivíduo dentro dessa faixa etária no conjunto de dados sem conhecer a sua identidade específica.

- Divulgação Inferencial: ocorre quando um invasor pode usar os dados divulgados para determinar o valor de uma característica de um indivíduo com mais precisão do que seria possível de outra forma. Por exemplo, usando um modelo de regressão altamente preditivo, um intruso pode inferir informações confidenciais sobre a renda de um indivíduo ao analisar os atributos registados nos dados.

[7]Um intruso é um utilizador de dados que procura associar um indivíduo a um registo de microdados ou atribuir informações sobre unidades populacionais específicas a partir de dados agregados. Fonte: Hundepool et al. (2010)

Segundo Benschop et al. (2022), os métodos de CDE têm como objetivo proteger tanto contra a divulgação de identidade como de atributos, no entanto, geralmente não são métodos considerados para a divulgação inferencial. Isto ocorre, pois, os microdados são normalmente distribuídos aos pesquisadores com o propósito da análise estatística e de inferência. Além disso, essas inferências geralmente visam prever o comportamento agregado em vez de dados individuais e são, portanto, preditores frequentemente imprecisos do comportamento individual.

1.2 Processo de Divulgação de Ficheiros de Dados

Neste subcapítulo será possível adquirirmos uma visão geral do processo de cinco etapas para controle da divulgação dos dados apresentado por Hundepool et al. (2010), começando com um arquivo de microdados original e originando um arquivo para uso externo. Cada etapa aborda aspetos específicos do processo, incluindo a necessidade de proteção de confidencialidade, características chave[8] e usos dos microdados, risco de divulgação, métodos de controle da divulgação e implementação, podemos observar os seus passos na Tabela 1.1.

Este processo envolve fazer escolhas, realizar análises e abordar problemas relacionados com cada etapa. De seguida serão descritas resumidamente as várias etapas deste processo, sendo que alguns destes temas são explicados com mais detalhe ao longo da dissertação:

• Necessidade de proteção da confidencialidade: Esta é a primeira etapa do processo de divulgação de ficheiros que explora a necessidade de proteger os dados. Existem dados que podem ser divulgados sem necessidade de serem protegidos e outros que necessitam dessa proteção, assim microdados que contêm variáveis confidenciais relacionadas a indivíduos ou empresas requerem proteção para garantir a confidencialidade.

• Características e uso dos microdados: Esta etapa envolve o estudo dos principais usos e características dos dados, distinguindo se é para uso público ou para fins de investigação. As diferenças nos tipos de utilizadores implicam diferentes necessidades, cenários de divulgação, tipos de análise que se espera realizar com os dados divulgados, diferentes estatísticas a serem mantidas e o nível de proteção desejado. A análise inclui considerar o tipo e a estrutura dos dados, trabalho preliminar sobre as variáveis, detalhes geográficos e coerência com tabelas publicadas.

[8] As características/variáveis chave são um conjunto de variáveis que, quando combinadas, podem levar à identificação dos indivíduos na base de dados. Fonte: Templ et al. (2014)

Etapas para o processo de divulgação	Análises a efetuar/Problema resolvido ↓ Resultados esperados
Necessidade de proteção da confidencialidade	Os dados referem-se a pessoas singulares ou coletivas? ↓ Precisamos proteger a unidade estatística
Características chave e usos dos microdados	Análise do tipo/estrutura dos dados ↓ Visão clara das unidades que necessitam de proteção
	Análise das metodologias de pesquisa ↓ Tipo de base de amostragem, amostras/enumeração completa dos estratos, análise mais aprofundada da metodologia de pesquisa, calibragem
	Análise dos objetivos dos Institutos de Estatística ↓ Tipo de divulgação (Ficheiros de Uso Público (FUP), Ficheiros de Investigação (FI)), políticas de divulgação, peculiaridades do fenómeno; coerência entre as várias divulgações (FUP, FI), coerência com as tabelas divulgadas e a base de dados online, etc.
	Análise das necessidades dos utilizadores ↓ Variáveis prioritárias, tipos de análise, etc.
	Análise de questionários ↓ Lista das variáveis a ser removidas, variáveis a ser incluídas, ideias sobre o nível de detalhe das variáveis estruturais
Risco de divulgação	Cenário de divulgação ↓ Lista de variáveis identificativas
	Definição de risco
	Avaliação do risco ↓ Se o risco é demasiado alto, são necessários métodos de limitação da divulgação
Métodos de controlo da divulgação	Análise do tipo de dados envolvidos, políticas dos Institutos Nacionais de Estatística e necessidades dos utilizadores ↓ Identificação do método de limitação da divulgação
	Análise da perda de informação
Implementação	Escolha do software, parâmetros e limites dos diferentes métodos

Tabela 1.1: Processo de divulgação de ficheiros de dados. Fonte: Adaptado de Hundepool et al. (2010)

• Risco de divulgação: No processo de divulgação de microdados, é necessário proteger a divulgação das informações, especialmente aquelas relacionadas a indivíduos ou empresas. Os dados podem ser classificados de acordo com sua natureza e utilização, sendo importante considerar as necessidades dos diferentes tipos de usuários, os cenários de divulgação, as análises esperadas e as estatísticas relevantes. É crucial manter a coerência entre diferentes arquivos e evitar a obtenção de mais informações do que o necessário. A avaliação de riscos desempenha um papel fundamental, definindo limites para determinar quando um arquivo ou unidade apresenta um risco aceitável ou inaceitável. A escolha dos cenários e o nível de risco aceitável dependem da cultura, políticas estatísticas e abordagens de análise de cada país. Não há um consenso sobre a melhor metodologia de avaliação de riscos, embora diferentes métodos possam produzir resultados semelhantes.

• Métodos de controle de divulgação: Se a avaliação de risco revelar um alto risco de divulgação, métodos de CDE são aplicados para produzir um arquivo de microdados protegido. Estes métodos têm abordagens perturbativas e não perturbativas, sendo usados para modificar ou reduzir o detalhe do conjunto de dados originais. A escolha do método depende da política do instituto, do tipo de dados e das necessidades do utilizador, sendo que análise da perda de informações deve ser realizada durante o processo de seleção para avaliar o impacto dos métodos de proteção.

• Implementação: A etapa final envolve a implementação de todo o procedimento, incluindo a possível seleção de um software, definição de parâmetros e determinação de riscos aceitáveis.

1.3 Classificação de Variáveis

Para a aplicação dos métodos de CDE, é necessário estabelecer a categorização das variáveis envolvidas, assim, podemos considerar uma variável como uma propriedade, atributo ou característica de um conjunto de dados, que pode ter valores distintos para cada indivíduo. De seguida, é possível observar uma figura representativa da classificação das variáveis.

As **variáveis identificativas** contêm informações que podem levar à reidentificação dos indivíduos, permitindo que um invasor associe os dados a um indivíduo ou grupo de indivíduos. Estas variáveis geralmente são consideradas confidenciais e requerem proteção especial para evitar a divulgação de identidade. Em contrapartida, **variáveis não identificativas** são variáveis que não podem ser usadas

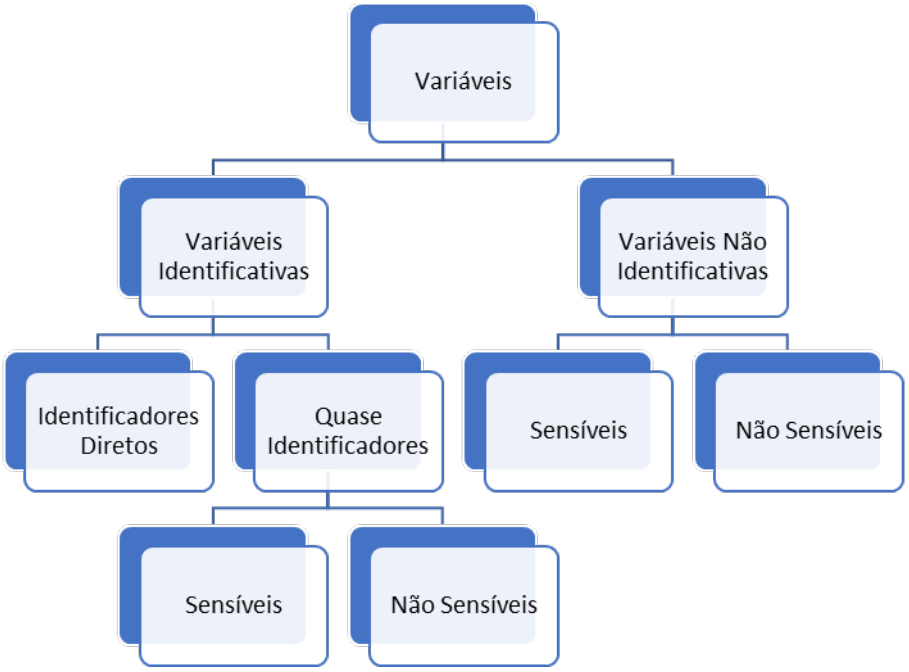

FIGURA 1.2: Classificação de Variáveis. Fonte: Adaptado de Benschop et al. (2022)

para reidentificação e, portanto, não são consideradas sensíveis a esse respeito. No entanto, as variáveis não identificativas ainda podem conter informações confidenciais/sensíveis, o que pode ser prejudicial caso a divulgação ocorra como resultado da divulgação da identidade com base nas variáveis identificativas. As variáveis não identificativas também são importantes no processo de CDE porque podem ser usadas em combinação com outras fontes de dados para fazer inferências sobre informações confidenciais (Hundepool et al. (2010) e Templ et al. (2014)).

Como podemos observar na Figura 1.2, as variáveis identificativas dividem-se em **identificadores diretos** e **quase identificadores** (ou variáveis-chave), os identificadores diretos são variáveis que identificam direta e inequivocamente um indivíduo, como números de passaporte e números do cartão de cidadão. Estes identificadores devem ser removidos do conjunto de dados antes de serem divulgados para proteger a privacidade dos indivíduos. Contrariamente, as **variáveis-chave** são um conjunto de variáveis que, quando combinadas, podem levar à reidentificação dos indivíduos no conjunto de dados. Essas variáveis incluem raça, data de nascimento, sexo e código postal, que, quando aliados a informações externas, podem identificar os indivíduos na base de dados. Quase identificadores são importantes para a análise e não podem ser simplesmente removidos. Portanto, o CDE anonimiza essas variáveis, preservando as suas informações para evitar a reidentificação. A combinação de valores de vários quase identificadores é chamada de chave, originando a

possibilidade de um registo único, isto é, identificável (Templ et al., 2014).

A classificação das variáveis como identificadores diretos ou quase identificadores depende, em parte, da possibilidade de conjuntos de dados externos conterem informações que, combinadas com os dados atuais, poderiam resultar em divulgação. A identificação e classificação de variáveis como quase identificadores depende da disponibilidade de informações em conjuntos de dados externos, entre outros fatores. Definir uma lista de possíveis cenários de divulgação com base em como os quase identificadores podem ser combinados entre si e as informações externas disponíveis é uma etapa essencial no processo de CDE. Aprofundamos os cenários de divulgação com mais detalhes no Capítulo Cenários de Divulgação.

Para a utilização dos métodos de CDE é útil classificar as variáveis como contínuas, categóricas e semicontínuas, consoante a sua classificação conseguimos escolher o melhor método de CDE. Assim, podemos definir variável contínua como sendo uma variável numérica que tem ainda a capacidade de realizar operações aritméticas (Domingo-Ferrer & Torra, 2001), por exemplo, volume, tempo e peso, entre outras. Em contrapartida, numa variável categórica não é possível realizar operações aritméticas (Domingo-Ferrer & Torra, 2001), por exemplo, o tipo de sangue ou o partido político de um indivíduo. Em adição, variáveis semicontínuas são variáveis continuas, mas que assumem valores limitados, por exemplo, a idade dos seres humanos que está dentro do conjunto que vai dos 0 aos 110 anos (Benschop et al., 2022).

Para além das classificações mencionadas, o processo CDE envolve a categorização de variáveis com base na sua sensibilidade ou confidencialidade. Quase identificadores e variáveis não identificativas podem ser classificados como sensíveis ou não sensíveis. Não se torna necessário dividir os identificadores diretos, pois são variáveis que pela sua natureza são sensíveis e imediatamente removidas da base de dados. Posto isto, segundo Templ et al. (2014), estas variáveis podem ser descritas da seguinte forma:

- **Variáveis Sensíveis**: contêm informações confidenciais que não devem ser divulgadas sem tratamento adequado usando métodos de CDE para reduzir o risco de divulgação. Exemplos são renda, religião, afiliação política, variáveis relativas à saúde, antecedentes criminais, comportamento sexual e registos médicos. Se uma variável é sensível depende do contexto: uma determinada variável pode ser considerada sensível em um país e não sensível em outro. A divulgação de variáveis confidenciais ainda pode levar à divulgação de atributos, mesmo que a divulgação de identidade seja impedida.

- **Variáveis Não Sensíveis**: contêm informações não confidenciais do indivíduo, como local de residência ou residência rural/urbana. No entanto, a classificação de uma variável como não sensível não significa que ela não precise de ser considerada no processo de CDE. Variáveis não sensíveis ainda podem servir como quase identificadores quando combinadas com outras variáveis ou outros dados externos.

Assim, na primeira fase, é necessário remover dos registos os identificadores diretos, ou seja, as variáveis que permitem a identificação direta do indivíduo, como nome, endereço, número de identificação (Bilhete de Identidade, Cartão de Cidadão, Número de Identificação Fiscal, etc.). Caso não seja suficiente, na segunda fase, é preciso identificar as combinações raras de variáveis chave que, se não forem eliminadas, podem possibilitar a identificação de algumas unidades estatísticas.

1.4 Cenário de Divulgação

Um cenário de divulgação é uma etapa fundamental na criação de um arquivo de microdados seguro, que envolve a definição de suposições realistas sobre o que um intruso pode saber sobre certos indivíduos e como ele poderia utilizar essas informações para fazer uma identificação e divulgação. Normalmente, diferentes tipos de divulgação de dados exigem diferentes cenários de divulgação e definições de risco (Benschop et al., 2022).

Para desenvolver um cenário de divulgação, os analistas devem descrever as informações que estão potencialmente disponíveis para um intruso e como podem usá-las para identificar um indivíduo. Isso envolve considerar diferentes fontes de informação que possam estar disponíveis e avaliar o risco de reidentificação por reconhecimento espontâneo ou reidentificação por correspondência de registos (Benschop et al., 2022).

Um exemplo de cenário de divulgação é o cenário "Vizinho intrometido". Neste cenário, um intruso tem conhecimento pessoal sobre um ou mais indivíduos no arquivo de microdados e tenta combinar as suas informações pessoais com os microdados para identificá-los. O arquivo externo disponível para o intruso neste cenário contém um ou alguns registos com informações pessoais detalhadas. Os meios e a estratégia de ataque do intruso dependem das informações sobrepostas para corresponder os identificadores diretos a um registo no arquivo de microdados. Outro exemplo de cenário de divulgação é o cenário "Arquivo externo". Neste cenário, um invasor tenta identificar indivíduos no arquivo de microdados associando as

informações fornecidas pelo arquivo externo ao arquivo de microdados. O arquivo externo é assumido como um registo público com informações pessoais disponíveis ao nível individual (Hundepool et al., 2010).

A escolha cuidadosa das variáveis chave é essencial para a avaliação do risco de divulgação num cenário de divulgação. Como mencionado anteriormente, diferentes tipos de divulgação de dados exigem diferentes cenários de divulgação e definições de risco. Para desenvolver um cenário de divulgação, os analistas devem descrever as informações que estão potencialmente disponíveis para um intruso e como podem usá-las para identificar um indivíduo. A partir destes cenários, as variáveis-chave podem ser identificadas e usadas para determinar o risco de divulgação.

