Universidade do Minho
Escola de Ciências

Jorge Morais

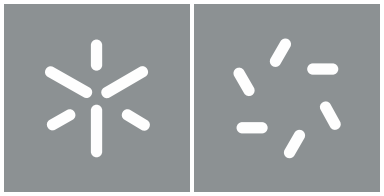
Comparação de Métodos Perturbativos: utilidade e perda de informação em bases de microdados

outubro de 2022

Comparação de Métodos Perturbativos: utilidade e perda de informação em bases de microdados

Jorge Morais

UMinho | 2022



Universidade do Minho
Escola de Ciências

Jorge Morais

Comparação de Métodos Perturbativos: utilidade e perda de informação em bases de microdados

Dissertação de Mestrado
Estatística para Ciência de Dados

Trabalho efetuado sob a orientação da
Professora Doutora Susana Faria
Doutora Rita Sousa

outubro de 2022

Despacho RT - 31 /2019 - Anexo 3

Declaração a incluir na Tese de Doutoramento (ou equivalente) ou no trabalho de Mestrado

DIREITOS DE AUTOR E CONDIÇÕES DE UTILIZAÇÃO DO TRABALHO POR TERCEIROS

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

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

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

Licença concedida aos utilizadores deste trabalho

Atribuição
CC BY

https://creativecommons.org/licenses/by/4.0/

[Esta licença permite que outros distribuam, remixem, adaptem e criem a partir do seu trabalho, mesmo para fins comerciais, desde que lhe atribuam o devido crédito pela criação original. É a licença mais flexível de todas as licenças disponíveis. É recomendada para maximizar a disseminação e uso dos materiais licenciados.]

Agradecimentos

A realização desta dissertação não teria sido possível sem o apoio de certas pessoas que tornaram possível o meu percurso académico. Assim, através de algumas palavras, pretendo agradecer às pessoas que me suportaram e ajudaram durante esta jornada da minha vida.

Em primeiro lugar, gostaria de agradecer às minhas orientadoras, a Professora Susana Faria e a Doutora Rita Sousa. Estou grato pelos seus valiosos conselhos e orientações, mas sobretudo pela disponibilidade e dedicação. Para mim foi um enorme privilégio poder trabalhar com a Professora Susana Faria e com a Doutora Rita Sousa e seria um prazer poder continuar a colaborar com ambas. A vocês, o meu muito sincero e honesto obrigado pela experiência transmitida e pela oportunidade.

Agradeço à minha família, por todo o amor dado e por me fornecerem condições para realizar o mestrado nestes últimos dois anos, pois sem eles não seria possível.

Agradeço à minha companheira, Diana, por ter estado sempre presente durante todo o meu percurso académico. Obrigado pela tua paciência, compreensão, dedicação, pela tua força e pelo teu amor. E ainda, obrigado por seres a pessoa que mais me incentiva e apoia a lutar pelos meus sonhos, sem o teu apoio não teria conseguido.

Por fim, não posso deixar de agradecer aos meus amigos o apoio que me deram durante os últimos anos.

DECLARAÇÃO DE INTEGRIDADE

Declaro ter atuado com integridade na elaboração do presente trabalho académico e confirmo que não recorri à prática de plágio nem a qualquer forma de utilização indevida ou falsificação de informações ou resultados em nenhuma das etapas conducente à sua elaboração. Mais declaro que conheço e que respeitei o Código de Conduta Ética da Universidade do Minho.

Guimarães, outubro de 2022

Jorge Rogério Silva Morais

(Jorge Rogério Silva Morais)

Resumo

Comparação de Métodos Perturbativos: utilidade e perda de informação em base de microdados

A procura por informação de alta qualidade por parte dos investigadores e do público em geral vem crescendo rapidamente nos últimos anos. Nesse sentido, é importante estabelecer um compromisso entre a disponibilização de informação estatística de qualidade e o cumprimento da legislação de proteção de dados. Técnicas de Controlo de Divulgação Estatística (CDE) sugerem métodos capazes de modificar dados sem revelar informação confidencial que possa ser vinculada a indivíduos específicos.

Este projeto pretende descrever e aplicar os vários métodos perturbativos de CDE, demonstrando os passos a efetuar de forma a que seja possível a perturbação dos dados e ainda comparar os diferentes métodos de CDE avaliando a sua utilidade face à perda de informação e face ao risco de identificação.

Numa fase inicial descrevem-se os diversos métodos de CDE apresentando-se as vantagens e desvantagens para cada um dos métodos, concluindo-se que a nível teórico o modelo *Exact General Additive Data Perturbation* (EGADP) e o modelo *Data Shuffling* produzem o menor risco de identificação e a maior utilidade nos dados. Para além da descrição dos métodos apresentam-se diversas medidas para o cálculo do risco de identificação e para a perda de informação.

Utilizando a linguagem de programação R aplicam-se os métodos numa base de microdados fornecida pelo Laboratório de Investigação em Microdados do Banco de Portugal (BPLIM). Para além da aplicação dos métodos descreve-se o *package* sdcMicro em R, que é essencial na aplicação dos métodos de CDE.

Com a aplicação a uma base de microdados real os resultados obtidos permitem concluir que a escolha do método pode variar consoante o objetivo do responsável da base de microdados. Neste caso, os métodos que apresentaram resultados mais favoráveis foram os modelos Aditivos de Ruído.

Assim, na aplicação da base de microdados PT2020 fornecida pelo BPLIM conclui-se que:

- Caso o objetivo do responsável seja obter o melhor compromisso entre a perda de informação e o risco de identificação, então a escolhe deve ser o modelo Aditivo de Ruído Correlacionado;
- Caso o objetivo do responsável seja obter a menor perda de informação e um risco de identificação não muito elevado, então a escolha deve ser o modelo Aditivo de Ruído Independente;
- Caso o objetivo do responsável seja obter o menor risco de identificação, independentemente da perda de informação, então a escolha deve ser o modelo *Exact General Additive Data Perturbation* (EGADP).

Conclusões que contrariam em parte a literatura, no entanto, ao longo da dissertação é referido por várias vezes que a escolha do melhor método dependerá muito do objetivo do responsável da base de microdados e que não é possível referir apenas um método capaz de satisfazer os diversos objetivos dos diferentes responsáveis.

Nesta dissertação faz-se também uma abordagem aos métodos perturbativos com dados longitudinais, no entanto esta ainda é uma área muito primitiva que precisa de ser mais desenvolvida, tanto a nível teórico como prático.

Palavras Chave: Controlo de Divulgação Estatística (CDE), Perturbação dos dados, *Package* sdcMicro, Risco de Identificação, Utilidade dos Dados

Abstract

Comparison of Perturbation Methods: utility and information loss in microdata

The demand for high-quality information from researchers and the public, in general, has been growing rapidly in recent years. In that sense, it is essential to establish a compromise between the availability of quality statistical information and compliance with data protection legislation. Statistical Disclosure Control (SDC) techniques suggest methods to modify data so that they can be published without revealing confidential information that can be linked to specific respondents.

This project aims to describe and apply the various perturbation methods of SDC, showing the steps to be taken to make it possible to perturb the data and also comparing the different methods of SDC, evaluating their data utility and disclosure risk.

Initially, the different SDC methods are described, presenting the advantages and disadvantages for each one of the methods, concluding that at a theoretical level the *Exact General Additive Data Perturbation* (EGADP) model and the *Data Shuffling* present the lowest disclosure risk and the highest data utility. In addition to the description of the methods, several measures are presented for calculating the disclosure risk and information loss.

Using the R programming language, the methods are applied in a microdata base provided by BPLIM. In addition to the application of the methods, the *package* sdcMicro in R is described, which is essential in the application of CDE methods.

With the results obtained in the application to a real data set, it is clear that the method choice depends on the goals of the person responsible for the microdata base. In this case, the methods that presented the most desirable results were the noise additive models.