1.5 Risco de Divulgação

Como mencionado para avaliar o risco de divulgação, primeiro é necessário fazer suposições realistas sobre o que um intruso pode saber sobre os Indivíduos e que informações estarão disponíveis para ele comparar com os microdados e potencialmente fazer uma identificação e divulgação. Intuitivamente, podemos considerar o risco como uma forma de medir a raridade de uma unidade na amostra ou na população, sendo que se considera que uma unidade está em risco se formos capazes de destacá-la das restantes (Benschop et al., 2022).

Por forma a evitar a divulgação, os arquivos de microdados não podem conter variáveis identificativas, como nome completo, morada ou número de identificação. No entanto, outras variáveis nos arquivos de microdados podem ser usadas como identificadores indiretos (Benschop et al., 2022).

Para garantir que o risco de divulgação seja mantido em níveis aceitáveis, a fonte responsável pelos dados deve tomar medidas apropriadas. Desta forma, é possível maximizar a utilidade dos dados, mesmo que a limitação da divulgação resulte na perda de informações. À medida que a perda de informações aumenta, a utilidade dos dados diminui, mas, ao mesmo tempo, o risco de divulgação também é reduzido (Mendes, 2010). A Figura 1.3 ajuda a visualizar a evolução entre a perda de informação e o risco de divulgação auxiliando na decisão das diferentes medidas de risco de divulgação.

Existe uma clara necessidade de obter medidas de risco de divulgação quantitativas e objetivas para o risco de reidentificação nos microdados. Para microdados que têm registos públicos, o risco de divulgação é conhecido, pois temos todas as variáveis de identificação disponíveis para toda a população. No entanto, para

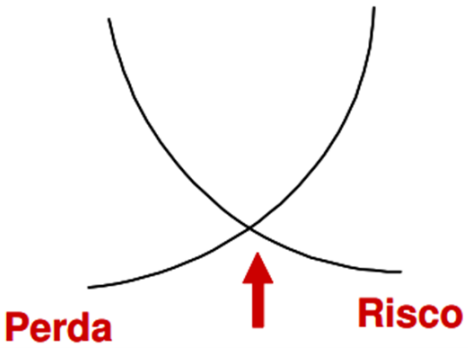

FIGURA 1.3: Comparação entre a perda de informação e o risco de divulgação. Fonte: Mendes (2010)

microdados de amostras da população, a base populacional é desconhecida ou parcialmente conhecida através da distribuição de probabilidades das variáveis contidas nesse subconjunto. Portanto, a modelação probabilística (técnica estatística usada tendo em conta o impacto de eventos ou ações aleatórias prevendo a ocorrência potencial de resultados futuros) ou a heurística (técnicas baseadas na experiência de várias tarefas, tal como, investigação, resolução de problemas, descoberta e aprendizagem) são usadas para estimar medidas de risco de divulgação a nível populacional, com base nas informações disponíveis na amostra. Este subcapítulo fornece uma visão geral dos métodos e ferramentas disponíveis para estimar as medidas de risco de divulgação (Hundepool et al., 2010).

Segundo Hundepool et al. (2010), podemos classificar as medidas de risco de divulgação em três tipos: medidas de risco baseadas em chaves[9] da amostra, aquelas baseadas em chaves na população efetuadas por modelos estatísticos ou heurísticas para estimar as quantidades de interesse e medidas baseadas na teoria de ligação de registos. Enquanto as duas primeiras classes são dedicadas à avaliação de risco para variáveis de identificação categóricas, a terceira pode ser usada para variáveis categóricas e contínuas.

Primeiramente, as medidas baseadas em chaves da amostra, considera-se que uma unidade/indivíduo está em risco se a sua combinação de pontuações na variável de identificação estiver abaixo de um determinado limite. De seguida, no segundo tipo de abordagem a maior preocupação é o risco de reidentificação, neste caso a unidade/indivíduo estão em risco se estiverem acima de um determinado limite. Por fim, quando variáveis identificativas são contínuas, não é possível explorar

[9]Uma chave, no contexto deste trabalho, é uma combinação de variáveis-chave. Fonte: Hundepool et al. (2010).

o conceito de raridade das chaves e transformamos tal conceito em raridade na vizinhança do registo, assim uma forma de medir a raridade da vizinhança é por meio de métodos de ligação de registos, isto é, através do processo que une diferentes fontes de dados num único arquivo (Hundepool et al., 2010).

1.5.1 Frequência das variáveis-chave

O foco da avaliação de risco para variáveis-chave está centrado na divulgação da identidade e na probabilidade de reidentificar corretamente os indivíduos nos dados divulgados. O risco é maior quando uma combinação de valores de variáveis-chave é rara dentro da amostra, pois os invasores que tentam corresponder indivíduos com chaves raras em conjuntos de dados externos têm uma probabilidade maior de encontrar uma correspondência correta, sendo que indivíduos com chaves mais raras têm sempre um maior risco de divulgação, mesmo quando os identificadores diretos são removidos do conjunto de dados (Hundepool et al., 2012).

O número de indivíduos que compartilham a mesma combinação de variáveis-chave determina a probabilidade de correspondência precisa em conjuntos de dados externos. Indivíduos com menor quantidade de variáveis-chave iguais têm maior probabilidade de serem correspondidos, indicando maior risco de divulgação. As frequências das chaves (f_k) fornecem uma base para a avaliação de risco, onde $f_k = 1$ representa a existência de uma combinação única de variáveis-chave num só indivíduo. Ao compreender a relação entre a raridade das chaves e o risco de divulgação, medidas eficazes podem ser implementadas para proteger as identidades dos indivíduos ao divulgar dados (Benschop et al., 2022).

Ao examinar os dados da amostra, torna-se valioso considerar a frequência dessa mesma combinação de variáveis na população, denotada como F_k, que representa o número de indivíduos na população com uma chave específica k. Sob suposições específicas, o valor esperado das frequências populacionais pode ser calculado usando o peso do desenho da amostra w_i para cada indivíduo. F_k é obtido somando os pesos amostrais de todos os registos com a mesma chave k. Semelhante a f_k, F_k permanece constante para todos os registos que partilham a mesma chave. O risco de reidentificação correta, ou a probabilidade de corresponder uma chave ao indivíduo correto na população, é determinado por $\frac{1}{F_k}$, que representa a probabilidade do pior cenário e serve como um indicador do risco de divulgação. Notavelmente, quando F_k é igual a 1, a chave é única na amostra e na população, apresentando um risco de divulgação de 1 (Benschop et al., 2022).

A medida de risco r_k, semelhante a f_k e F_k, mantém-se constante para todos os indivíduos que partilham o mesmo padrão de valores das variáveis-chave, sendo conhecida como risco individual. Os valores de r_k podem ser compreendidos como a probabilidade de divulgação para esses indivíduos ou como a probabilidade de correspondência bem-sucedida com indivíduos selecionados aleatoriamente de um ficheiro de dados externo que possui os mesmos valores das variáveis-chave (Benschop et al., 2022). Posto isto, o objetivo é impedir que um intruso obtenha $f_k = F_k = 1$, pois isto significa que é possível identificar um indivíduo único na população.

1.5.2 *K-Anonimity*

Na análise de dados, é frequentemente observado que certas variáveis ou combinações de variáveis chave possuem valores únicos para apenas um pequeno número de indivíduos. Nestes casos, o risco de reidentificação para esses indivíduos é maior em comparação com os restantes (Benschop et al., 2022).

No conceito de *k-anonymity*, um conjunto de dados protegido é considerado satisfazer o *k-anonymity* (onde $k > 1$) se, para cada combinação de atributos chave, existirem pelo menos k registos no conjunto de dados que partilham essa combinação. Se for assumido que o *k-anonymity* oferece proteção suficiente para os indivíduos, o foco pode ser na minimização da perda de informação, garantindo ao mesmo tempo que o *k-anonymity* seja mantido. Esta abordagem oferece uma solução clara para o compromisso entre a proteção de dados e a utilidade dos dados (Domingo-Ferrer & Torra, 2008).

O *k-anonymity* funciona como uma medida de risco determinada com base nos microdados a serem divulgados, considerando apenas a amostra. Um indivíduo viola o *k-anonymity* se a contagem de frequência (f_k) para uma chave específica k for inferior ao limite designado k. Através deste método, a medida de risco é o número de observações que viola o anonimato de k para um determinado valor de k, que é (Benschop et al., 2022):

$$\sum_i I(f_k < k) \tag{1.1}$$

onde I é a função indicadora e i refere-se ao número de registos. Isto é simplesmente uma contagem do número de indivíduos com uma frequência de amostra da chave menor que k, sendo a contagem maior para um k maior.

1.6 Qualidade da Informação e Correlação das Variáveis

Conforme mencionado por Karr et al. (2005), a qualidade dos dados é medida pela sua capacidade de serem utilizados de forma eficiente, económica e rápida para informar e avaliar a tomada de decisões. A qualidade dos dados é uma medida multidimensional, que vai além do nível de registo e inclui fatores como acessibilidade, relevância, atualidade, metadados[10], documentação, capacidade e expectativas dos utilizadores.

Para minimizar a perda de informação, é importante manter o conjunto de dados transformados semelhante ao conjunto de dados original, garantindo assim que os dados sejam analiticamente válidos e úteis. Segundo Winkler (2005), um conjunto de dados é considerado analiticamente válido se as médias e covariâncias nos subconjuntos de dados, os valores marginais em tabelas de dados e pelo menos uma característica da distribuição forem preservados.

Existem formas complementares de avaliar a preservação da estrutura do conjunto de dados original, incluindo a comparação entre os dados originais e os dados alterados, a comparação de algumas estatísticas dos conjuntos de dados e a análise do comportamento dos métodos de CDE usados para medir o impacto sobre a estrutura do conjunto de dados original (Domingo-Ferrer & Torra, 2001). Para este trabalho, serão efetuadas análises comparativas entre os dados originais e os dados transformados, bem como a avaliação de determinadas estatísticas nos conjuntos de dados, para avaliar o impacto dos métodos de CDE nos dados.

1.6.1 Correlação de Pearson, Kendall e Point-Biserial

Nesta secção da revisão da literatura, iremos apresentar e discutir três medidas de correlação comumente utilizadas: a correlação de Pearson, a correlação de Kendall e a correlação Point-Biserial. A análise de correlação é uma técnica estatística utilizada para avaliar a relação entre duas ou mais variáveis. Na sua essência, existem inúmeras possibilidades para a relação entre um par de variáveis aleatórias. É comummente razoável assumir, que uma aproximação adequada, é a linear, por exemplo ao examinar um par de variáveis aleatoriamente relacionadas X e Y, um gráfico de dispersão que exibe as observações conjuntas tende a agrupar-se ao longo

[10]Metadados referem-se a dados que fornecem informações sobre outros dados, descrevendo várias características dos dados, tais como a sua estrutura, tabelas e atributos, bem como valores ausentes e incompletos. Fonte: Karr et al. (2005)

de uma linha reta. Por outro lado, se não houver uma relação linear, o gráfico de dispersão afastar-se-á dessa linha (Newbold et al., 2022).

Os vários coeficientes de correlação captam diferentes aspetos de uma relação monótona, isto é, da força e direção das associações entre variáveis e são interpretados de forma distinta em análise estatística. Quando aplicados a dados contínuos, estes coeficientes retratam o grau de associação entre duas variáveis de maneiras ligeiramente diferentes. Tipicamente, uma associação monótona fortemente crescente (ou decrescente) entre as variáveis está associada a valores positivos (ou negativos) em todos os coeficientes de correlação simultaneamente (Chok, 2010).

O coeficiente de correlação de Pearson, amplamente utilizado para medir a associação entre duas variáveis contínuas, é calculado como a covariância das duas variáveis (X e Y) dividida pelo produto dos seus desvios padrão (σ). Este coeficiente é comumente representado pela letra grega ρ (rho):

$$\rho = \frac{Cov(X, Y)}{\sigma_X \sigma_Y} \tag{1.2}$$

O coeficiente de correlação de Pearson está limitado entre -1 e 1. Uma associação monótona positiva, onde duas variáveis tendem a aumentar ou diminuir juntas, resulta em $\rho > 0$, enquanto uma associação monótona negativa, onde uma variável tende a aumentar enquanto a outra diminui, resulta em $\rho < 0$. Um valor de ρ igual a 0 indica a ausência de uma associação monótona, ou a ausência de qualquer associação (Chok, 2010).

Em contrapartida, o coeficiente de correlação de Kendall é especialmente projetado para medir a associação entre duas variáveis ordinais[11], que não precisam necessariamente de ser variáveis intervalares, tendo em conta relações não lineares e a presença de *outliers*[12] nos dados. A estimativa do coeficiente de correlação de Kendall considera (x_i,y_i) e (x_j,y_j) duas observações de variáveis aleatórias contínuas, obtidas numa amostra de n observações, expressa da seguinte forma (Chok, 2010):

$$\tau = \frac{\sum_{i=1}^{n} \sum_{j=1}^{n} sgn(x_i - x_j) sgn(y_i - y_j)}{n(n-1)} \tag{1.3}$$

[11]Um variável ordinal é um subconjunto das variáveis categóricas, sendo que assume valores num intervalo ordenado de categorias. Fonte: Hundepool et al. (2012).

[12]Outlier é um valor incomum que foi reportado corretamente, mas não está de acordo com o padrão dos restantes valores. Por exemplo, uma base de dados com as idades de uma população onde o intervalo entre a pessoa mais velha e a restante população é de 20 anos. Fonte: Brand (2002).

Onde as funções sinal,

$$\text{sgn}(x_i - x_j) = \begin{cases} 1 & \text{se } (x_i - x_j) > 0 \\ 0 & \text{se } (x_i - x_j) = 0 \\ -1 & \text{se } (x_i - x_j) < 0 \end{cases} ; \text{sgn}(y_i - y_j) = \begin{cases} 1 & \text{se } (y_i - y_j) > 0 \\ 0 & \text{se } (y_i - y_j) = 0 \\ -1 & \text{se } (y_i - y_j) < 0 \end{cases} \quad (1.4)$$

No entanto, o coeficiente de correlação Point-Biserial é uma medida de associação entre uma variável dicotómica e uma variável contínua, equivalente ao coeficiente de correlação de momento do produto de Pearson. Os valores variam de 1, indicando uma relação positiva perfeita, passando por zero, significando nenhuma associação, até -1, representando uma correlação negativa perfeita (Kornbrot, 2014).

Assim, esta correlação é calculada considerando uma variável aleatória contínua Y e outra variável aleatória dicotómica X, assumindo apenas dois valores, assume-se ainda que n observações (Y_k,X_k), com k = 1,2,... n estão disponíveis. Esta correlação é expressa (NCSS, s.d.):

$$r_{pb} = \frac{\bar{Y}_1 - \bar{Y}_0}{s_Y} \sqrt{\frac{np_0(1-p_0)}{n-1}} \quad (1.5)$$

Onde,

$$s_Y = \sqrt{\frac{\sum_{k=1}^{n}(Y_k - \bar{Y})^2}{n-1}}; \bar{Y} = \frac{\sum_{k=1}^{n} Y_k}{n}; p_0 = 1 - \frac{\sum_{k=1}^{n} X_k}{n} \quad (1.6)$$

Compreender estas medidas de correlação é crucial, pois permitem a avaliação das relações entre as variáveis, fornecendo informações sobre padrões, dependências e possíveis conexões causais. No capítulo 4, será possível explorar a aplicação e interpretação destas medidas de correlação na base de dados.