Thus, in the application to a real dataset provided by BPLIM, it is concluded that:

- If the responsible person's objective is to obtain the best compromise between the information lost and the disclosure risk, then the choice should be the Additive Correlated Noise model;
- If the responsible person's objective is to obtain the least loss of information and a not very high disclosure risk, then the choice should be the Independent Noise Additive model;
- If the responsible person's objective is to obtain the lowest disclosure risk, regardless of the information lost, then the choice should be the *Exact General Additive Data Perturbation* (EGADP) model.

Conclusions that partially contradict the literature, however, throughout this paper it is mentioned several times that the choice of the best method depend a lot on the objective of the person responsible for the microdata base and that it is not possible to mention only one method capable of satisfying the various objectives of the differents responsibles.

In this dissertation, an approach is also made to perturbative methods with longitudinal data, however, this is still a very primitive area that needs to be further developed, both theoretically and practically.

Keywords: Statistical Disclosure Control (SDC), Data Perturbation, Disclosure Risk, Data Utility, Package sdcMicro

Conteúdo

1 Introdução **1**
1.1 BPLIM - Banco de Portugal . . . 1
1.2 Objetivos . . . 2
1.3 Estrutura do Relatório . . . 3

2 Conceitos Fundamentais **4**
2.1 Classificação de Variáveis . . . 4
2.2 Tipos de Identificação . . . 4
2.3 Risco de Identificação e Perda de Informação . . . 5
2.4 Processo de Proteção . . . 6

3 Medidas de Risco de Identificação **8**
3.1 Medidas de Risco para Variáveis Categóricas . . . 8
3.2 Medidas de Risco para Variáveis Numéricas . . . 11
3.3 Risco de Identificação Individual . . . 12
3.4 Risco de Identificação Global . . . 12

4 Métodos de Controlo de Divulgação Estatística **14**
4.1 Métodos Não Perturbativos . . . 14
4.2 Métodos Perturbativos . . . 15
4.2.1 Métodos Perturbativos para Variáveis Categóricas . . . 16
4.2.2 Métodos Perturbativos para Variáveis Numéricas . . . 17
4.2.3 O Paradigma de Permutação . . . 27
4.3 Perturbação de Dados Longitudinais . . . 28
4.3.1 Mapeamento Inverso de Dados Longitudinais . . . 29
4.3.2 Perda de Informação e Risco de Identificação . . . 31
4.4 Geração de Dados Sintéticos . . . 32
4.5 Comparação dos Métodos de CDE . . . 33
4.5.1 Comparação de Modelos Lineares e Não Lineares de Ruído . . . 33
4.5.2 Comparação de Modelos com Métodos Perturbativos . . . 34
4.5.3 PRAM, *Rank Swapping* e *Shuffling* . . . 35
4.6 Conclusão . . . 35

5 Perda de Informação e Utilidade dos Dados **36**
5.1 Medidas de Perda de Informação para Variáveis Categóricas . . . 36
5.2 Medidas de Perda de Informação para Variáveis Numéricas . . . 39

6 Ambiente R **42**
6.1 Ferramentas para CDE . . . 42
6.2 *Package* sdcMicro . . . 43
6.2.1 Objeto CDE . . . 44
6.2.2 Medição de Risco de Identificação . . . 46

6.2.3 Aplicação de Métodos de CDE 49
6.2.4 Avaliação de Perda de Informação 54
6.2.5 Extração de Resultados 55
6.3 Exemplo EUSILCS 56

7 Caso de Estudo 62
7.1 Descrição da Base de Microdados 62
7.2 Escolha das variáveis chave 71
7.3 Avaliação do Risco de Identificação 73
7.4 Aplicação dos Métodos Perturbativos em Variáveis Categóricas 74
7.4.1 Supressão Local 74
7.4.2 PRAM 77
7.4.3 Avaliação dos Métodos 80
7.5 Métodos para Variáveis Numéricas 82
7.5.1 Modelos Lineares de Ruído 82
7.5.2 Modelos Não Lineares de Ruído 85
7.5.3 Outros Métodos Perturbativos 87
7.5.4 Avaliação dos Métodos 91

8 Conclusão e Trabalho Futuro 111

Anexo I - Gráficos das variáveis sensíveis originais 115

Anexo II - Gráficos das variáveis sensíveis perturbadas pelo modelo escolhido 117

Acrónimos

ACP - Análise em Componentes Principais;
AICEP - Agência para o Investimento e Comércio Externo de Portugal;
AIP-CCI- Associação Industrial Portuguesa- Câmara de Comércio e Indústria;
ANI- Agência Nacional de Inovação;
BPLIM - Laboratório de Investigação em Microdados do Banco de Portugal;
CAE- Classificação da Atividade Económica;
CAP- Confederação dos Agricultores de Portugal;
CCP - Confederação do Comércio e Serviços de Portugal;
CDE - Controlo de Divulgação Estatística;
CEC- Concelho Empresarial do Centro;
CTP- Confederação do Turismo de Portugal;
DIS-SUDA - *Data Intrusion Simulante - Special Uniques Detection Algorithm*;
EAM - Erro Absoluto Médio;
EGADP - *Exact General Additive Data Perturbation*;
EQM - Erro Quadrático Médio;
ERDF- *European Regional Development Fund*;
ESF- *European Social Fund*;
GADP - *General Additive Data Perturbation*;
I&T - Informação e Tecnologia;
IAPMEI- Agência da Competividade e Inovação;
IPSO - *Information Preserving Statistical Obfuscation*;
MDVM - Microagregação por Máxima Distância ao Valor Médio;
MSU - *Minimal Sample Unique*;
OT- Objetivo Temático;
PI- Prioridade de Investimento;
PO - Programa Operacional;
QI PME- Incentivos de Qualificação para as pequenas e médias empresas;
PME- Pequenas e Médias Empresas;
POCI- Programa Operacional da Competitividade e Internacionalização;
PRAM - *Post Randomization Method*;
RCI- *Resorts Condominiums International*;
ROMM - *Random Orthogonal Matrix Masking*;
SGO- Sistema de Informação do COMPETE 2020;
SI- Sistemas de Informação;
SIFSE- Sistema de integração do Fundo Social Europeu;
SUDA - *Special Uniques Detection Algorithm;*
TI- Tipologia da Intervenção;
TIC - Tecnologias de Informação e Comunicação;
TP- Turismo de Portugal

Glossário

Anonimização: Método de preservação de informações privadas ou confidenciais por meio de exclusão ou codificação de variáveis identificadoras presentes na base de dados original;

Base de microdados original: Base de microdados antes da aplicação de uma perturbação;

Base de microdados perturbada/alterada: Base de microdados resultante da aplicação de uma perturbação;

Confidencialidade: Confidencialidade nos dados é uma propriedade dos dados, geralmente resultante de medidas legislativas, que impede a divulgação não autorizada;

Controlo de Divulgação Estatística: Técnica estatística de controlo da divulgação de dados, pode ser definida como um conjunto de métodos para reduzir o risco de informações que identifique indivíduos, empresas ou outras organizações. Tais métodos estão apenas relacionados com a etapa de divulgação e normalmente são baseados em restrições aplicadas à quantidade de informação a disponibilizar ou a modificar;

Dados Longitudinais: Dados gerados por medidas repetidas ao longo do tempo em diferentes indivíduos;

Dados Confidenciais: Dados que permitem, de uma forma direta/indireta a identificação de um indivíduo;

Estrutura Hierárquica: A base de dados é constituída por observações ligadas entre si, isto é, as observações individuais pertencem a um grupo, por exemplo, agregados familiares ou empregados de empresas;

Identificação: Ocorre quando uma pessoa ou organização reconhece ou aprende algo que desconhecia, relativamente a uma pessoa ou organização, através dos dados disponibilizados;

Risco de Identificação: O risco de identificação de uma observação ocorre se uma estimação da informação confidencial de uma observação é possível, correspondendo à probabilidade de identificação de uma observação [3];

Valor Limite: Valor abaixo do qual uma observação é considerada segura para divulgação;

Variáveis Chave: Um conjunto de variáveis que quando combinadas podem ser ligadas a informação externa de forma a possibilitar a identificação de uma observação;

Variáveis confidenciais/sensíveis: Variáveis cujos valores não devem ser descobertos. A determinação de variáveis sensíveis é alvo de preocupações legais e éticas e deve ser aplicado tendo em consideração a legislação nacional relativamente à lei de proteção de dados em vigor.

Variáveis Identificadoras: Variáveis que identificam inequivocamente o individuo/entidade;

Variáveis Identificadoras Indiretas: Variáveis que que possibilitam deduzir as unidades estatísticas a partir de informação que não conste das variáveis identificadoras diretas;

Métodos Determinísticos: Métodos de CDE que seguem um certo algoritmo e produzem os mesmos resultados quando aplicados repetidamente nos mesmos dados;

Microdados: A menor unidade de informação processada por uma base de dados. Uma base de microdados contém informação a nível individual.

Observação *outlier*: Observação que possuí um valor demasiado distante dos valores das outras observações;

Perda de Informação: Refere-se à redução de informação na base de microdados perturbada pelos métodos de CDE;

Utilidade dos dados: Refere-se à semelhança na informação dos dados perturbados face aos dados originais.

Notação

Para este projeto utiliza-se a seguinte notação:

- N: Número total de observações da população;
- n: Número total de observações da amostra;
- X: Conjunto das P variáveis confidenciais originais;
- x_{ip}: Valor da observação original i da variável confidencial p;
- S: Conjunto das Q variáveis não confidenciais originais;
- s_{iq}: Valor da observação original i da variável não confidencial q;
- W: Base de dados original com todas as variáveis;
- w_{ij}: Valor da observação original i da variável j;
- Y: Conjunto das P variáveis confidenciais perturbadas;
- y_{ip}: Valor da observação perturbada i da variável confidencial p;

Índice de Tabelas

Tabela 1 - Métodos Não Perturbativos;
Tabela 2 - Métodos Perturbativos;
Tabela 3 - Exemplo do paradigma da permutação;
Tabela 4 - Exemplo da aplicação de Mapeamento Inverso;
Tabela 5 - Variável *fundo*;
Tabela 6 - Variável *instrumento*;
Tabela 7 - Variável *profinan*;
Tabela 8 - Variável *ot*;
Tabela 9 - Variável *pi*;
Tabela 10 - Variável *medida*;
Tabela 11 - Variável *organismo*;
Tabela 12 - Variável *ti*;
Tabela 13 - Variável *to*;
Tabela 14 - Variável *dom_interv*;
Tabela 15 - Variável *dstanl*;
Tabela 16 - Variável *fonte*;
Tabela 17 - Variável *estadofse*;
Tabela 18 - Variável *transacionavel*;
Tabela 19 - Variável *intensidadetecnologica*;
Tabela 20 - Variável *tic*;
Tabela 21 - Observações que não cumprem as condições da medida *k-anonymity*;
Tabela 22 - Riscos de Identificação Individuais através dos *SUDA-scores*;
Tabela 23 - Combinações únicas;
Tabela 24 - Tabela de contingência entre as variáveis *proj_i40* e *pestadosfe*;
Tabela 25 - Tabela de contingência entre as variáveis *fundo* e *estadosfe*;
Tabela 26 - Tabela de contingência entre as variáveis *fundo* e *proj_i40*;
Tabela 27 - Valores da medida UT1;
Tabela 28 - Medidas de perda de informação e risco de identificação;
Tabela 29 - Estatísticas principais da variável *dec_investaprov*;
Tabela 30 - Estatísticas principais da variável *dec_investeleg*;
Tabela 31 - Estatísticas principais da variável *dec_incentivoapov*;
Tabela 32 - Estatísticas principais da variável *investcand*;
Tabela 33 - Estatísticas principais da variável *totpagam_realizado*;
Tabela 34 - Medidas de perda de informação e risco de identificação com as restrições;
Tabela 35 - Estatísticas principais da variável *dec_investaprov* para os métodos com restrições;
Tabela 36 - Estatísticas principais da variável *dec_investeleg* para os métodos com restrições;
Tabela 37 - Estatísticas principais da variável *dec_incentivoapov* para os métodos com restrições;
Tabela 38 - Estatísticas principais da variável *investcand* para os métodos com restrições;
Tabela 39 - Estatísticas principais da variável *totpagam_realizado* para os métodos com restrições;
Tabela 40 - Gráficos entre as Variáveis perturbadas e as variáveis originais;
Tabela 41 - Gráficos entre as Variáveis perturbadas e as variáveis originais;

Tabela 42 - Variâncias das variáveis perturbadas e originais do modelo de Ruído Independente;
Tabela 43 - Coeficientes de correlação entre as variáveis originais e as perturbadas;
Tabela 44 - Coeficiente de Gini;

Índice de Figuras

Figura 1 - Instalações do Banco de Portugal, Porto;

Figura 2 - Processo de proteção, retirado de [24];

Figura 3 - Exemplo de *l-diversity*, retirado de [24];

Figura 4 - Comparação entre as diferentes opções de aplicação de metodologias de CDE, retirado de [24];

Figura 5 - Interface gráfica da função **sdcApp**;

Figura 6 - Histograma dos Riscos de Identificação Individuais;

Figura 7 - Gráfico de valores em falta antes da supressão;

Figura 8- Gráfico de valores em falta após a supressão;

Figura 9 - Gráficos de Barras das variáveis antes e após o método PRAM;

Figura 10 - Histograma dos riscos de identificação individuais;

Figura 11 - Perda de Informação *vs* Risco de Identificação (Modelo de Ruído Independente);

Figura 12 - Perda de Informação *vs* Risco de Identificação (Modelo de Ruído Correlacionado);

Figura 13 - Perda de Informação *vs* Risco de Identificação (Modelo de Ruído Multiplicativo);

Figura 14 - Perda de Informação *vs* Risco de Identificação (Microagregação pela distância de Mahalanobis);

Figura 15 - Perda de Informação *vs* Risco de Identificação (Microagregação por Máxima Distância ao Valor Médio);

Figura 16 - Perda de Informação *vs* Risco de Identificação (Microagregação com base em Análise de Componentes Principais);

Figura 17 - Perda de Informação *vs* Risco de Identificação (Microagregação pelo método de *Ranking* Individual);

Figura 18 - Perda de Informação *vs* Risco de Identificação (*Rank Swapping*);

Figura 19 - Gráficos de Perda de Informação (IL1s e EQM) *vs* Risco de Identificação;

Figura 20 - Gráficos entre as variáveis originais sensíveis;

Figura 21 - Gráficos entre as variáveis perturbadas sensíveis;