Capítulo 2

Controlo de Divulgação Estatística

Uma parte significativa dos dados obtidos por agências estatísticas não pode ser publicada diretamente devido a preocupações com a privacidade e confidencialidade, preocupações estas que abrangem aspetos tanto legais como éticos. O CDE procura modificar e transformar os dados de forma que seja possível publicá-los ou divulgá-los sem revelar as informações confidenciais que eles contêm. Ao mesmo tempo, esforça-se para minimizar a perda de informações resultante do processo de anonimização (Benschop et al., 2022).

Como mencionado na Introdução, a crescente ênfase na proteção de dados é uma preocupação que se estende não apenas ao ambiente civil, mas também ao âmbito da Marinha Portuguesa. Nesse contexto, uma série de procedimentos e estratégias é adotada para garantir a salvaguarda das informações, dentre as quais se incluem a atribuição de códigos anonimizados e a eliminação de dados classificados como sensíveis. Detalhes sobre a implementação desses métodos foram fornecidos pela Direção de Análise e Gestão da Informação (DAGI), uma entidade especializada encarregada da gestão e análise da informação.

Este capítulo encontra-se dividido em dois subcapítulos, no primeiro subcapítulo será feita uma descrição sobre métodos perturbativos, por sua vez, no segundo subcapítulo serão descritos os métodos não perturbativos.

Como mencionado na Introdução, existem dois tipos de métodos de CDE, estes métodos visam alterar a informação de tal modo que quando os dados são divulgados o risco de identificação de indivíduos, empresas ou organizações seja diminuto (Mendes, 2010). Segundo Hundepool et al. (2010), "os métodos de CDE minimizam o risco de divulgação a um nível aceitável, com a divulgação do máximo de informação possível".

Ainda, podemos caracterizar os métodos de CDE como a análise e perturbação dos dados originais para eliminar/reduzir ao mínimo a possibilidade de identificação direta ou indireta das pessoas envolvidas, sem incorrer em esforços ou custos excessivos (Instituto Nacional de Estatística, 2019). Sendo que, com a utilização destes métodos pretende-se que os dados sejam analiticamente válidos e úteis, mas com o mínimo de perda de informação e risco de divulgação.

2.1 Métodos Perturbativos

Métodos perturbativos distorcem os dados originais antes da sua divulgação, introduzindo um erro propositalmente por razões de confidencialidade, permitindo a divulgação do conjunto completo de microdados com valores perturbados em vez de valores exatos (Hundepool et al., 2012).

A utilização de métodos perturbativos serve para modificar os valores das variáveis antes da sua publicação, sendo de referir que este método pode ser usado tanto para variáveis contínuas como categóricas (Mendes, 2010). Do mesmo modo que consideramos qual a melhor estratégia a adotar consoante as variáveis que temos ao nosso dispor, devemos ainda ter em mente que o resultado da análise estatística dos dados perturbados não deve diferir significativamente da análise estatística calculada a partir dos dados originais (Domingo-Ferrer & Torra, 2001), ou seja, devemos considerar que apesar de certos atributos terem sido alterados, os cálculos estatísticos efetuados com os dados originais devem ser equivalentes aos dados transformados.

Na Tabela 2.1 podemos observar os diferentes métodos perturbativos, que informam se são mais adequados para dados contínuos ou categóricos.

Segundo Liew et al. (1985), a distorção de dados pela probabilidade de distribuição é um método que permite gerar amostras distorcidas com a mesma função densidade e propriedades estatísticas das amostras originais.

A microagregação, conforme descrita por Hundepool et al. (2012), envolve agrupar registos individuais em clusters, garantindo que nenhum indivíduo predomine no grupo, e substituindo os valores individuais por valores calculados através de microagregações antes da publicação.

Ainda, Hundepool et al. (2012) considera que reamostragem é uma técnica estatística que envolve a repetição da seleção de amostras aleatórias de um conjunto de dados original, sendo que cria t amostras independentes dos valores originais

Método	Dados Contínuos	Dados Categóricos
Adição de Ruído	X	
Distorção de dados pela probabilidade de distribuição		X
Microagregação	X	
Reamostragem	X	
Compressão com perdas	X	
Imputação Múltipla	X	
Camuflagem	X	
PRAM		X
Troca de classificação	X	X
Arredondamento	X	

Tabela 2.1: Métodos Perturbativos. Fonte: Adaptado de Domingo-Ferrer e Torra (2001)

das variáveis, cada uma com n registos, que são ordenados com base num critério comum. A partir destas amostras, uma variável camuflada é formada utilizando a média dos valores ordenados em cada uma delas.

Segundo Rubin (1987), imputação múltipla consiste numa técnica estatística projetada para substituir um valor ausente ou deficiente por dois ou mais valores aceitáveis. Posteriormente, Rubin (1993) propôs utilizar estas imputações múltiplas numa base de dados original e divulgar apenas a base de dados sintética[13] obtida.

A compressão com perdas é uma técnica de compressão de dados recomendada para dados contínuos. Neste método, são deliberadamente introduzidas alterações ou modificações nos dados originais, mesmo que isso resulte na perda ou distorção de alguma informação. Um exemplo ilustrativo deste método é tratar os microdados numéricos de maneira semelhante a uma imagem, aplicando modificações na representação dos dados, alterando assim a resolução da imagem (Domingo-Ferrer & Torra, 2001).

Ainda, Domingo-Ferrer e Torra (2001) menciona que a camuflagem utiliza otimização para ocultar registos sensíveis num conjunto infinito, resultando em respostas intervalares, sendo este método considerado adequado para dados contínuos pois transforma respostas pontuais em intervalares.

Segundo Hundepool et al. (2010), PRAM (*Post Randomization Method*) é um método que tem como objetivo induzir o intruso a classificar erradamente os

[13]Os dados sintéticos são gerados aleatoriamente, preservando estatísticas específicas ou relações internas encontradas na base de dados original. Podemos ainda dividir os dados sintéticos em três tipos: dados totalmente sintéticos, dados parcialmente sintéticos e dados híbridos. Fonte: Hundepool et al. (2010)

dados utilizando um mecanismo conhecido e predeterminado. Basicamente, a cada registo num ficheiro de microdados ocorre uma alteração em uma ou mais variáveis categóricas considerando uma certa probabilidade.

A troca de classificação é um método onde os valores das variáveis selecionadas são trocados nos registos, preservando efetivamente as classificações dos valores enquanto alteram seus valores reais (Hundepool et al., 2012).

Considerando os objetivos deste trabalho e os dez métodos apresentados por Domingo-Ferrer e Torra (2001), foram selecionados e descritos com mais detalhe os métodos mais simples de utilizar e de compreender considerando a base de dados em questão e o seu nível de confidencialidade: adição de ruído e arredondamento.

2.1.1 Adição de Ruído

A adição de ruído consiste em adicionar "pequenos" valores aleatórios nos valores originais de uma variável. Este método é mais eficaz na proteção de variáveis contínuas e impede a correspondência exata, sendo que o benefício da adição de ruído é a natureza contínua do ruído que terá média zero, o que torna impossível realizar uma correspondência exata com arquivos externos. No entanto, dependendo da magnitude do ruído adicionado, uma correspondência aproximada por intervalo ainda pode ser possível. Ao utilizar a adição de ruído para proteger os dados, é crucial considerar fatores como o tipo de dados, o uso pretendido dos dados e as propriedades dos dados antes e após a adição de ruído, incluindo a análise da distribuição, especialmente a média, covariância e correlação entre os conjuntos de dados perturbados e originais (Brand, 2002).

Para dados que possuem outliers, é importante observar que quando a distribuição de dados perturbada é semelhante à distribuição de dados original, a adição de ruído não protegerá os outliers. Como podemos observar na Figura 2.1 após a adição de ruído, esses outliers geralmente ainda podem ser detetados como outliers e, portanto, facilmente identificados (Brand, 2002).

Como mencionado em Hundepool et al. (2010), existem vários algoritmos de adição de ruído, sendo os principais: adição de ruído não correlacionado, adição de ruído correlacionado, adição de ruído e transformação linear e adição de ruído e transformação não linear. Contudo Domingo-Ferrer e Torra (2001), consideram que a única forma de preservar a correlação será através da adição de ruído aleatório com a mesma estrutura de correlação dos dados originais. Mais ainda, segundo

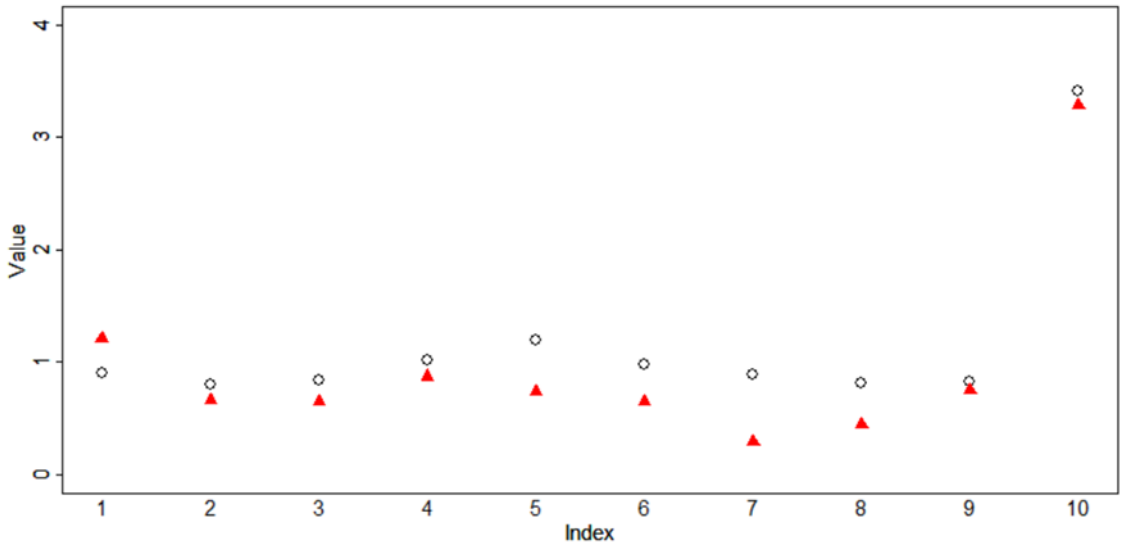

FIGURA 2.1: Ilustração do efeito da adição de ruído. Fonte: Brand (2002)

Flossmann e Lechner (2006), a maneira mais simples de proteger os dados seria adicionando ruído nas covariâncias.

Foi utilizado o método de adição de ruído não correlacionado por ser considerado um método mais simples e direto de implementar em comparação com os outros métodos mencionados. Ao adicionar ruído não correlacionado, a preocupação principal é preservar as médias e as relações entre as variáveis, tornando-o uma escolha conveniente quando a simplicidade e a rapidez de aplicação são importantes.

A proteção dos dados através da adição de ruído não correlacionado consiste em substituir cada valor no conjunto de dados original adicionando um número aleatório. Esse número aleatório é escolhido a partir de uma distribuição normal e a sua variância é proporcional à variância da variável original. Este tipo de adição de ruído pode ser descrito por (Hundepool et al., 2012):

$$z_j = x_j + \epsilon_j, \tag{2.1}$$

considerando x_j o valor original, z_j o novo valor, $\varepsilon_j \sim N(0, \sigma^2_{\varepsilon_j})$ e $\sigma_{\varepsilon_j} = \alpha \times \sigma_j$, considerando σ_j o desvio padrão dos dados originais e α uma constante positiva usada para variar a "quantidade de ruído", normalmente entre 0,5 e 2. Apesar deste método manter as médias e as relações entre as variáveis, não preserva as variâncias exatas nem os coeficientes de correlação, podendo assim, demonstrar alterações na dispersão de dados ou perda de informações relevantes.

A adição de ruído correlacionado também preserva as médias e adicionalmente permite a preservação dos coeficientes de correlação. A diferença com o método anterior é que a matriz de covariância dos erros agora é proporcional à matriz de covariância dos dados originais (Hundepool et al., 2012).

Em contrapartida, Kim (1986) menciona que através do método de transformações lineares é possível assegurar a estimativa da matriz de covariância em variáveis camufladas usando adição de ruído.

Por fim, Sullivan (1989) propõe um método que combina adição de ruído e transformações não lineares, as etapas incluem cálculo da distribuição, suavização, conversão em variáveis uniformes e padrão, adição de ruído e transformação reversa.

2.1.2 Arredondamento

O arredondamento envolve ajustar os valores de todas as células de uma tabela para uma base específica, criando incerteza em relação ao valor real de cada célula. Introduz-se uma pequena, mas geralmente aceitável, quantidade de distorção nos dados originais. O arredondamento é considerado um método eficaz para proteger tabelas de frequência, especialmente quando várias tabelas são geradas a partir de um único conjunto de dados. Ele fornece proteção para frequências baixas e células vazias. O método é simples de implementar, pois os dados são visivelmente perturbados (Hundepool et al., 2012). Exemplos da utilização do método de arredondamento é removendo números decimais ou arredondando ao valor mais próximo (Benschop et al., 2022).

Existem vários tipos de arredondamento, incluindo arredondamento convencional, arredondamento aleatório, arredondamento controlado e arredondamento semi-controlado. No arredondamento convencional, cada célula é arredondada para o múltiplo mais próximo de uma base especificada. As vantagens deste arredondamento são que este método arredonda independentemente os valores agregados (somas ou totais) de linhas e colunas de dados dos valores individuais dentro de cada célula, garantindo que a soma geral dos dados permaneça consistente após o arredondamento. Além disso, ao lidar com várias bases de dados ou tabelas que representam diferentes aspetos da mesma população, o arredondamento convencional garante que as células que representam os mesmos registos nestas tabelas tenham valores arredondados idênticos, mantendo assim a integridade dos dados e protegendo a privacidade dos registos individuais (Hundepool et al., 2012).

2.2 Métodos Não Perturbativos

Os métodos não perturbativos não distorcem os dados originais, em vez disso, reduzem a quantidade de informações divulgadas através de supressões parciais ou diminuição do detalhe no conjunto de dados original (Hundepool et al., 2012).

A utilização de métodos não perturbativos pretende não afetar tão drasticamente os dados, suprimindo parcialmente ou reduzindo o detalhe no conjunto de dados originais (Domingo-Ferrer & Torra, 2001). Os métodos não perturbativos têm a capacidade de ser utilizados com variáveis categóricas, mas só alguns podem ser utilizados para variáveis contínuas, como podemos observar na Tabela 2.2.

Método	**Dados Contínuos**	**Dados Categóricos**
Amostragem		X
Recodificação Global	X	X
Codificação Superior e Inferior	X	X
Supressão Local		X

Tabela 2.2: Métodos Não Perturbativos. Fonte: Adaptado de Domingo-Ferrer e Torra (2001)

A amostragem refere-se à prática de publicar um subconjunto dos registos de microdados originais em vez de todo o conjunto de dados. Este subconjunto, denotado como amostra S, é selecionado para representar uma fração dos registos originais. Embora a amostragem possa oferecer proteção para variáveis contínuas sob certas suposições, o risco de exposição permanece quando dados administrativos externos estão envolvidos (Hundepool et al., 2012).

A codificação superior e inferior, ao contrário da recodificação global que se aplica a todos os valores, combina valores extremos acima e abaixo de um limite numa nova categoria. Esta técnica é benéfica quando a maioria dos valores está concentrado no centro da distribuição e as categorias periféricas contêm poucas observações (Benschop et al., 2022).

Considerando os objetivos deste trabalho e os quatro métodos apresentados por Domingo-Ferrer e Torra (2001), foram selecionados e descritos com mais detalhe os métodos mais simples de utilizar e de compreender considerando a base de dados em questão e o seu nível de confidencialidade: recodificação global e supressão local.