Índice de Códigos

Código 1 - Aplicação da função **createSdcObj**;
Código 2 - Aplicação da função **slotNames**;
Código 3 - Aplicação do *Slot* sdc@originalRisk;
Código 4 -Aplicação da função **addGhostVars**;
Código 5 - Aplicação da função **measure_risk**;
Código 6 - Aplicação da função **suda2**;
Código 7 - Aplicação da função **ldiversity**;
Código 8 - Aplicação da função **ldiversity**;
Código 9 - Aplicação da função **dRisk**;
Código 10 - Aplicação da função **dRiskRMD**;
Código 11 - Aplicação da função **createNewID**;
Código 12 - Aplicação da função **addNoise**;
Código 13 - Aplicação da função **argus_microaggreagation**;
Código 14 - Aplicação da função **microaggregation**;
Código 15 - Aplicação da função **argus_rankswap**;
Código 16 - Aplicação da função **rankSwap**;
Código 17 - Aplicação da função **shuffle**;
Código 18 - Aplicação da função **pram**;
Código 19 - Aplicação da função **localSupp**;
Código 20 - Aplicação da função **undolast**;
Código 21 - Aplicação da função **dUtility**;
Código 22 - Aplicação da função **gini**;
Código 23 - Aplicação da função **print**;
Código 24 - Aplicação da função **extractManipData**;
Código 25 - Estrutura da base de dados EusilcS;
Código 26 - Cálculo das medidas de risco de identificação;
Código 27 - Aplicação do método PRAM;
Código 28 - Aplicação do método de Adição de Ruído;
Código 29 - Calculo dos SUDA *scores* e medidas de perda de informação;
Código 30 - Cálculo do risco de identificação final;
Código 31 - Criação do objeto SDC e cálculo do risco de identificação original;
Código 32 - Calculo do risco de Identificação Global;
Código 33 - Aplicação de Supressão Local;
Código 34 - Medida de *k-anonymity* e DIS-SUDA *scores*;
Código 35 - Aplicação do método PRAM;
Código 36 - Cálculo dos riscos de identificação da base de microdados perturbada;
Código 37 - Alterações provocadas pelo método PRAM;
Código 38 - Aplicação do modelo de Ruído Independente;
Código 39 - Aplicação do modelo de Ruído Correlacionado;
Código 40 - Criação dos modelos de regressão;
Código 41 - Cálculo das medidas de utilidade e de risco de identificação do Modelo EGADP;

Código 42 - Cálculo das medidas de utilidade e de risco de identificação do modelo multiplicativo;
Código 43 - Aplicação do modelo *Data Shuffling*;
Código 44 - Aplicação de Microagregação pela distância de Mahalonobis;
Código 45 - Aplicação de Microagregação por Máxima Distância ao Valor Médio;
Código 46 - Aplicação de Microagregação com base em componentes principais;
Código 47 - Aplicação de Microagregação pelo método de *ranking* individual;
Código 48 - Aplicação do método *Rank Swapping*;
Código 49 - Aplicação do método *Rank Swapping*.

1. Introdução

Atualmente, existe uma grande quantidade de dados sobre determinadas unidades estatísticas que se designam por microdados, isto é, dados que contêm informação a nível individual, por exemplo, pessoas, empresas ou grupos financeiros. A procura por bases de microdados tem aumentando significativamente, pois análises económicas e empíricas ganham cada vez mais importância, e tais análises apenas serão possíveis na presença de informação detalhada na base de microdados em estudo. Por outro lado, esta procura resulta em diversos desafios legais, éticos e técnicos. Assim, as instituições estatísticas deparam-se com o grande desafio de assegurar a confidencialidade das unidades estatísticas na divulgação de uma base de microdados.

Controlo de Divulgação Estatística (CDE) é uma área em grande crescimento nos últimos anos, que permite às instituições publicar os seus dados de forma segura e eficiente do ponto de vista do utilizador. Como grande parte dos dados recolhidos pelas instituições estatísticas não podem ser diretamente publicados, devido a questões de privacidade e confidencialidade, é necessário garantir o cumprimento da legislação sobre a proteção de dados e ao mesmo tempo fornecer informação estatística de qualidade aos utilizadores. Nesse contexto, os métodos de CDE foram criados com o intuito de serem aplicados à base de microdados antes da sua divulgação.

Os métodos criados de CDE procuram trabalhar e alterar as bases de microdados originais de modo a obter uma base de microdados modificada ou perturbada que será divulgada sem revelar informação confidencial, procurando ao mesmo tempo limitar a perda de informação resultante do processo de perturbação dos dados. O objetivo da perturbação é obter uma base de microdados cujo risco de identificação não ultrapasse um determinado limiar, que geralmente é definido por quem cria a base de microdados ou por questões legais. Desde 25 de maio de 2018 que se encontra em vigor, em todo o território da União Europeia, o Regulamento Geral de Proteção de Dados (RGPD) que estabelece regras relativas ao tratamento, por uma pessoa, uma empresa ou uma organização, de dados pessoais relativos a pessoas na União Europeia. O RGPD estabelece novas regras de forma a garantir a confidencialidade dos dados e a minimizar a possibilidade de identificação de um indivíduo/instituição nas bases de microdados divulgadas [14].

No âmbito do Mestrado em Estatística para Ciência de Dados, ministrado na Universidade do Minho, foi realizado um estágio curricular no departamento de Estudos Económicos do Banco de Portugal, mais precisamente no Laboratório de Investigação em Microdados (BPLIM), situado na cidade do Porto. Neste primeiro Capítulo é realizada uma descrição do Banco de Portugal, em particular do Laboratório de Investigação em Microdados, com uma breve contextualização da instituição, as suas principais funções, os seus órgãos e como se encontra organizada. Serão também descritos os principais objetivos do estágio, a estrutura do relatório e a linguagem de programação utilizada.

1.1 BPLIM - Banco de Portugal

O Banco de Portugal (BdP) é o banco central da República Portuguesa, fundado em 1846 em Lisboa, local da sua sede. São órgãos do BdP o Governador, o Conselho de Administração, o Conselho de Auditoria e o Conselho Consultivo. Faz parte do Sistema Europeu de Bancos Centrais, do Eurosistema, do Mecanismo Único de Supervisão e do Mecanismo Único de Resolução.

O BdP é a entidade emissora da moeda nacional, mas o seu trabalho não se centra apenas nesse ramo. Tem como principais missões a estabilidade da economia portuguesa, gerir parte das reservas cambiais do Banco Central Europeu, supervisionar as instituições de crédito, sociedades financeiras e instituições de pagamento, tendo o poder de aplicar medidas preventivas e sancionatórias. Outra das suas funções principais e a mais conectada

com o presente tema em estudo é a sua função como Autoridade Estatística Nacional. É responsável pela recolha e elaboração das estatísticas monetárias, financeiras, cambiais e da balança de pagamentos. Estas estatísticas são publicadas no Boletim Estatístico e no portal BPStat- Estastísticas online (www.bportugal.pt).

Figura 1:Instalações do Banco de Portugal, Porto

Laboratório de Investigação em Microdados (BPLIM)

O BPLIM foi criado em 2016 e está localizado no Porto. É uma unidade autónoma dentro do departamento de estudos económicos e a sua missão principal é de ajudar a produção de projetos de investigação e estudos sobre a economia portuguesa. Através do BPLIM, investigadores internos e externos conseguem obter acesso a base de microdados anonimizadas e perturbadas de acordo com as suas necessidades particulares. O BPLIM distingue-se de outras instituições que disponibilizam microdados, na medida em que não só os disponibilizam como dão suporte científico e computacional aos investigadores através de formações, workshops, seminários e atividades regulares. Este apoio tenciona colmatar e solucionar eventuais problemas relacionados com o suporte científico e computacional, com o *software* estatístico utilizado e com exploração de bases de microdados.

1.2 Objetivos

O tema proposto no âmbito desta dissertação é *"Comparação de Métodos Perturbativos: utilidade e perda de informação em bases de microdados"*. Como já referido este tema abrange uma área de grande interesse nos últimos anos, pois cada vez mais existe a necessidade de proteger os dados de modo a estes serem disponibilizados a diversos utilizadores.

Este projeto pretende descrever os vários métodos perturbativos de CDE, descrever os passos a efetuar de forma a que seja possível a perturbação dos dados e ainda comparar os diferentes métodos de CDE avaliando a sua utilidade face à perda de informação. Assim os principais objetivos propostos para este tema serão:

- Comparação dos diferentes métodos de Controlo de Divulgação Estatística, com base nos seguintes indicadores:
  - Utilidade da informação;
  - Risco de Identificação.
- Aplicação e análise destes métodos numa base de microdados confidencial.