2.2.1 Recodificação Global

A recodificação é um método determinístico utilizado para reduzir o número de categorias ou valores distintos associados a uma variável. O objetivo é alcançado através da combinação ou agrupamento de categorias para variáveis categóricas ou pela criação de intervalos para variáveis contínuas. Ao contrário de outras técnicas utilizadas para mitigar o risco de divulgação, a recodificação é aplicada a todas as observações de uma variável específica, não apenas às consideradas em risco. Existem dois tipos de recodificação, estas são a recodificação global e a codificação superior e inferior (Benschop et al., 2022).

Em adição, este processo reduz efetivamente o número de categorias disponíveis nos dados, potencialmente diminuindo o risco de divulgação, especialmente para categorias com poucas observações. No entanto, é importante observar que a recodificação global também resulta na perda do detalhe das informações disponíveis para análise (Benschop et al., 2022).

Como mencionado, neste método, uma variável categórica V_i combina várias categorias originando uma nova variável, denominada V_i'. No caso de variáveis contínuas, a recodificação global envolve a substituição de V_i por uma versão discretizada, V_i', que representa um intervalo finito mapeado a partir do intervalo potencialmente infinito de V_i. Esta técnica de recodificação é particularmente adequada para microdados categóricos, pois permite a proteção de registos que apresentem combinações incomuns de variáveis categóricas (Hundepool et al., 2012).

2.2.2 Supressão Local

Segundo Hundepool et al. (2012), supressão local é um método que diminui o risco de divulgar informação sobre indivíduos ou empresas suprimindo valores individuais em variáveis identificativas. Em adição, De Waal e Willenborg (1999), referem que uma combinação num conjunto de microdados que não ocorre com frequência suficiente pode ser protegida por supressão local, ou seja, um ou mais valores nessa combinação podem ser excluídos, sendo que a supressão local do valor de uma variável não deverá ser aplicada a todos os registos num conjunto de microdados, mas apenas a alguns dos registos.

Este método é fundamentalmente orientado para variáveis categóricas, no entanto também pode ser utilizado em variáveis contínuas. A abordagem mais simples para determinar que valores das variáveis devem ser suprimidos localmente é

examinando cada combinação individualmente e lidar com cada registo separadamente. Existem dois métodos possíveis para fazer isso. Primeiramente, quando um valor é suprimido, ele é imediatamente marcado como ausente, originando um conjunto de microdados utilizado para determinar se a combinação é segura. Alternativamente, o conjunto de microdados original pode ser usado para avaliar a segurança das combinações. No entanto, ambas as abordagens apresentam desafios e complicações, tal como, o aparecimento de combinações que parecem, incorretamente, não ocorrer com muita frequência ou demasiada supressão dos dados (Willenborg & De Waal, 1996).

Com base no trabalho realizado por Willenborg e De Waal (1996), se o objetivo for minimizar o número de supressões locais, não é possível decidir separadamente que valores devem ser suprimidos para cada combinação insegura e registo. Em vez disso, é necessário tomar uma decisão simultânea. Quando há um grande número de combinações inseguras, é recomendado usar a supressão local para suprimir algumas dessas combinações.

Durante a análise dos dados, após utilização de métodos de CDE, por forma a realmente garantir a proteção dos dados divulgados, sugeriu-se a utilização do método de supressão local em alguns dos casos únicos.

Capítulo 3

Tratamento e Implementação da Proteção dos Dados

No presente capítulo será descrita a forma como se procedeu ao tratamento e análise dos dados. Em primeira instância, será abordada a forma como foi obtida a base de dados. De seguida, será apresentada uma descrição das variáveis, separando as mesmas entre variáveis identificativas e variáveis-chave. E por fim, serão descritas todas as ações de pré-processamento inerentes ao desenvolvimento da base de dados transformada, isto é, os métodos de CDE usados.

3.1 Base de Dados

A base de dados que servirá de caso de aplicação foi pré-processada e fornecida pela ASPOF M Frazão Vala no âmbito da realização da sua dissertação de mestrado. Para mais detalhe, sobre a realização e pré-processamento da base de dados fornecida pela mesma pode ser consultado o artigo de Moura et al. (2023) na plataforma Mar-IA.

Esta base de dados é composta por 10553 registos e possui 51 variáveis, registos estes obtidos através da junção de três fontes de dados: a Direção-Geral de Recursos Naturais, Segurança e Serviços Marítimos de Portugal, o Código das Nações Unidas para Locais de Comércio e Transporte e o Registo da Frota da UE.

Por forma, a obter uma melhor perceção destas variáveis e da sua classificação entre variável categórica ou contínua é possível observar o Apêndice A. A descrição de cada uma das variáveis da base de dados será efetuada com base no Regulamento da Comissão da UE 2017/218, relativo ao registo da frota de pesca da UE:

- Reg_number: matrícula da embarcação;

- Latitude: latitude onde ocorreu a fiscalização;
- Longitude: longitude onde ocorreu a fiscalização;
- GDH (Grupo-Data-Hora): conjunto de caracteres que expressam data e hora;
- IRCS (*International radio call sign*) indicator: foi atribuída a letra Y se a embarcação apresenta a bordo um rádio internacional, caso contrário aparece a letra N;
- Licence indicator: foi atribuída a letra Y se a embarcação apresenta licença de pesca em conformidade, caso contrário aparece a letra N;
- VMS (*Vessel Monitoring System*) indicator: foi atribuída a letra Y se a embarcação apresenta um sistema de monitorização por satélite, caso contrário aparece a letra N;
- ERS (*Electronic Reporting System*) indicator: foi atribuída a letra Y se a embarcação apresenta um sistema de comunicação eletrónica (diário de bordo), caso contrário aparece a letra N, em adição a célula pode aparecer em branco caso não se tenha informação;
- AIS (*Automatic Identification System*) indicator: foi atribuída a letra Y se a embarcação apresenta um sistema de identificação automática, caso contrário aparece a letra N, em adição a célula pode aparecer em branco caso não se tenha informação;
- MMSI (*Maritime Mobile Service Identity*): número identificativo da embarcação de serviço móvel marítimo;
- Unit: unidade de Marinha anonimizado que fiscalizou a embarcação em questão;
- Sub_Type: subtipos da embarcação de pesca registados durante as fiscalizações (15 subtipos). Os subtipos existentes são: arrastão, armadilhas, polivalente, emalhar/tresmalho, palangreiro, ganchorra, navio pesca à linha, outras, cercador, salto e vara, navio de apoio, apanha de algas, desconhecido, navio fábrica e não identificado;
- Gear: arte em uso durante a fiscalização (23 artes de pesca na base de dados);

• Result: Resultado da fiscalização, isto é, se foi considerada legal (LEGAL) ou presumível infração (PRESUM);

• Infraction: Infração cometida pela embarcação, pode ser classificada de 1 a 14 em numeração romana: I - Diário de bordo inexistente, II - Diário de bordo preenchido incorretamente, III - Artes proibidas, IV - Pesca em zona proibida ou interdita, V - Pesca proibida por potência motora ou arqueação excessiva, VI - Capturas indevidas por pesca interdita, VII - Capturas indevidas por captura acessória, VIII - Capturas indevidas por pescado de tamanho inferior ao mínimo legal, IX - Atividade exercida sem licença ou autorização, X - Sinalização e/ou identificação indevida (s) das artes de pesca, XI - Sinalização e/ou identificação indevida (s) da embarcação, XII - Diversos: Certificados Inválidos, XIII - Diversos: Inscrição marítima inexistente/inválida e XIV - Diversos: Outros (exemplos: falta de documentos a bordo, falta de pirotécnicos, falta de meios de salvação, extintores caducados, entre outros.) (Pinto, 2017);

• CFR (*Common Fleet Register*): número de registo da embarcação, constituído por três letras representantes do país e nove números (2285 embarcações);

• Vessel_Type: representa o tipo de embarcação relativamente à sua arte segundo a abreviação padrão da *International Standard Statistical Classification of Fishery Vessels*, apresentamos de seguida a definição dos 12 códigos que aparecem na base de dados em questão: DB (*using boat dredge*/embarcação ganchorra), FX (*fishing vessels not specified*/embarcação não especificada), LL (*longliners*/embarcação palangreiro), LP (*pole and line vessels*/embarcações de salto e vara), MO (*multipurpose vessels*/embarcação polivalente), MOX (*multipurpose vessels nei*[14]/embarcação polivalente não incluído em outro lugar), SN (*seiner netters*/embarcações com rede de arrasto), SP (*purse seiners*/embarcação com rede de cerco), TTP (*stern trawlers factory*/navio fábrica de arrasto), TTW (*stern trawlers wet-fish*/ navio de peixe fresco de arrasto), TU (*outrigger trawlers*/embarcações de arrasto) e WOX (*trap setters nei*/embarcação de armadilhas não incluído em outro lugar);

• Main fishing gear: equipamento de pesca mais utilizado a bordo do navio durante um período de atividade anual ou por campanha de pesca. De seguida serão descritos os 11 códigos presentes na base de dados: DRB (*towed dredges*/arrasto com draga rebocada (ganchorra)), FPO (*pots*/armadilhas nassas (covos)), GNS (*set gillnets (anchored)*/rede de emalhar fundeada), GTR (*trammel nets*/tresmalho), LHP

[14]NEI significa "*not elsewhere included*". Utiliza-se esta sigla para indicar que uma embarcação não se enquadra em nenhuma outra categoria definida.

(*handlines and hand-operated pole-and-lines*/linhas de mão e linhas de vara (operadas manualmente)), LLD (*drifting longlines*/palangres de deriva), LLS (*set longlines*/palangres fundeadas), OTB (*single boat bottom otter trawls*/redes de arrasto pelo fundo), PS (*purse seines*/redes de cerco com retenida), SB (*beach seines*/redes de alar para a praia) e TBB (*beam trawls*/redes de arrasto pelo fundo de vara), segundo *International Standard Statistical Classification of Fishing Gear*;

- Subsidiary fishing gears 1-5: equipamento de pesca subsidiário a bordo do navio, esta variável encontra-se dividida em cinco colunas.

- LOA (*length overall*): comprimento fora a fora da embarcação, em metros, isto é, comprimento máximo de uma embarcação da proa à popa;

- LBP (*length between perpendiculars*): LBP ou LPP corresponde ao comprimento entre perpendiculares da embarcação, em metros, isto é, distância medida paralelamente à linha de água, entre as perpendiculares a vante e a ré;

- Tonnage GT (*Gross Tonnage*): corresponde à arqueação bruta da embarcação, em toneladas;

- Other tonnage: nos termos da Convenção de Oslo ou de acordo com uma definição a precisar pelo Estado-Membro, em toneladas;

- Power of main engine: potência do motor principal da embarcação, em kilowatts;

- Power of auxiliary engine: potência auxiliar da embarcação, em kilowatts;

- GTs: aumento de arqueação permitido por razões de segurança;

- Hull material: codificação do material do casco, são números de 1 a 7, correspondendo a madeira, metal, fibra de vidro/plástico, outro, desconhecido, poliéster e alumínio, respetivamente;

- Date of entry into service: data em que a embarcação entrou ao serviço;

- Year of construction: ano em que a embarcação foi contruída;

- Real_local: código relativamente ao local de registo da embarcação;

- Local_Name: nome da localidade onde foi registada a embarcação (44 localizações portuguesas);

- I – XIV: variáveis indicativas dos diferentes tipos de infrações;

- Number_infracs: o número de infrações registadas para a embarcação;

• NUTS II_Code: O código NUTS II representa a região onde a embarcação foi registada, considerando apenas para as embarcações portuguesas.

Com base no que foi descrito no Capítulo 1 sobre a classificação de variáveis e o risco de divulgação das mesmas, podemos deduzir que nesta base de dados o número de registo, o CFR, o MMSI e a unidade fiscalizadora são identificadores diretos das embarcações, sendo a remoção deste tipo de variáveis essencial para garantir a privacidade e anonimato dos indivíduos. Contudo, por forma a manter as informações necessárias para uma análise adequada, o CFR e as unidades fiscalizadoras serão utilizadas após a sua anonimização.

Adicionalmente, apesar de ser difícil descobrir a identidade das embarcações através destas variáveis de uma maneira singular, ao analisar a combinação de valores de variáveis-chave podemos obter a identificação da embarcação fiscalizada. Como exemplos, temos as variáveis *GDH*, *IRCS indicator*, *LOA*, *LBP*, year of construction, entre outros.

Consequentemente, as seguintes variáveis foram selecionadas para serem removidas ou transformadas por serem consideradas variáveis-chave:

• ***GDH*** - a hora exata da fiscalização quando combinada com a latitude e longitude, têm potencial de ser verificadas online e identificar a embarcação de pesca. Como a base de dados apresenta outras variáveis de tempo, esta variável foi eliminada;

• **Localização** – as variáveis referentes ao local onde as embarcações de pesca foram registadas são consideradas cruciais para a análise da base de dados, portanto, não foram eliminadas;

• ***IRCS indicator*, *Licence indicator*, *VMS indicator*, *ERS indicator* e *AIS indicator*** – estas variáveis apesar de serem binárias (apenas com valor de Y ou N) quando combinadas, por exemplo, com a variável *Main Fishing Gear* podem levar à identificação de 29 embarcações, posto isto, e considerando a importância reduzida das mesmas foram removidas da base de dados;

• ***Date of entry into service* e *Year of construction*** – a combinação destas duas variáveis pode levar à identificação das embarcações, por forma, a mitigar o risco elevado manteve-se apenas o ano;

• ***Subsidiary fishing gears 1–5*** – estas cinco variáveis apresentam um elevado número de valores omissos, em adição a sua combinação leva à identificação de embarcações inequivocamente, isto é, com um conjunto de artes secundárias único

é possível associar uma embarcação de pesca. Assim, considerando o nível de risco e a existência da variável *Main Fishing Gear* que nos dá informações suficientes, optou-se por eliminar estas variáveis;

• ***LOA*, *LBP*, *Tonnage GT*, *Other tonnage*, *Power of main engine* e *Power of auxiliary engine*** – estas variáveis têm muitos valores únicos e por isso requerem ser transformadas;

• ***GTs*** – a grande quantidade de valores omissos representa um risco significativo para as embarcações que têm algum valor nesta variável, deste modo removeu-se a variável da base de dados.

Como mencionado anteriormente é necessário manter a variável CFR e a unidade fiscalizadora, como tal, primeiramente dividiu-se a variável CFR em dois valores distintos, isto é:

• ***CFR_#*** - este valor é atribuído às embarcações que têm um *CFR* conhecido. O *CFR* original é substituído por *CFR_#*, onde # representa um número atribuído a cada embarcação. Isto possibilita a análise de embarcações com CFR conhecidos, possivelmente levando à identificação de infratores reincidentes e à avaliação da frequência das fiscalizações;

• **NOCFR_#** - Este valor é atribuído a embarcações sem um número CFR conhecido. Nestes casos, uma variável recebe o valor NOCFR_#, onde # representa um número único atribuído a cada embarcação.

De seguida, a unidade fiscalizadora foi modificada atribuindo um código anonimizado (*unit_#*), onde cada navio recebe um número exclusivo #, mantendo o anonimato dos navios da Marinha Portuguesa, e ainda, permitindo a identificação e análise das unidades individualmente.

3.2 Aplicação dos Métodos de Controlo de Divulgação Estatística

Tendo em consideração as variáveis *Date of entry into service* e *Year of construction*, reconhece-se que estas mantêm características significativas mesmo quando representadas apenas pelo ano. Para proteger a confidencialidade, pode ser aplicado um método de arredondamento ao *Year of construction*, arredondando-o para a década mais próxima. Da mesma forma, a *Date of entry into service* pode ser arredondada para o ano mais próximo utilizando a mesma técnica. No entanto,

é importante salientar que, apesar da aplicação do arredondamento, ainda existem 13 registos (para $k = 3$) em que a combinação das variáveis-chave *Main Fishing Gear*, *Date of entry into service* e *Year of construction* pode representar potenciais riscos de reidentificação. De modo a colmatar este risco foi adicionado ruído, como mencionado na Equação 2.1.

Para salvaguardar a privacidade com maior eficácia, o valor divulgado na base de dados será a representação arredondada de z_j para o número inteiro mais próximo. Ao utilizar este método, os dados originais são distorcidos aleatoriamente e o desvio padrão é ajustado como uma proporção do desvio padrão da variável original, garantindo que os novos valores apresentem uma diferença não significativa em relação aos valores reais. Esta abordagem preserva a privacidade ao permitir informações e análises dos dados significativas a partir do conjunto de dados modificado.

Após consideração das várias combinações de variáveis, foi determinado que as variáveis *LOA, LBP, Tonnage GT, Other tonnage, Power of main engine* e *Power of auxiliary engine* representam um risco potencial de identificação das embarcações. Para resolver este problema, estas variáveis foram submetidas a transformações específicas para proteger a privacidade e a confidencialidade do conjunto de dados.

Foram implementadas transformações específicas para mitigar os riscos identificados associados às variáveis *Other tonnage* e *Power of auxiliary engine*. A variável *Other tonnage* foi arredondada para o número inteiro mais próximo, enquanto a variável *Power of auxiliary engine* foi arredondada para a dezena mais próxima. Esse processo de arredondamento pode ser visto como uma forma de recodificação global, em que os valores das variáveis são ajustados para ficarem dentro de um intervalo com uma amplitude de 1 ou 10, respetivamente. O resultado do arredondamento será, na verdade, o centro desse intervalo.

Para a variável *Power of main engine*, foi aplicada o arredondamento para a dezena mais próxima, contudo, ainda se observavam muitos valores únicos, por esta razão, optou-se por adicionar ruído aleatório à variável usando a Equação 2.1. Adicionalmente, se o valor resultante dessa transformação fosse zero, ele seria substituído pelo valor mínimo da variável original *Power of main engine*, e se o valor transformado fosse negativo este seria convertido para o seu valor absoluto, dada a natureza da variável. Um procedimento semelhante foi usado para lidar com o risco de identificação associado à variável *Tonnage GT*, onde ruído aleatório também foi adicionado usando a equação 2.1, com as mesmas inclusões relativamente aos valores nulos e negativos.

Para as variáveis *LOA* e *LBP*, o simples arredondamento para o número inteiro mais próximo observou-se não ser suficiente. Para tal, foi introduzido um ruído aditivo nos valores originais usando a equação 2.1, arredondando para a casa decimal mais próxima para resolver esse problema. Dada a natureza das variáveis considerou-se apenas valores positivos, seguindo o mesmo procedimento usado.

Após a aplicação destas transformações, foi necessário aplicar a técnica de supressão local aos dados relacionados com quatro embarcações onde uma das variáveis categóricas continha valores únicos que as identificava. Obviamente, fomos forçados a remover informação destas embarcações do conjunto de dados para garantir ainda mais a necessária confidencialidade. Além disso, foi usada a técnica de imputação parcial para substituir valores específicos de variáveis contínuas que apresentavam um risco de identificação de 100%. O valor imputado foi o resultado da média dos três valores dessa mesma variável mais próximos do valor a ser substituído, revelando que esta imputação parcial foi apenas necessária para duas embarcações entre todos os supostos infratores.

Modificações adicionais foram feitas para mitigar o risco de identificação da embarcação com base na combinação da variável GDH e das coordenadas geográficas. O horário das fiscalizações foi removido, e apenas o ano (*Year*), mês (*Month*), dia (*Day*) e período do dia (*Period*) foram mantidos. O período do dia foi dividido em quatro categorias: das 0h às 6h, das 6h às 12h, das 12h às 18h e das 18h às 24h.

Além disso, um procedimento de deslocamento geográfico foi implementado, isto é, a introdução de um deslocamento aleatório na latitude e na longitude, visto como uma perturbação aleatória. Os valores de deslocamento variavam de 0 a 0,5 milhas náuticas. O novo valor representa uma localização aleatória dentro de um retângulo de 1 milha de largura, cujo centro é a localização real. A Figura 3.1 inclui um conjunto selecionado de registos próximos à região de Sagres em Portugal para ilustrar o deslocamento geográfico final juntamente com a localização original.

Ao implementar essas medidas, o conjunto de dados é protegido ainda mais contra os possíveis riscos de identificação da embarcação, e ao mesmo tempo mantendo a integridade dos dados para fins de análise.

Consequentemente, obtemos uma base de dados constituída por 10 553 registos e 41 variáveis, sendo estas: *Latitude, Longitude, Unit, Sub_Type, Gear, Result, Infraction, CFR, Vessel_Type, Main Fishing Gear, LOA, LBP, Tonnage GT, Other Tonnage, Power of main engine, Power of auxiliary engine, Hull material, Date of entry into service, Year of construction, real_local, Local_Name, variáveis*

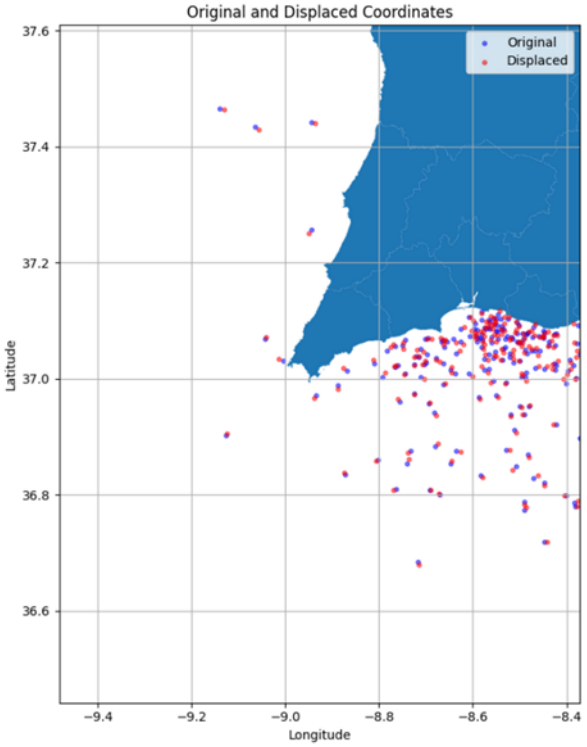

FIGURA 3.1: Coordenadas originais vs Coordenadas deslocadas

de I – XIV, number_infracs, Year, Month, Day, Period e *NUTS II_Code*. Podemos ainda observar na tabela 3.1 as variáveis que sofreram transformações e a transformação respetiva.

Variáveis transformadas	**Transformação sofrida**
GDH	Fragmentação em ano, mês, dia e período
Latitude/Longitude	Deslocamento geográfico
CFR	Atribuição de código anonimizado
Unit	Atribuição de código anonimizado
Date of entry	Arredondamento e Adição de Ruído
Year of construction	Arredondamento e Adição de Ruído
LOA	Adição de Ruído e Arredondamento
LBP	Adição de Ruído e Arredondamento
Tonnage GT	Adição de Ruído
Other Tonnage	Recodificação Global
Power of main engine	Arredondamento e Adição de Ruído
Power of auxiliary engine	Recodificação Global

TABELA 3.1: Variáveis transformadas

3.2.1 Avaliação do Risco

Depois de todas as transformações nas variáveis selecionadas, foi realizada uma avaliação de risco completa, considerando todas as variáveis da base de dados do Registo da Frota da UE disponíveis publicamente. Foi determinado que o conjunto de dados continha apenas variáveis sensíveis relacionadas à presença de presumíveis infrações e o tipo de infrações, denotando que os registos com resultado "LEGAL" são não sensíveis. Assim, a base de dados após estas transformações apresentava 8 917 registos sem infrações e 1 636 registos considerados presumíveis infrações.

Tendo isto em mente, o foco foi direcionado para as embarcações com pelo menos uma infração. Entre estas embarcações, apenas oito poderiam ser potencialmente identificadas, no entanto, mesmo que estas oito embarcações fossem identificadas, a identificação não poderia ser garantida pois seriam baseadas em variáveis que tinham sofrido uma perturbação causada pela adição de ruído, não sendo possível ter a certeza se o valor ali representado era o real ou outro valor diferente do original perturbado pela transformação imposta na variável.

Apenas as variáveis que não foram alteradas devido à sua importância e aquelas que sofreram transformações de arredondamento foram consideradas para maior refinamento da avaliação de risco. Usando apenas essas variáveis, antes da aplicação da supressão local foi possível identificar apenas duas embarcações. No entanto, deve-se destacar que estas duas embarcações já não são diretamente identificáveis devido ao processo de substituição de supressão local descrito anteriormente. Assim, os dados divulgados não possuem mais este risco. Portanto, na base de dados final nenhuma embarcação pode ser identificada exclusivamente através das bases de dados disponíveis publicamente, tornando cada embarcação indistinguível das outras em termos de variáveis-chave.

Em conclusão, o conjunto de dados final garante que nenhuma entrada individual possa identificar diretamente uma embarcação fiscalizada que se presume ter um registo de infração, sendo este um aspeto crucial para a proteção da privacidade. Posto isto, o conjunto de dados fornece um bom nível de confidencialidade e protege as identidades das embarcações fiscalizadas, garantindo a conformidade com os regulamentos e padrões de privacidade.

Capítulo 4

Avaliação da Qualidade dos dados protegidos

O presente capítulo é dedicado à apresentação e validação dos resultados obtidos após comparação entre a base de dados original e a base de dados protegida, recorrendo à utilização de *heatmaps* dos dados omissos, histogramas das variáveis originais selecionadas para sofrer alterações confrontados com os histogramas dessas mesmas variáveis após a transformação elencada no Capítulo anterior e matrizes de correlações de Pearson, de Kendall e de *Point-Biserial*.

Durante o processo de implementação de medidas de controle de divulgação na base de dados, é inevitável que haja algum comprometimento na qualidade dos dados. No entanto, o objetivo nas transformações aplicadas foi encontrar um equilíbrio entre preservar a melhor qualidade possível dos dados e garantir a proteção contra a divulgação de identificação direta.

Embora ainda haja algum risco inerente envolvido, é importante enfatizar que esse risco não é absoluto ou aplicável a todas as embarcações. O nível de risco varia dependendo das variáveis específicas e das suas respetivas transformações. Para as embarcações que nunca foram associadas a uma presumível infração, o risco de identificação direta é limitado devido à ausência de informações sensíveis relacionadas a elas.

No conjunto de dados protegido, certas variáveis não possuem valores devido à indisponibilidade de detalhes específicos nas bases de dados originais, além disso, um pequeno número de variáveis pode ter intencionalmente valores em falta como parte das medidas de controle de divulgação. Um *heatmap* apresentado na Figura 4.1 fornece uma visão geral da extensão dos dados em falta (representado a branco), concentrando-se nas variáveis com valores em falta diferentes de zero (representado a branco). É importante destacar que variáveis como *real_local, Local_Name* e

NUTSII_Code apresentam relativamente poucos registos de valores em falta, sendo que a maioria dos dados em falta restantes pode ser atribuída à ausência de registos nos dados do Registo da Frota da UE para determinadas variáveis.

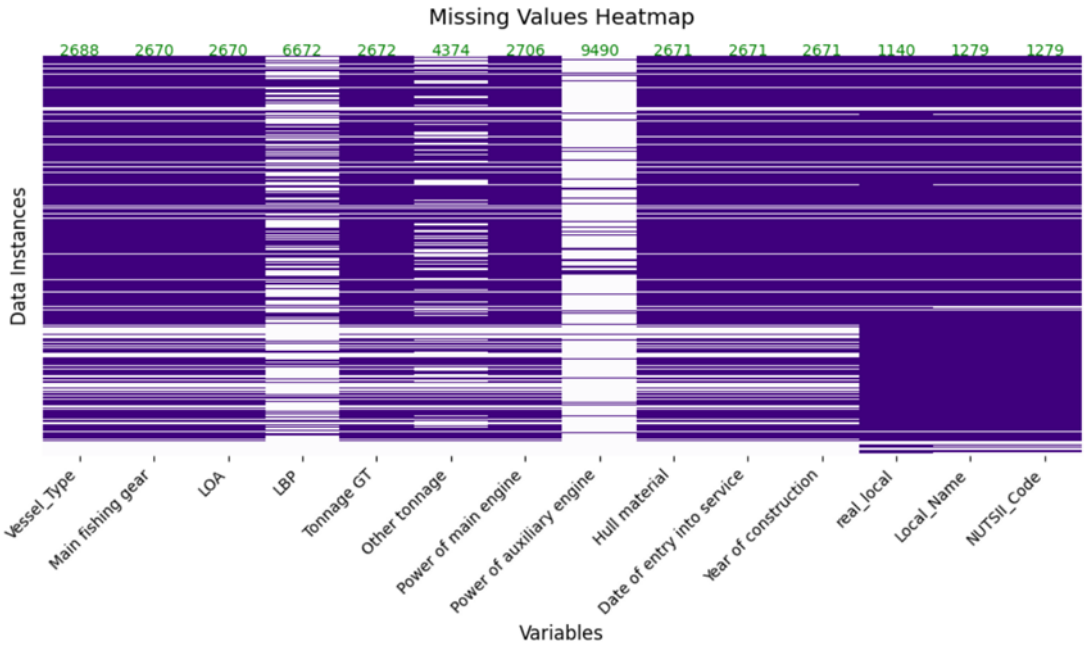

FIGURA 4.1: Heatmap dos valores omissos presentes nas variáveis da base de dados protegida

O arredondamento aplicado às variáveis pode ser classificado como um método de anonimização com efeito perturbativo ou não perturbativo, dependendo do nível de arredondamento e da amplitude dos dados transformados. No entanto, neste caso específico, a transformação de arredondamento terá um impacto mínimo na utilidade ou conteúdo informativo das variáveis. Arredondar para o número inteiro mais próximo ou para a casa decimal simplesmente ajusta os valores para um nível de precisão mais amplo sem alterar a distribuição geral ou os padrões dos dados. Portanto, após a transformação de arredondamento, não é necessário avaliar ainda mais a utilidade dessas variáveis.

O mesmo se aplica às variáveis modificadas pela supressão local. A remoção de valores específicos por meio da supressão local foi realizada para proteger informações sensíveis e reduzir o risco de identificação. No entanto, como o número de valores suprimidos é relativamente pequeno, isso não afeta significativamente a utilidade ou o conteúdo informativo dos dados.

Em relação aos valores imputados parcialmente, a quantidade é mínima e foram utilizados para substituir valores ausentes ou não divulgados no conjunto de

dados. O impacto dessas imputações nos dados é insignificante, tornando-as não perturbativas.

Algumas medidas de utilidade podem ser realizadas nas variáveis perturbadas por adição de ruído, onde todos os registos foram modificados. A análise destes dados passa primeiramente por uma análise visual e posteriormente uma análise estatística, isto é, comparando médias, variâncias e correlações. Nas figuras 4.2 e 4.3 podemos observar os histogramas, também conhecidos como distribuições de frequências, das variáveis *LOA* e *LBP*, respetivamente, comparando os dados originais e os dados protegidos.