De referir ainda que, para a aplicação dos métodos de CDE e para o cálculo das diversas medidas, utiliza-se a linguagem de programação R sendo essencialmente utilizado o *package* **sdcMicro**, como descrito no Capítulo 7.

1.3 Estrutura do Relatório

Esta dissertação é composta por 9 capítulos. No Capítulo 1 introduz-se o tema em estudo, a contextualização da dissertação realizada e os objetivos propostos. No Capítulo 2 apresentam-se os conceitos fundamentais para o estudo de uma base de microdados na área de Controlo de Divulgação Estatística (CDE), e ainda é descrito os passos a realizar no processo de proteção. O Capítulo 3 descreve algumas medidas de risco de identificação para os diferentes tipos de variáveis e alguns métodos para a estimação do risco de identificação global de uma base de microdados. No Capítulo 4 estão descritos os métodos de CDE, apresentando-se uma breve explicação sobre os métodos não perturbativos, os métodos geradores de dados sintéticos e descrevendo-se com maior detalhe os métodos perturbativos. É ainda apresentado um método perturbativo para os dados longitudinais, bem como medidas para o calculo do risco de identificação e de perda de informação para este tipo de dados. Para além da descrição dos métodos, é realizada uma comparação a nível teórico, onde se conclui sobre qual o melhor método. O Capítulo 5, descreve as medidas de perda de informação para variáveis numéricas e categóricas, apresentam-se diversas medidas capazes de analisar a perda de informação entre as duas bases de microdados em estudo (original e perturbada). No Capítulo 6 apresenta-se uma descrição das funções existentes no *package* sdcMicro utilizando a linguagem de programação R. Efetua-se ainda a implementação dos métodos e das medidas na base de dados EUSILCS. O Capítulo 7 contém o caso de estudo. Inicialmente descreve-se as variáveis existentes na base de microdados em estudo (PT2020) e todos os passos e decisões realizadas para a perturbação da base de microdados, e no final é apresentada uma comparação entre os diversos métodos aplicados. Por fim, conclui-se quais os métodos que apresentam os melhores resultados para os diferentes cenários possíveis. No Capítulo 8, apresentam-se as conclusões retiradas sobre a literatura e sobre o caso de estudo apresentado e o trabalho futuro.

2. Conceitos Fundamentais

Neste Capítulo serão introduzidos alguns conceitos importantes para a aplicação dos métodos de Controlo de Divulgação Estatística (CDE).

2.1 Classificação de Variáveis

Na aplicação dos métodos CDE a uma base de microdados é importante identificar as variáveis de acordo com o risco de identificação. As variáveis de uma base de microdados podem ser separadas em três categorias:

- **Identificadores Diretos:** Variáveis que permitem a identificação de um indivíduo sem margem de duvidas (exemplo: Número de segurança social, Número de identificação fiscal);
- **Variáveis Chave:** Variáveis que são identificadores indiretos e correspondem ao conjunto de variáveis que quando combinadas entre si permitem a identificação de um indivíduo (exemplo: Género, idade, região);
- **Variáveis Não Identificadores:** Variáveis que não são variáveis chave ou identificadores diretos. Um exemplo, são variáveis que não estão incluídas em documentos externos, no entanto, continuam a ser variáveis de grande importância, pois podem conter informação confidencial ou sensível.

Para além desta classificação de variáveis, os métodos CDE classificam ainda as variáveis com base na sua sensibilidade ou confidencialidade. Apenas as variáveis chave ou as variáveis não identificadoras podem ser classificadas como sensíveis ou não sensíveis. [3]

- **Variáveis Sensíveis**: Variáveis cujo valores não devem ser identificados para qualquer indivíduo na base de microdados. A identificação destas variáveis depende normalmente de questões de ética e legalização. Por exemplo, variáveis que contenham informação criminal, comportamento sexual, registos médicos ou salário mensal são normalmente consideradas variáveis sensíveis. Se uma variável é sensível ou não tudo depende do contexto, do país e do próprio enquadramento legal;
- **Variáveis Não Sensíveis**: Variáveis que não contêm informação confidencial, tal como género, país, entre outras. No entanto, não significa que estas variáveis não, sejam importantes para efeitos de investigação e portanto para a aplicação dos métodos CDE.

Esta classificação de variáveis depende, essencialmente, da disponibilização de base de dados externas que possam conter informação que quando combinada permita a identificação de um indivíduo.

Em todos estes tipos de variáveis poderão existir variáveis contínuas e categóricas, no entanto, os métodos de divulgação estatística diferem consoante o tipo de variável.

2.2 Tipos de Identificação

Identificação ocorre quando um intruso revela informação anteriormente desconhecida sobre um indivíduo, através da base de dados divulgada.

Antes de aplicar os métodos CDE é necessário assegurar que a identificação dos indivíduos não aconteça. Para isso, são considerados três tipos de identificação:

- **Identificação da Identidade:** Ocorre quando um utilizador identifica corretamente um indivíduo com base em informação disponível. Esta identificação é possível através de identificadores diretos, de combinações raras de variáveis chave e do conhecimento do valor exato das variáveis chave contínuas em base de dados externas;

- **Identificação de Atributos:** Neste caso, o utilizador é capaz de determinar algumas características de um indivíduo com base em informação disponível na base de dados divulgada. Por exemplo, se numa dada região as pessoas com idades compreendidas entre os 56 e 60 anos de etnia árabe estão todas desempregadas, se considerar as variáveis chaves como Etnia, Idade e Região, e a Situação de Emprego como variável sensível, então nesta região, todas as pessoas com as características, etnia árabe, idade entre os 50 e 60 anos são identificadas como pessoas em situação de desemprego. A combinação de variáveis chave escolhidas permitiu a identificação da variável sensível [24].

- **Identificação por Inferência:** Neste caso, com a aplicação de Modelos de Regressão na base de dados divulgada, um utilizador obtém informação sensível de um indivíduo. Por exemplo, com um modelo de regressão, um utilizador consegue inferir sobre as variáveis sensíveis de um indivíduo, utilizando os atributos comuns entre base de dados externas e a base de dados divulgada [24].

Os métodos CDE normalmente previnem a identificação da identidade e de atributos [3].

Um passo importante no processo de CDE é definir uma lista de possíveis cenários de identificação, tendo em consideração as possíveis combinações de variáveis chave e a informação externa disponível em base de dados e por fim tratar os dados para prevenir a sua identificação.

Um cenário de identificação avalia a informação disponível em base de microdados externas e de que forma estas possibilitam a identificação de um indivíduo pertencente à base de dados divulgada. Esta avaliação é feita através das variáveis comuns entre a base de microdados divulgada e as bases de microdados externas. O risco de identificação depende diretamente da inclusão ou da exclusão de variáveis no conjunto de variáveis chave.

O principal e mais importante resultado da avaliação destes cenários é a identificação das variáveis chaves às quais serão aplicados os métodos CDE, pois essas variáveis serão baseadas na informação disponível publicamente.

A estrutura dos dados também é uma característica que influencia a probabilidade de identificar um indivíduo da base de microdados divulgada. Caso os dados possuam uma estrutura hierárquica, isto é, as unidades individuais estão distribuídas por grupos, a possibilidade de identificar um indivíduo aumenta, por duas razões [24]:

- Se um indivíduo pertencente a um grupo é identificado, a estrutura hierárquica permite a identificação de todos os membros do mesmo grupo;

- Valores de variáveis de um membro do grupo que são comuns em todos os membros, poderão ser usados para a re-identificação de um outro indivíduo do mesmo grupo.

Estes grupos poderão representar, por exemplo, agregados familiares, escolas, hospitais ou grupos económicos.

2.3 Risco de Identificação e Perda de Informação

Após a aplicação dos métodos CDE é necessária uma avaliação da alteração provocada na base de dados original. Nesta secção serão discutidos apenas dois métodos de avaliação com grande importância devido à relação existente entre eles:

- **Risco de Identificação:** É definido com base em possíveis cenários de identificação, e avalia o risco de um utilizador identificar as características de um indivíduo pertencente à base de microdados. Medidas de risco de identificação serão apresentadas no Capítulo 3;

- **Perda de Informação:** Após a aplicação dos métodos é necessário avaliar a informação que foi perdida com o processo de perturbação dos dados, comparando os valores da base de microdados alterada com a base de microdados original. Antes da aplicação dos métodos nos dados originais, assume-se que estes dados têm zero informação perdida. Existem várias medidas para avaliar a informação perdida no processo de perturbação, que serão descritas no Capítulo 4.

O objetivo dos métodos CDE será atingir um equilíbrio entre o risco de identificação e a perda de informação, ou seja, por um lado obter uma base de dados segura para divulgação, por outro lado, útil para os utilizadores da mesma [25]. Quanto menor for o risco de identificação maior será a perda de informação, ou seja, menor a utilidade dos dados para os utilizadores. Para o risco de identificação diminuir é necessário a existência de maior perturbação, no entanto se existe maior perturbação, as diferenças entre a base de dados original e a base de dados perturbada serão maiores. Os métodos de CDE são caracterizados pelo equilíbrio entre o risco de identificação e a utilidade dos dados para os utilizadores [3].

Para que seja possível estabelecer um limiar para o risco de identificação e para a perda de informação é necessário conhecer a finalidade da base de microdados a divulgar, e que tipo de análises serão realizadas pelos utilizadores. Por exemplo, caso uma base de microdados seja divulgada ao público em geral, o nível de risco de identificação deve ser o mais baixo possível, pois o objetivo é não revelar informação confidencial a qualquer custo. Por outro lado, caso a base de microdados seja divulgada num ambiente mais seguro, e apenas a uma parte do público, já não existe tanta preocupação na identificação de um indivíduo, mas sim na utilidade da informação para análises.

Na aplicação dos métodos de CDE, existem dois procedimentos de forma a controlar o risco de identificação [8]:

- **Utilidade Primeiro:** Aplica-se um método de perturbação, que se pretende que provoque pouca perda de informação. Caso o risco de identificação calculado seja demasiado elevado, então é necessário a aplicação do mesmo método mas com maior perturbação, sacrificando um pouco mais a utilidade da informação;

- **Confidencialidade Primeiro:** Neste caso, aplica-se um método de perturbação que permita a definição de um limiar para o risco de identificação. Se a perda de informação é demasiado alta, é necessário a aplicação do mesmo método, mas com um limiar menos rigoroso para o risco de identificação.

Um passo importante para a avaliação do risco de identificação é a escolha das variáveis chaves, pois as combinações dessas variáveis servem como estimativas para o risco de identificação. Na avaliação da perda de informação dos dados perturbados é importante verificar se não existem alterações significativas nas relações entre as variáveis, caso contrário, os utilizadores podem questionar sobre a credibilidade dos dados perturbados.

2.4 Processo de Proteção

Na proteção de uma base de microdados, o objetivo é escolher um método que permita obter uma base de microdados perturbada com informação útil para os utilizadores, e simultaneamente não revelar informação confidencial. Todo o processo depende de diferentes regulamentos de acordo com o país e as bases de

microdados em estudo. Nesta dissertação, apresenta-se uma descrição dos métodos, das medidas de avaliação, e ainda uma comparação entre os diversos métodos apresentados.

Na Figura 2 ilusta-se o funcionamento do processo de proteção de uma base de microdados [24].

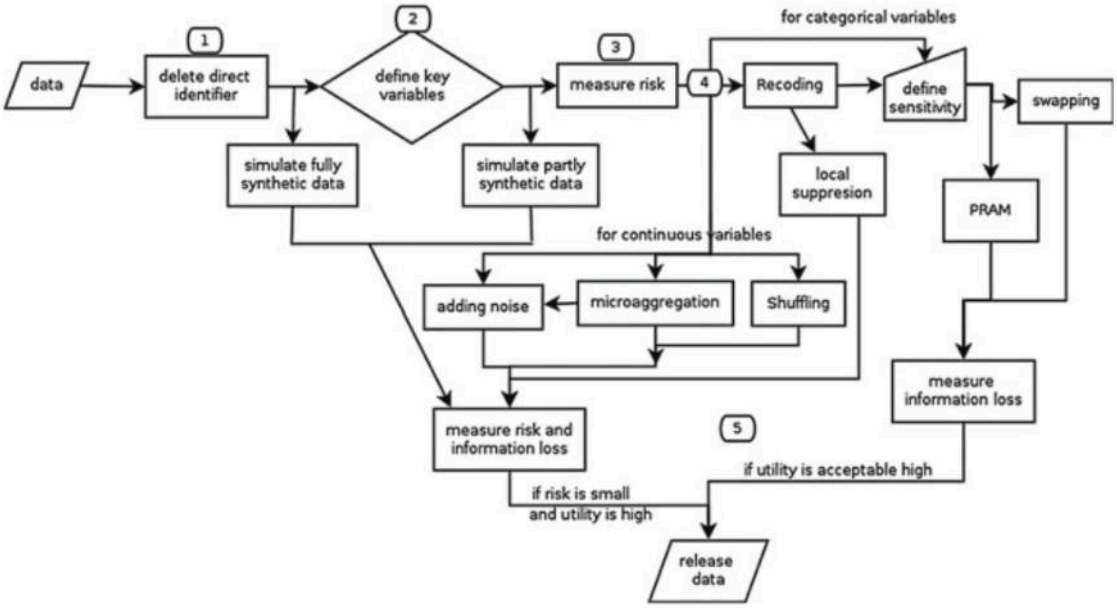


Figura 2: Processo de proteção, retirado de [24]

Os pontos marcados com números na Figura 2, representam:

1. O processo de proteção começa pela anonimização da base de microdados;
2. Para métodos não sintéticos, a escolha das variáveis chave é de grande importância;
3. Para variáveis chave categóricas, antes da aplicação de qualquer método, é aconselhável a estimação do risco de identificação global e individual. Para variáveis sensíveis contínuas é importante calcular as observações *outliers*;
4. Aplicam-se os diferentes métodos de CDE conforme as variáveis confidenciais/sensíveis;
5. A cada aplicação de um método, as medidas de risco de identificação e de perda de informação devem ser reportadas.

3. Medidas de Risco de Identificação

A medição do risco de identificação numa base de microdados é uma tarefa essencial no processo de perturbação. Esta medição pode ser realizada antes e/ou após a perturbação de uma base de microdados. É importante calcular o risco de identificação da base de microdados original, por forma a perceber qual o melhor método de CDE a utilizar de acordo com o risco calculado. Após a perturbação, o cálculo do risco de identificação é realizado por forma a assegurar que a base de microdados não divulga informação confidencial. Em termos de risco de identificação C.R. Rao [20] refere que após a aplicação do método de CDE, o utilizador não é capaz de inferir acerca da identidade de uma observação ou obter uma estimativa correta do valor das variáveis sensíveis. Dalenius [5] refere que a identificação ocorre quando a base de microdados divulgada permite ao utilizador obter uma estimativa adequada dos valores das variáveis confidenciais.

A definição de risco de identificação pode ser formalizada da seguinte forma [20]: assume-se que os utilizadores contêm alguma informação sobre as variáveis confidenciais X e não confidenciais S, bem como o acesso a base de microdados externas onde as variáveis não confidenciais S estão presentes. Assim, antes da divulgação da base de microdados perturbada, o risco de identificação é definido como a probabilidade de um utilizador prever os valores de X através da função de densidade condicional $f(X|S)$. Quando o conjunto de variáveis confidenciais perturbadas Y é divulgado, os utilizadores possuem agora informação adicional, e podem utilizar a função de densidade condicional $f(X|S,Y)$ de forma a obter valores preditos de X. No entanto, se $f(X|S,Y) = f(X|S)$, então o acesso aos dados perturbados não fornece informação adicional aos utilizadores. Assim, o risco de identificação é ideal quando:

$$f(X|S,Y) = f(X|S),$$

ou seja, dado o conjunto de variáveis não confidenciais S, as variáveis confidenciais originais X e as variáveis confidenciais perturbadas Y são independentes.