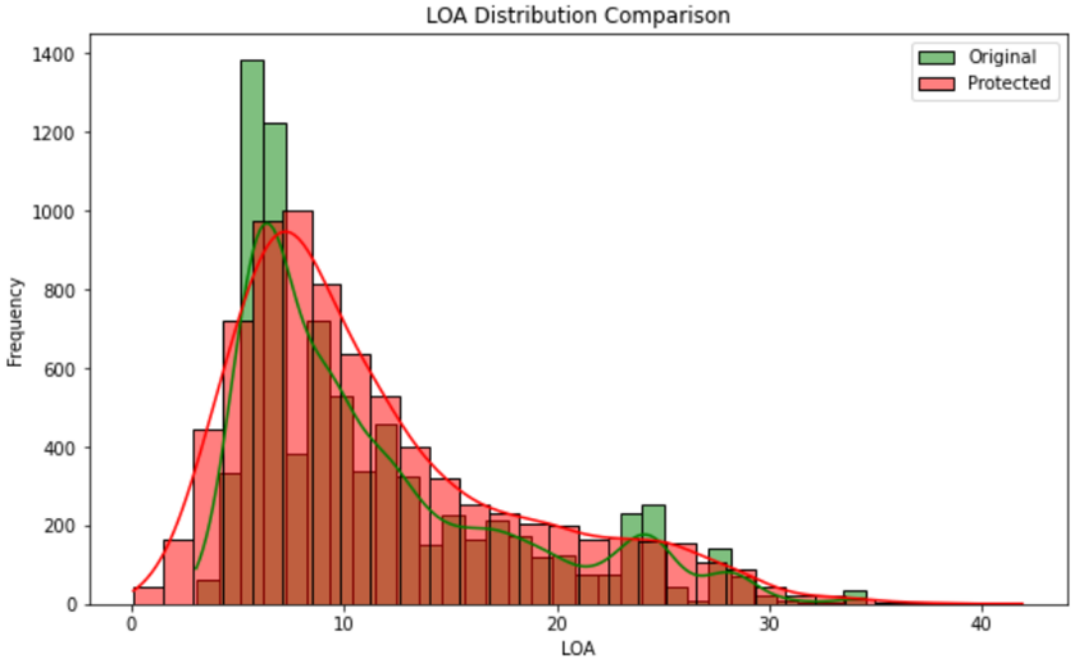

FIGURA 4.2: Histograma comparando a distribuição de LOA da base de dados original e protegida com uma estimativa de densidade por Kernel

Observando as figuras 4.2 e 4.3 percebemos que há um maior desvio dos dados protegidos para os dados originais nos valores mais extremos, diminuindo a amplitude da frequência destes dados. Com apenas o arredondamento destas variáveis, diversas embarcações apresentariam o mesmo comprimento e iriamos notar uma possível suavização do gráfico, contudo foi a adição de ruído que levou a uma suavização dos *outliers*.

Nestes gráficos, a dispersão tanto dos dados originais como protegidos exibe uma distribuição normal, sendo que considerando as características e requisitos específicos das embarcações de pesca comercial é compreensível uma maior centralização

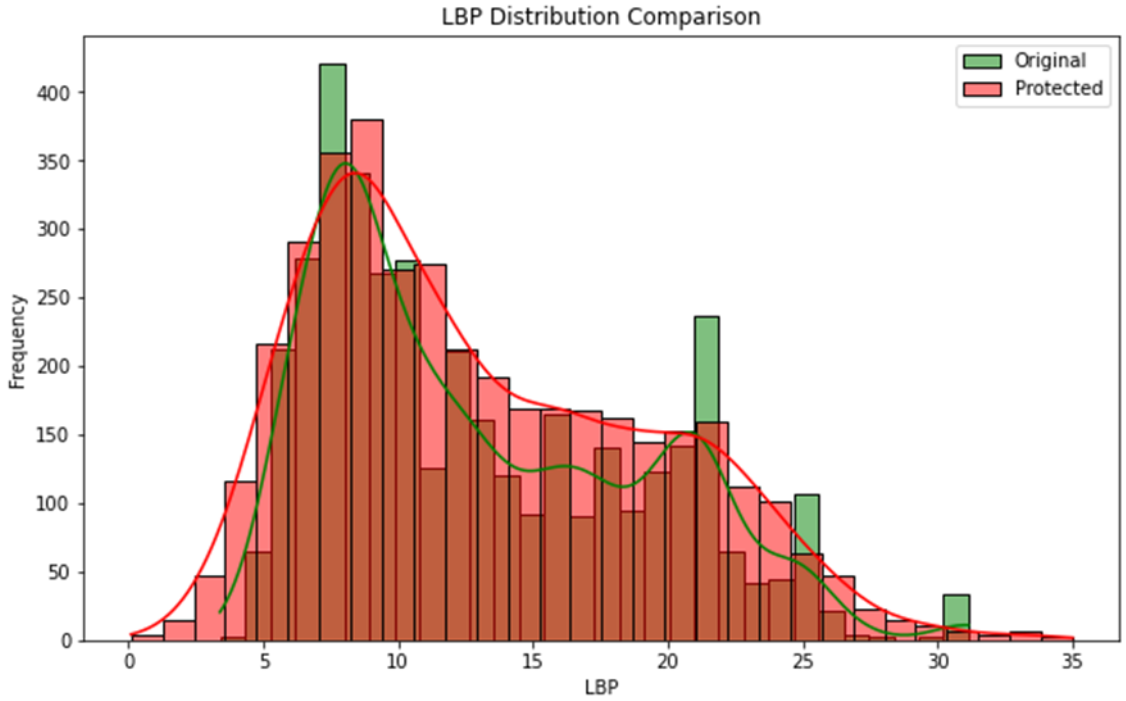

FIGURA 4.3: Histograma comparando a distribuição de LBP da base de dados original e protegida com uma estimativa de densidade por Kernel

dos comprimentos das embarcações.

Ao comparar a estimativa de densidade kernel da distribuição das duas variáveis em confronto, verifica-se que o efeito das transformações nestas variáveis é mínimo, dada a notável similaridade entre as estimativas.

Nas figuras 4.4 e 4.5 podemos observar os histogramas, também conhecidos como distribuições de frequências, das variáveis *Tonnage GT* e *Power of main engine*, respetivamente, comparando os dados originais e os dados protegidos.

A partir da análise das figuras 4.4 e 4.5, podemos observar que as variáveis perturbadas assemelham-se às originais nas suas distribuições. No entanto, a adição de ruído suavizou as variáveis perturbadas em comparação com os dados originais. Esse efeito de suavização é esperado, pois o ruído segue uma distribuição normal. A introdução desta suavização por meio da adição de ruído pode ser benéfica para melhorar a qualidade dos dados, reduzindo o impacto de valores atípicos e aprimorando a integridade dos dados. No entanto, também é importante ressaltar que isso pode levar à perda de informações detalhadas e à introdução de possíveis distorções.

Será importante notar que as variáveis *Tonnage GT* e *Power of main engine*, como ilustrado nas figuras 4.4 e 4.5, apresentam um ligeiro deslocamento para a

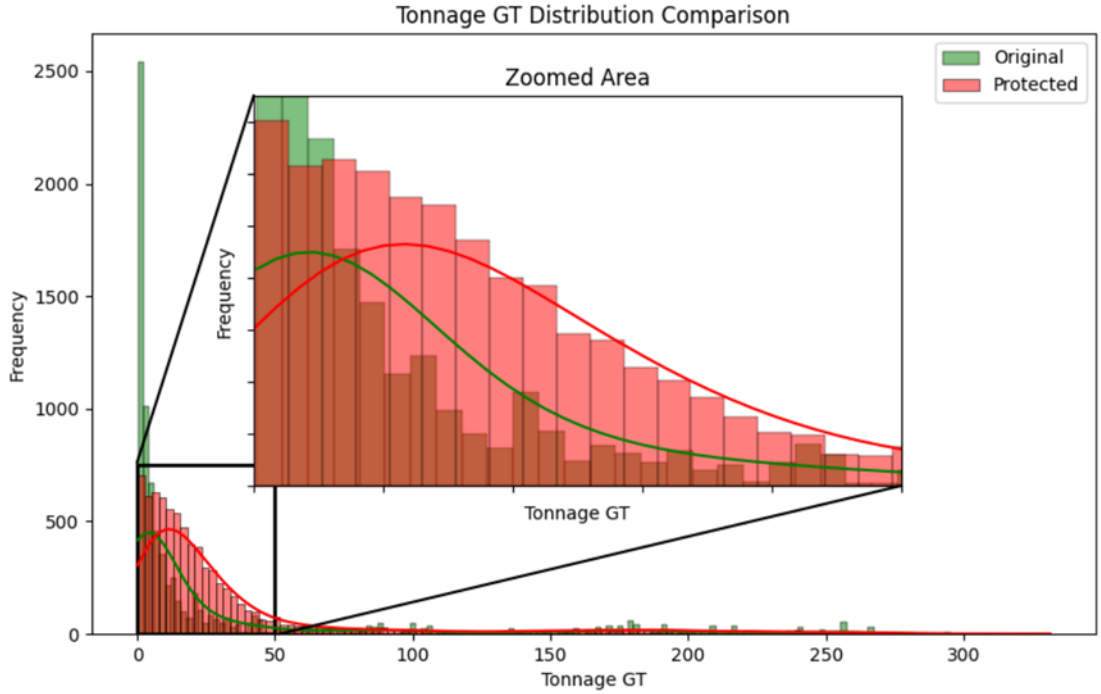

FIGURA 4.4: Histograma comparando a distribuição de Tonnage GT da base de dados original e protegida com uma estimativa de densidade por Kernel

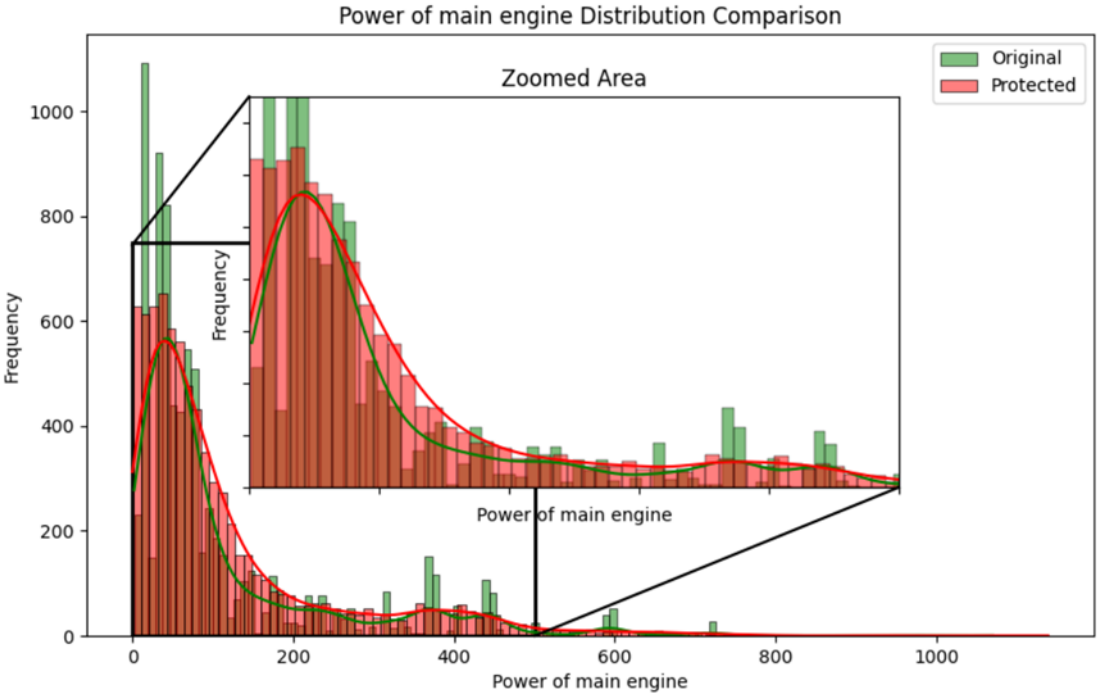

FIGURA 4.5: Histograma comparando a distribuição de Power of main engine da base de dados original e protegida com uma estimativa de densidade por Kernel

direita nas distribuições quando observadas em intervalos mais pequenos e próximos de zero. Este deslocamento era esperado, pois recodificamos os valores para não serem nulos ou negativos, atribuindo o valor mínimo registado de zero e o valor absoluto para valores negativos. No entanto, é importante destacar que este ajuste não alterou significativamente a distribuição geral das variáveis.

Podemos observar uma semelhança entre os histogramas da variável *Tonnage GT* e *Power of main engine*, pois à medida que as embarcações aumentam de tamanho, a sua arqueação bruta também tende a aumentar, refletindo o seu peso total e capacidade de carga. Da mesma forma, embarcações maiores geralmente requerem motores principais mais potentes para propulsioná-las de forma eficiente pela água e atingir velocidades desejadas.

Nas figuras 4.6 e 4.7 podemos observar os histogramas, também conhecidos como distribuições de frequências, das variáveis *Date of entry into service* e *Year of construction*, respetivamente, comparando os dados originais e os dados protegidos.

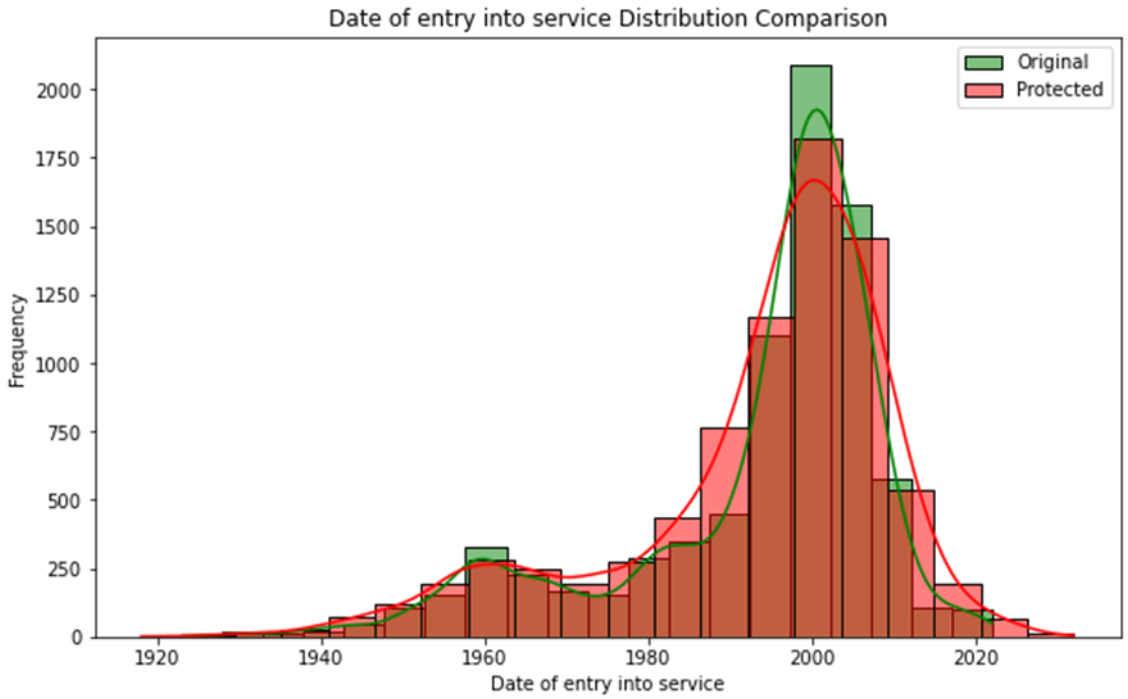

FIGURA 4.6: Histograma comparando a distribuição de Date of entry into service da base de dados original e protegida com uma estimativa de densidade por Kernel

Analisando as variáveis *Date of entry into service* e *Year of construction* compreendemos que também os seus histogramas demonstram uma distribuição normal, manifestando valores quase iguais, isto devesse a estas variáveis estarem intrinsecamente relacionadas.

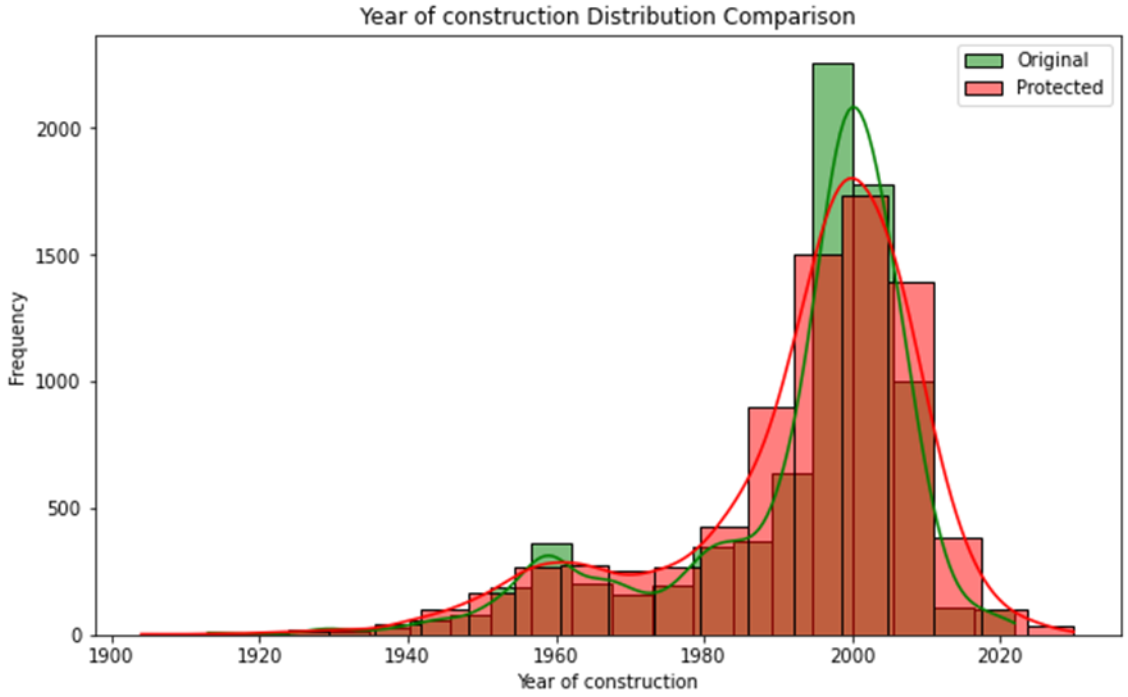

FIGURA 4.7: Histograma comparando a distribuição de Year of construction da base de dados original e protegida com uma estimativa de densidade por Kernel

Variáveis	**Média Original**	**Média Protegida**	**Variância Original**	**Variância Protegida**
LOA	11.467893	11.460244	42.572205	46.282576
LBP	13.031482	13.059856	35.808518	39.730192
Tonnage GT	29.095551	36.692501	3180.355289	2950.528324
Power of main engine	113.076253	119.412629	18284.676221	18666.126755
Date of entry into service	1993.418422	1993.402055	260.613887	285.513118
Year of construction	1992.816798	1992.797640	279.031780	305.537906

TABELA 4.1: Comparação das médias e variâncias das variáveis perturbadas calculadas a partir da base de dados original e base de dados protegida

Como mencionado no subcapítulo 1.6 uma forma de avaliar a qualidade dos dados é comparar as estatísticas das variáveis entre os conjuntos de dados protegidos e originais, tais como médias e variâncias. Esta comparação permite-nos avaliar quão bem os métodos de anonimização de dados preservaram as propriedades estatísticas dos dados originais. A tabela 4.1 apresenta a média e a variância das variáveis perturbadas calculadas a partir dos conjuntos de dados originais e protegidos.

Através de uma análise das estatísticas apresentadas na tabela 4.1, torna-se evidente que as variáveis que apresentam as maiores disparidades também possuem variâncias maiores, uma vez que a perturbação introduzida pela adição de ruído é responsável por estas variações e é proporcional à variância de cada variável. No entanto, é importante enfatizar que as diferenças gerais entre as medidas estatísticas dos conjuntos de dados originais e de dados protegidos permanecem relativamente pequenas ao considerar a distância normalizada entre os dois conjuntos de dados, expressa como:

$$d_i = \frac{\bar{x_i} - \bar{x_i^*}}{\sigma_i} \tag{4.1}$$

onde $\bar{x_i}$ e $\bar{x_i^*}$ representam as médias das variáveis originais e protegidas, respetivamente, e σ_i é o desvio padrão da variável original. Os valores das distâncias normalizadas foram:

- *LOA* – 0.001172
- *LBP* – 0.004742
- *Tonnage GT* – 0.134710
- *Power of main engine* – 0.046859
- *Date of entry into service* – 0.001014
- *Year of construction* – 0.001147

Note-se que todos os valores estão muito próximos de zero, indicando que os métodos de anonimização de dados preservaram com relativo sucesso as características essenciais dos dados originais e ao mesmo tempo protegeram a privacidade das embarcações.

Além de analisar as médias, variâncias e a estimativa de densidade por Kernel, outra medida relevante a ser considerada é a alteração da correlação entre estas variáveis e o número de infrações registadas. A correlação de Pearson, que avalia a relação linear entre as variáveis, fornece informações sobre esta mudança, conforme apresentado na tabela 4.2. No entanto, considerando que as variáveis perturbadas podem conter *outliers*, também é crucial considerar a correlação de Kendall.

Ao examinar os valores de correlação apresentados na tabela 4.2, é evidente que os conjuntos de dados original e protegido apresentam padrões de correlação comparáveis entre as variáveis numéricas e a variável categórica *number_infracs*,

Variáveis	**Número de Infrações**			
	Pearson		**Kendall**	
	Dados Originais	**Dados Protegidos**	**Dados Originais**	**Dados Protegidos**
LOA	0.183503	0.175143	0.174021	0.158154
LBP	0.110379	0.098568	0.118285	0.103662
Tonnage GT	0.120172	0.105599	0.178250	0.104214
Power of main engine	0.154544	0.143508	0.173864	0.141693
Date of entry into service	-0.006641	-0.006642	-0.016354	-0.018762
Year of construction	-0.003901	-0.001887	-0.019317	-0.019739

TABELA 4.2: Comparação entre a correlação de Pearson e de Kendall, referente ao número de infrações detetadas e as variáveis transformadas dos dados originais e protegidos

considerando as diferenças nos valores de correlação entre os dois conjuntos de dados insignificantes. Portanto, em termos de correlação, os conjuntos de dados original e protegido fornecem informações idênticas e não apresentam distinções relevantes. A incorporação da correlação de Kendall permitiu obter uma compreensão mais aprofundada da associação entre as variáveis perturbadas e o número de infrações, tanto antes quanto depois do processo de proteção de dados.

Além disso, podemos examinar a correlação entre as variáveis perturbadas e a ocorrência de cada uma das catorze infrações, que é a variável sensível neste conjunto de dados. Por esta razão, as matrizes de coeficientes de correlação de Pearson e Kendall são apresentadas nas próximas figuras.

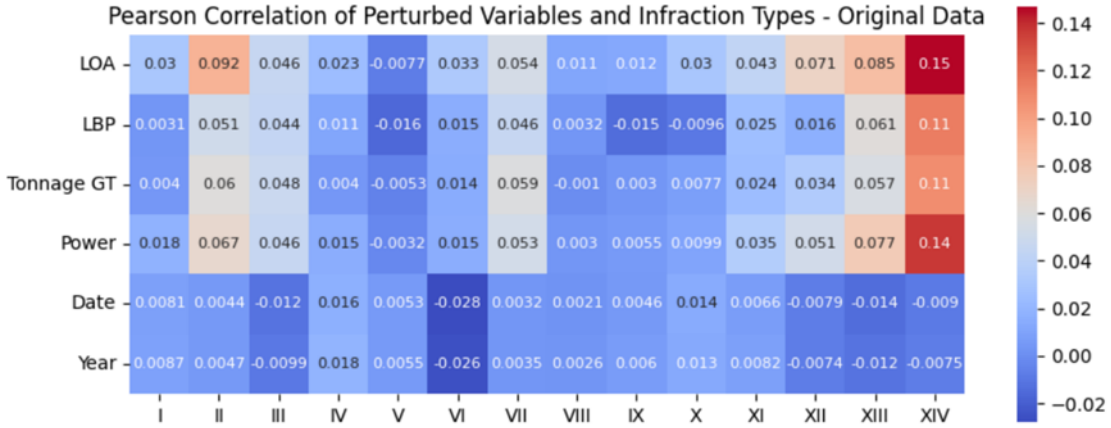

FIGURA 4.8: Matriz de correlação de Pearson entre as variáveis perturbadas e o tipo de infrações na base de dados original

Em relação à correlação de Pearson entre as variáveis perturbadas e os tipos de infrações, observamos padrões de correlação semelhantes tanto nos Dados Originais quanto nos Dados Protegidos. Os coeficientes de correlação indicam a

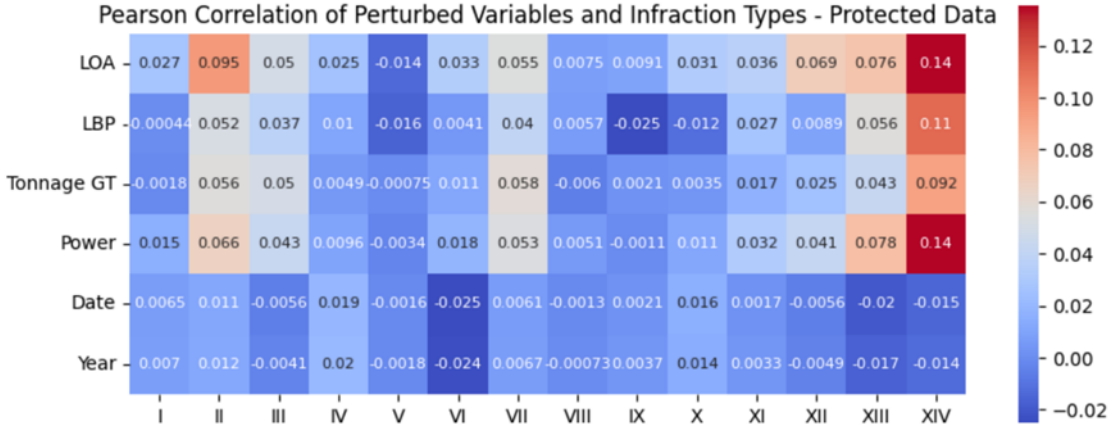

FIGURA 4.9: Matriz de correlação de Pearson entre as variáveis perturbadas e o tipo de infrações na base de dados protegida

intensidade e direção da relação linear entre as variáveis. No entanto, é importante observar que os Dados Protegidos apresentam uma magnitude de correlação ligeiramente menor em comparação aos Dados Originais. Essa diferença pode ser atribuída aos métodos de perturbação aplicados para proteger os dados, enquanto preservam a sua utilidade.

Além disso, é possível notar uma correlação mais elevada entre a infração XIV e as variáveis perturbadas, como mencionado anteriormente, a infração XIV, que abrange uma variedade de situações como falta de documentos a bordo, falta de pirotécnicos, falta de meios de salvação, extintores caducados, entre outros, é caracterizada por sua natureza abrangente. Portanto, não é surpreendente que esta infração apresente uma correlação mais forte com as variáveis em questão.

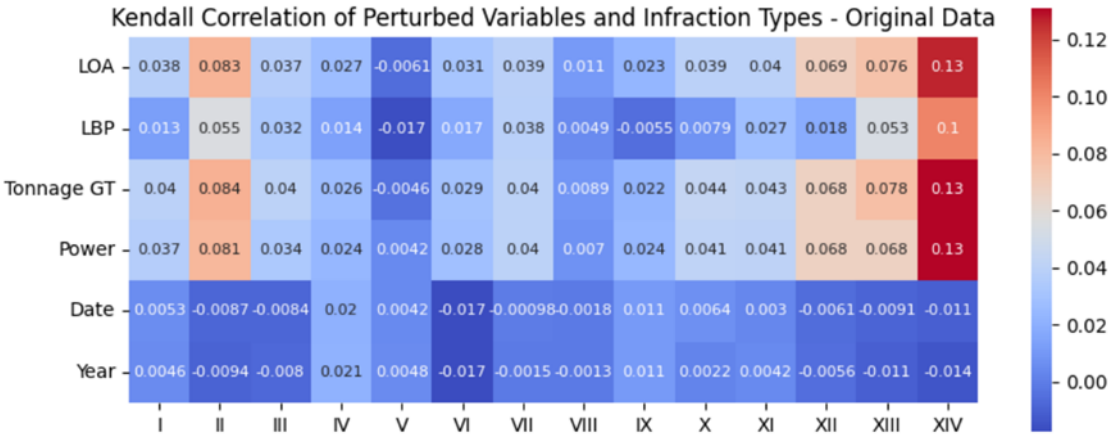

FIGURA 4.10: Matriz de correlação de Kendall entre as variáveis perturbadas e o tipo de infrações na base de dados original

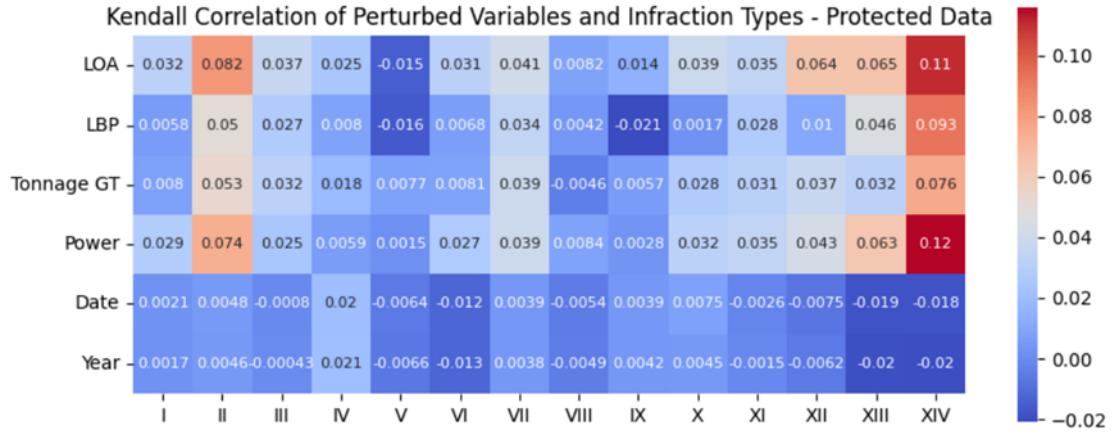

FIGURA 4.11: Matriz de correlação de Kendall entre as variáveis perturbadas e o tipo de infrações na base de dados protegida

No que diz respeito à correlação de Kendall, esta avalia a força e direção da associação entre variáveis, tendo em consideração as relações não lineares e a presença de *outliers*.

Sendo esta correlação considerada menos sensível a valores extremos ou *outliers* nos dados em comparação com algumas outras medidas de correlação, como a correlação de Pearson. Isso acontece por causa do método de cáclulo da correlação de Kendall e por causa do seu foco na ordem dos dados, em vez dos seus valores absolutos. Assim, ao analisar a correlação de Kendall entre as variáveis perturbadas e os tipos de infrações, podemos observar padrões consistentes tanto nos Dados Originais quanto nos Dados Protegidos. Sendo que, os coeficientes de correlação de Kendall fornecem uma melhor perceção sobre as relações entre as variáveis, sendo que apresenta grandes similaridades à correlação de Pearson.

Após analisar as figuras, como evidenciado as matrizes de correlação apresentam valores semelhantes no geral. No entanto, é crucial enfatizar que, neste caso, o objetivo é quantificar a relação entre uma variável contínua e uma variável binária.

A correlação Point-Biserial é especificamente projetada para este propósito, permitindo-nos capturar e avaliar efetivamente o grau de associação entre as variáveis perturbadas e a ocorrência de diferentes infrações. As figuras 4.12 e 4.13 ilustram a correlação Point-Biserial entre as variáveis perturbadas e a ocorrência de várias infrações. Estas análises fornecem uma representação clara da força e direção da associação entre as variáveis perturbadas contínuas e os tipos de infração binários, e mais uma vez, os valores mostram diferenças mínimas.