3.1 Medidas de Risco para Variáveis Categóricas

Para variáveis categóricas, o risco de identificação normalmente baseia-se no conceito de singularidade na amostra e/ou população, isto é, o foco está em observações que possuem combinações raras de variáveis chave. De seguida apresentam-se algumas medidas utilizadas para calcular o risco de identificação em variáveis categóricas.

Frequency Counts

O cálculo de *Frequency Counts* serve de base a muitos métodos de estimação do risco de identificação [24]. Define-se f_k como o número de observações da amostra com combinação k de variáveis chave e F_k como o número de observações da população com combinação k de variáveis chave, ou seja, f_k ou F_k, correspondem ao número total de indivíduos que possuem a mesma combinação k de variáveis chave na amostra e na população, respetivamente.

Caso $f_k = 1$, então a combinação k de variáveis chave apenas está associada a uma observação, e designa-se por combinação única de variáveis chave na amostra, ou seja, mais nenhuma outra observação partilha a mesma combinação de variáveis chave.

Caso $F_k = 1$, a combinação k de variáveis chave é única tanto na população como na amostra, ou seja, em toda a população apenas uma observação apresenta a combinação k e o risco de identificação é máximo. Combinações únicas populacionais merecem uma atenção especial quando se avalia o risco de identificação.

Em geral, quanto menor o número de observações com a mesma combinação de variáveis chave, maior será o risco de identificação de cada observação. Supondo que a base de dados divulgada contém uma observação com uma combinação rara de variáveis chave, um utilizador que tente associar esta observação com uma base de dados externa, onde a mesma combinação de variáveis está presente, a probabilidade de identificação da observação é elevada. Se a mesma combinação de variáveis chave fosse comum a diversas observações, a probabilidade de identificação seria menor.

Princípios de K-anonymity e l-diversity

Uma base de dados satisfaz a condição de *K-anonymity* se cada combinação k de variáveis chave é comum a pelo menos K observações, ou seja,

$$F_k \geq K, \forall k \in \{1, ..., C\}\,, \tag{3.1}$$

onde C representa o número total de combinações possíveis para as variáveis chave escolhidas.

A medida de risco é representada pelo número total de observações que não satisfazem a condição de *K-anonymity* para um certo valor K:

$$\sum_{i=1}^{N} I_i(F_k < K), \tag{3.2}$$

onde i se refere à i-ésima observação da base de dados, N o número total de observações, F_k representa a *frequency count* populacional da combinação k de variáveis chave associada à observação i e I_i representa a função indicatriz da observação i, definida como

$$I_i = \begin{cases} 1 & \text{se } F_k < K \\ 0 & \text{se } F_k \geq K \end{cases}, \tag{3.3}$$

Por exemplo, considere-se uma amostra constituída por 4 indivíduos, onde se escolhe como variáveis chave Etnia, Género e Escolaridade. Supondo que dois indivíduos partilham a mesma combinação de variáveis chave $\{$Negra, Masculino, Secundário$\}$ e os outros dois indivíduos partilham a combinação de variáveis chave $\{$Negra, Feminino, Secundário$\}$, isto significa, que as combinações de variáveis chave apresentam $F_k = 2$, $k = 1, 2$. Então, a amostra satisfaz a condição de *2-anonymity* e não satisfaz a condição de *3-anonymity* pois cada combinação de variáveis chave apenas contém duas observações.

O método *K-anonymity* apresenta uma grande desvantagem, pois não é suficientemente rigoroso. Informação sensível pode ser descoberta mesmo que a condição de *K-anonymity* seja satisfeita, isto pode acontecer em base de dados que contenham informação sensível categórica mas não identificável [3]. Para resolver este problema foi criado o conceito de *l-diversity*.

Um grupo de observações que partilhe a mesma combinação de variáveis chave é considerado *l-diverse* se existem pelo menos l categorias diferentes para a variável sensível, isto é, o valor da variável sensível não pode ser comum para todas as observações que partilhem a mesma combinação de variáveis chave. O nível requerido de *l-diversity* depende das possíveis categorias nas variáveis sensíveis.

	Key variables		f_k	Sensitive variable	Distinct l-diversity
	Gender	Age group		Medical condition	
1	Male	30s	3	Cancer	2
2	Male	30s	3	Heart disease	2
3	Male	30s	3	Heart disease	2

Figura 3: Exemplo de *l-diversity*, retirado de [24]

Na Figura 3 é possível ver que os indivíduos 1, 2 e 3 cumprem a condição de *2-diversity*, pois partilham a mesma combinação de variáveis chave (género e grupo etário) e a variável sensível apresenta duas categorias distintas ($l = 2$). Neste caso, os dados são *2-diverse* e satisfazem a condição de *3-anonymity*.

Special Uniques Detection Algorithm

Os métodos apresentados anteriormente são baseados na procura de variáveis chave através de informação disponível em documentos externos, no entanto, na prática isso pode se tornar algo muito complexo. De forma a ultrapassar este problema, foi criado o algoritmo *Special Uniques Detection Algorithm* (SUDA). Os testes realizados com este algoritmo permitem concluir que produz estimativas de risco significativamente eficazes [3].

Para a aplicação do algoritmo (SUDA) é necessário a classificação de conjuntos de variáveis chave como *Minimal Sample Unique*.

Minimal Sample Unique (MSU) é o menor conjunto de combinações únicas de variáveis chave da amostra/população. Para classificar um conjunto de variáveis chave como um conjunto MSU de dimensão q, é necessário verificar se tal conjunto cumpre o requisito minimal. Caso exista algum subconjunto de combinações únicas de variáveis chave, com dimensão $q - 1$, contido no conjunto em avaliação, então este conjunto de dimensão q não pode ser classificado como MSU. Uma observação pode conter mais que um conjunto classificado como MSU [24].

Na aplicação do algoritmo SUDA, o primeiro passo é identificar todos os conjuntos de combinações únicas de variáveis chave classificados como MSU na amostra/população. De seguida, atribuir um *SUDA score* a cada observação da base de microdados. O cálculo do *SUDA score* tem em consideração:

- Quanto menor a dimensão de um conjunto MSU, maior é o risco de identificação da observação;
- Quanto maior o número de conjuntos MSUs de uma observação, maior o risco de identificação da observação.

O interesse deste algoritmo é maior em conjuntos de variáveis chave de menor dimensão, desta forma procura por MSUs é limitada para uma dimensão máxima dos conjuntos de variáveis chave. Para cada MSU v, de dimensão q, um *SUDA score* é calculado da seguinte forma:

$$s_v = \begin{cases} \frac{1}{Q!} \prod_{d=q}^{M} (Q - d) & \text{se } d \leq M \\ 0 & \text{caso contrário,} \end{cases} \tag{3.4}$$

onde M é a dimensão máxima de cada conjunto de MSU (estabelecida pelo responsável do processo de perturbação) e Q o número total de variáveis chave categóricas. O argumento $\frac{1}{Q!}$ provoca a normalização dos *SUDA scores*.

O *SUDA score* para cada observação, é calculado através da soma de todos os *SUDA scores* dos conjuntos classificados como MSU associados a essa observação.