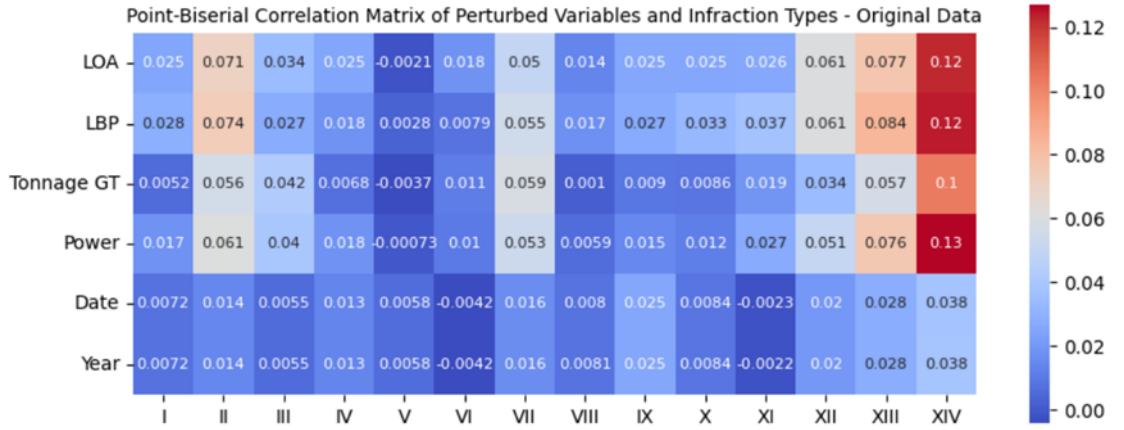

FIGURA 4.12: Matriz de correlação de Point-Biserial entre as variáveis perturbadas e o tipo de infrações na base de dados original

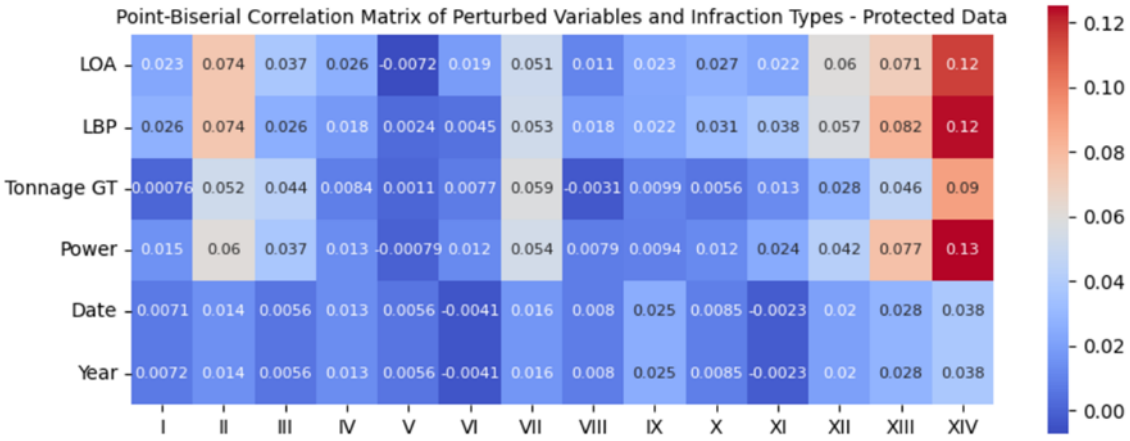

FIGURA 4.13: Matriz de correlação de Point-Biserial entre as variáveis perturbadas e o tipo de infrações na base de dados protegida

No geral, a avaliação da qualidade dos dados demonstrou que os dados divulgados mantiveram a sua utilidade para fins de análise e pesquisa. As técnicas de anonimização e transformações empregadas garantiram de forma eficaz a proteção da identificação das embarcações, enquanto preservaram a maior parte da integridade e utilidade dos dados. Compreendemos assim, que ao utilizar tais técnicas, os pesquisadores ainda podem extrair informações valiosas e chegar a conclusões válidas a partir dos dados perturbados, ao mesmo tempo em que mantêm a proteção da privacidade necessária.

Para melhor compreensão do código por trás do pré-processamento e das transformações das variáveis é possível obter o mesmo indo ao website https://github.com/ricardomourarpm/Fishery_Inspection_PT_2017_23. Contudo, como podemos observar na figura 4.14, certas informações foram excluídas do código público.

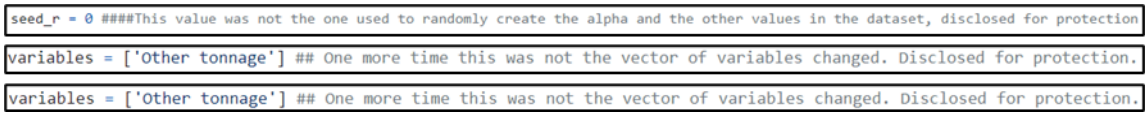

FIGURA 4.14: Exemplos de partes do código não completas, por motivos de privacidade

Esta seleção de informação foi necessária pois a partilha do código daria possibilidade de utilizar um processo inverso para obter as variáveis originais e eventualmente descobrir a identidade de várias embarcações de pesca.

Conclusão

Com base nas evidências e resultados obtidos durante esta dissertação de mestrado, é possível concluir que o tema do CDE desempenha um papel fundamental na proteção da confidencialidade dos dados. Ao estudar e comparar os diversos métodos de CDE, foi possível constatar a sua relevância na transformação de dados sensíveis em dados protegidos, sem comprometer a utilidade das informações. Além disso, a análise das propriedades estatísticas dos dados reais e protegidos revelou uma similaridade significativa, o que evidencia a eficácia dos métodos utilizados.

Esta dissertação de mestrado é apenas a ponta do iceberg no que toca CDE, a base de dados utilizada não tinha dados realmente confidenciais, mas sim sensíveis, contudo, podemos concluir que o conjunto de dados resultante do processo de anonimização e perturbação manteve a privacidade das embarcações fiscalizadas, atendendo aos regulamentos e padrões de privacidade. As técnicas empregadas garantiram a proteção adequada da identidade das embarcações, tornando-as indistinguíveis umas das outras. Ao mesmo tempo, os dados preservaram a sua utilidade para análise e pesquisa, permitindo a extração de informações valiosas e conclusões válidas. A combinação de medidas de anonimização, como remoção de identificadores diretos, arredondamento, recodificação e adição de ruído, contribuiu para garantir a confidencialidade e a proteção da privacidade das embarcações, ao mesmo tempo em que manteve a qualidade e a integridade dos dados. Estas técnicas demonstraram ser eficazes na redução do risco de reidentificação, garantindo assim a conformidade com os padrões de privacidade estabelecidos.

No contexto da Marinha Portuguesa, onde a segurança e a proteção dos dados são de extrema importância, esta dissertação demonstrou que existem métodos mais eficazes do que os utilizados atualmente. Embora a Marinha adote medidas como a atribuição de códigos anonimizados ou a eliminação de dados considerados sensíveis, os resultados obtidos neste estudo confirmaram que existem abordagens mais efetivas para garantir a utilidade dos dados, ao mesmo tempo em que reduzem significativamente o risco de identificação de variáveis sensíveis.

É crucial destacar que a evolução tecnológica e a crescente necessidade de

compartilhar informações exigem uma abordagem mais sofisticada e robusta no que diz respeito à proteção de dados. Os métodos de CDE analisados nesta dissertação fornecem um caminho promissor para garantir a confidencialidade dos dados, permitindo a continuidade da partilha de informações com segurança.

No entanto, é importante ressaltar que esta pesquisa representa apenas um passo inicial nessa área de estudo e que há espaço para futuros estudos e aprimoramentos. Novas abordagens e técnicas podem ser exploradas para aprimorar ainda mais a proteção de dados sensíveis na Marinha Portuguesa e em outras instituições. Além disso, a colaboração entre académicos, profissionais e especialistas em segurança de dados pode contribuir para o desenvolvimento de estratégias mais avançadas e eficazes de CDE.

Com este trabalho foi possível identificar algumas limitações e desafios que podem ser abordados em pesquisas futuras. Embora o estudo tenha utilizado uma base de dados real, é importante ressaltar que ela foi limitada a um contexto específico da Marinha Portuguesa com dados sensíveis e não dados confidenciais. Seria interessante explorar a aplicabilidade dos métodos de CDE em outros tipos de dados, para entender melhor sua eficácia e adaptabilidade em diferentes contextos, por exemplo, a utilização do método PRAM que induz o intruso a classificar erradamente os dados ou a distorção de dados pela probabilidade de distribuição por ser um método que gera dados sintéticos e distorce os dados de forma controlada.

Além disso, o estudo concentrou-se em quatro métodos de CDE específicos: adição de ruído, arredondamento, supressão local e recodificação global. No entanto, existem outros métodos e técnicas disponíveis na literatura que também podem ser explorados. Seria valioso ampliar a análise e a comparação desses métodos adicionais, para determinar sua eficiência e identificar qual deles é mais adequado para proteger a confidencialidade dos dados em diferentes situações.

Em suma, esta dissertação de mestrado contribuiu significativamente para o conhecimento no campo do CDE e para o avanço das estratégias de proteção de dados na Marinha Portuguesa. Os resultados obtidos reforçam a importância de abordagens mais sofisticadas para garantir a confidencialidade dos dados, protegendo tanto a segurança nacional como os indivíduos envolvidos. Espera-se que este estudo seja um ponto de partida para futuras investigações e para o aprimoramento contínuo da proteção de dados sensíveis no contexto militar.

Bibliografia

Alén-Savikko, A., & Pitkänen, O. (2016). Data Protection, Privacy and European Regulation in the Digital Age. *Unigrafia OY, Helsinki*.

Benschop, T., Machingauta, C., & Welch, M. (2022). Statistical disclosure control: a practice guide.

Brand, R. (2002). Microdata protection through noise addition. *Inference Control in Statistical Databases: From Theory to Practice*, 97–116.

Chok, N. S. (2010). *Pearson's versus Spearman's and Kendall's correlation coefficients for continuous data* (tese de doutoramento). University of Pittsburgh.

Comando Naval. (2020). *Relatos e Comunicados Operacionais*.

De Waal, A., & Willenborg, L. (1999). Information loss through global recoding and local suppression. *Netherlands Official Statistics*, *14*, 17–20.

Domingo-Ferrer, J., & Torra, V. (2001). Disclosure control methods and information loss for microdata. *Confidentiality, disclosure, and data access: theory and practical applications for statistical agencies*, 91–110.

Domingo-Ferrer, J., & Torra, V. (2008). A critique of k-anonymity and some of its enhancements. *2008 Third International Conference on Availability, Reliability and Security*, 990–993.

Duncan, G. T., Fienberg, S. E., Krishnan, R., Padman, R., & Roehrig, S. F. (2001). Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies.

Dupriez, O., & Boyko, E. (2010). *Dissemination of microdata files: principles procedures and practices*. International Household Survey Network.

Fitzgerald, B. (2001). Alexander Graham Bell: the BU years. *B.U. Bridge*, *5*. https://www.bu.edu/bridge/archive/2001/09-14/bell.html

Flossmann, A., & Lechner, S. (2006). Combining blanking and noise addition as a data disclosure limitation method. *Privacy in Statistical Databases: CENEX-SDC Project International Conference, PSD 2006, Rome, Italy, December 13-15, 2006. Proceedings*, 152–163.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J., Nordholt, E. S., Seri, G., & De Wolf, P.-P. (2010). *Handbook on statistical disclosure control*. Citeseer.

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Nordholt, E. S., Spicer, K., & De Wolf, P.-P. (2012). *Statistical disclosure control* (Vol. 2). Wiley New York.

Instituto Nacional de Estatística. (2019). Política de Confidencialidade Estatística. https://www.ine.pt/xportal/xmain?xpid=INE&xpgid=ine_inst_confestatistica&xlang=pt

Karr, A. F., Sanil, A. P., & Banks, D. L. (2005). Data quality: A statistical perspective. *Statistical Methodology*, *3*(2), 137–173.

Kim, J. (1986). A Method for Limiting Disclosure in Microdata based on Radom Noise and Transformation.

Kornbrot, D. (2014). Point biserial correlation. *Wiley StatsRef: Statistics Reference Online*.

Lambert, D. (1993). Measures of Disclosure Risks and Harm. *Journal of Official Statistics*, *9*(2), 313–331.

Lei nº58/2019. (2019). Diário da República n.º 151/2019, Série I de 2019-08-08, páginas 3-40. https://diariodarepublica.pt/dr/detalhe/lei/58-2019-123815982

Liew, C. K., Choi, U. J., & Liew, C. J. (1985). A data distortion by probability distribution. *ACM Transactions on Database Systems (TODS)*, *10*(3), 395–411.

Mendes, E. C. P. (2010). Confidencialidade de Dados: Aplicação e Comparação de Técnicas de Controlo da Divulgação Estatística.

Moura, R., Santos, N. P., Vala, A., Mendes, L., Castro Neto, M., Simões, P., & Lobo, V. (2023). Fisheries Inspection in Portuguese Waters from 2015 to 2023. *Scientific Data (submetido)*.

NCSS. (s.d.). Point-Biserial and Biserial Correlations. https://www.ncss.com/wp-content/themes/ncss/pdf/Procedures/NCSS/Point-Biserial_and_Biserial_Correlations.pdf

Newbold, P., Carlson, W. L., & Thorne, B. M. (2022). *Statistics for business and economics* (10th). Pearson.

Pinto, R. d. C. M. (2017). *Análise de dados da fiscalização da pesca – Estudo da evolução e exploração dos recursos piscícolas no ecossistema marinho português* (tese de doutoramento). Escola Naval, Marinha Portuguesa.

Rubin, D. B. (1987). *Multiple imputation for nonresponse in surveys* (Vol. 81). John Wiley & Sons.

Rubin, D. B. (1993). Statistical disclosure limitation. *Journal of official Statistics*, *9*(2), 461–468.

Sullivan, G. R. (1989). *The use of added error to avoid disclosure in microdata releases*. Iowa State University.

Templ, M., Meindl, B., Kowarik, A., & Chen, S. (2014). Introduction to statistical disclosure control (sdc). *Project: Relative to the testing of SDC algorithms and provision of practical SDC, data analysis OG*.

Willenborg, L., & De Waal, T. (1996). *Statistical disclosure control in practice* (Vol. 111). Springer Science & Business Media.

Winkler, W. E. (2005). Re-identification methods for evaluating the confidentiality of analytically valid microdata. *Statistics*, *9*, 1–14.

Apêndice A - Variáveis da Base de Dados

Variáveis	Tipo de Variável
Reg_number	Categórica
Latitude	Contínua
Longitude	Contínua
GDH	Categórica
IRCS	Categórica
Licence indicator	Categórica
VMS	Categórica
ERS	Categórica
AIS	Categórica
MMSI	Categórica
Unit	Categórica
Sub_Type	Categórica
Gear	Categórica
Result	Categórica
Infraction	Categórica
CFR	Categórica
Vessel_Type	Categórica
Main fishing gear	Categórica
Subsidiary fishing gears 1-5	Categórica
LOA	Contínua
LBP	Contínua
Tonnage GT	Contínua
Other tonnage	Contínua
Power of main engine	Contínua
Power of auxiliary engine	Contínua
GTs	Contínua
Hull material	Categórica
Date of entry into service	Contínua
Year of construction	Contínua
Real_local	Categórica
Local_Name	Categórica
I – XIV	Categórica
Number_infracs	Contínua
NUTS II_Code	Categórica