Para estimar o risco de identificação de uma observação, os *SUDA scores* podem ser usados em conjunto com a métrica *Data Intrusion Simulation* (DIS), obtendo o algoritmo DIS-SUDA, que é um método de avaliação que fornece o risco de identificação de cada observação. Este algoritmo calcula a probabilidade condicionada da identificação correta de uma observação na base de dados em estudo, sabendo da existência de conjuntos classificados como MSUs associados a essa observação. O valor desta probabilidade é definido como o risco de identificação para cada indivíduo.

3.2 Medidas de Risco para Variáveis Numéricas

As medidas de risco para variáveis contínuas são medidas à *posteriori*, ou seja, são aplicadas após a perturbação das variáveis contínuas, e calculam a distância entre os dados perturbados e os dados originais, muito semelhante aos métodos de perda de informação.

Record Linkage

Seja x_{ip} o valor da variável confidencial p na observação i da base de dados original e y_{ip} a mesma observação na base de dados perturbada. Para cada observação y_{ip}, é calculada a distância a todas as observações de X_p. De seguida, considera-se apenas a primeira e a segunda observação mais próxima de y_{ip}, x_{1p} e x_{2p}, respetivamente. Se x_{1p} ou x_{2p} é a observação original utilizada para gerar y_{ip} nos dados perturbados, então diz-se que a observação y_{ip} está 'ligada' à observação original. Repete-se este processo para todas as observações. Finalmente, o risco de identificação global é definido como a percentagem de observações y_{ip} 'ligadas' às observações originais x_{ip}, $i = 1, ..., N$ e $p = 1, ..., P$. *Record Linkage* é um método que avalia o número correto de ligações quando se comparam os valores perturbados com os valores originais e é baseado na distância entre os valores originais e os valores perturbados. Este processo é muito intensivo computacionalmente e como tal não é aconselhado a aplicação em base de dados com elevado número de observações.

Medida Intervalar

Uma aplicação bem-sucedida dos métodos de CDE deve resultar em valores perturbados não muito próximos dos valores originais, porque caso sejam muito semelhantes a identificação é possível.

Na aplicação desta medida, são criados intervalos em torno de cada valor perturbado e de seguida verifica-se se o valor original pertence ao intervalo definido. Valores que estejam dentro do intervalo são considerados demasiado próximos e como tal é necessária maior perturbação de forma a garantir que a identificação não ocorre. A amplitude do intervalo é calculada através do desvio padrão (σ_p) da variável confidencial p e de um parâmetro escalar α, definido pelo responsável da base de microdados. Os intervalos são obtidos da seguinte forma:

$$\left[y_{ip} - \sigma_p \cdot \alpha; y_{ij} + \sigma_p \cdot \alpha\right], \tag{3.5}$$

onde y_{ip} é a observação perturbada i da variável p.

O risco de identificação é calculado como a proporção dos valores originais que pertencem ao intervalo definido em (3.5). Para grande parte das observações é um método satisfatório, no entanto para observações *outliers* não é eficaz na perceção do risco de identificação desta observação, pois observações *outliers* contêm um risco de identificação maior [7].

Contagem de Outliers

Observações *outliers* são importantes na medição do risco de identificação em dados contínuos. Quando existem *outliers*, estas observações apresentam valores muito elevados ou muito baixos em comparação às restantes observações da base de dados, e nestes casos, mesmo com a aplicação de perturbação, existirá sempre um risco de identificação elevado.

Na prática a deteção das observações *outliers* é feita através da identificação de valores que são superiores a um determinado percentil. Dito isto, é de se esperar que no processo de perturbação essas observações sejam sujeitas a uma maior perturbação em comparação às restantes observações [24].

Esta medida estima o risco de identificação e é descrita da seguinte forma:

1. Calcula a Distância de Mahalanobis Robusta entre indivíduos, obtendo assim uma distância multivariada para cada observação i;

2. Estima os intervalos para cada observação i dos dados originais; a amplitude do intervalo depende das distâncias calculadas no primeiro passo e de um parâmetro escalar. Quanto maior a distância de Mahalanobis Robusta maior a amplitude dos intervalos;

3. Verifica se os valores perturbados pertencem aos intervalos calculados no passo 2. Caso o valor pertença ao intervalo então diz-se que essa observação está em risco de identificação.

Assim, uma base de dados perturbada é considerada segura para divulgação, quando todas as observações y_{ip} contenham pelo menos m observações dos dados perturbados na sua vizinhança. m é definido de acordo com as vizinhanças das outras observações, ou pelo responsável da base de microdados.

3.3 Risco de Identificação Individual

O Risco de Identificação Individual tem como base as variáveis chave categóricas. Assim, através das *frequency counts* calcula-se o risco de identificação individual da seguinte forma:

$$r_i = \frac{1}{F_k}, \tag{3.6}$$

onde F_k representa a *frequency count* populacional da combinação k de variáveis chave associada à observação i.

Esta medida é uma alternativa aos *SUDA scores* e conduz a resultados conservativos, isto é, os riscos de identificação estimados através desta medida podem conduzir a riscos de identificação desajustados. Para além disso, este risco será o mesmo para observações que partilham a mesma combinação de variáveis chave.

Nesse sentido, aconselha-se a análise de outras medidas complementares de risco, nomeadamente, as medidas de risco para variáveis sensíveis contínuas.

3.4 Risco de Identificação Global

No cálculo do risco de identificação global para uma base de dados perturbada, é possível a associação das diversas medidas de risco individual [3]. Estas medidas de risco global necessitam de um cuidado especial pois é possível a estimação de um risco global aceitável, e mesmo assim existirem observações com risco de identificação elevado que serão compensadas por outras observações com risco de identificação reduzido.

Média das Medidas do Risco de Identificação Individual

A forma mais simples de associar os riscos individuais é através da média aritmética de todos os riscos de identificação individuais das observações na amostra:

$$R = \frac{1}{N}\sum_{i=1}^{N} r_i, \tag{3.7}$$

onde r_i representa o risco individual da observação i e N o número total de observações na base de dados.

O risco de identificação global corresponde a uma média dos riscos individuais e indica a proporção de todas as observações que poderão ser identificadas por um utilizador.

Contagem de indivíduos com risco de identificação indiividual superior a um valor pré-defenido

Outra forma de expressar o risco global de uma base de dados perturbada é através da contagem do número total de observações que excedem um certo valor de risco individual [3]. A determinação deste valor limite pode ser absoluto, contabilizando o número de observações que contêm um risco de identificação superior, por exemplo, a $0,05$. Este cálculo também pode ser feito em termos relativos, determinando-se a proporção de observações que possuem risco de identificação individual superior ao terceiro quartil da função distribuição do risco de identificação individual. Assim o risco global seria a proporção de observações que satisfazem uma das condições descritas.

Risco de Identificação do Grupo

Como já foi referido anteriormente, as bases de dados podem ter uma estrutura hierárquica, onde as observações pertencem a grupos. Neste caso, a identificação de um indivíduo pertencente a um grupo pode levar à identificação de outros indivíduos do mesmo grupo. Assim, é fácil de perceber que quando esta estrutura está presente, o risco de identificação de um grupo h é o risco de pelo menos um membro ser identificado [3]:

$$r^h = P(A_1 \bigcup A_2 \bigcup ... \bigcup A_{N_h}) = 1 - \prod_{i=1}^{N_h}(1 - P(A_i)) \tag{3.8}$$

onde A_i é o evento que permitiu a identificação do indivíduo i e $P(A_i) = r_i$ é o risco individual de identificação da observação i e N_h corresponde ao número de indivíduos pertencentes ao grupo h.

Como risco de identificação do grupo é definido como a probabilidade de pelo menos um elemento do grupo ser identificado por um utilizador, então este valor não pode ser inferior ao risco de identificação individual.

4. Métodos de Controlo de Divulgação Estatística

Os métodos de CDE têm como principal objetivo modificar os dados estatísticos de modo a permitir a sua divulgação, sem disponibilizar informação que possibilite a identificação de uma observação a partir da base de dados perturbada. A maior dificuldade neste processo é atingir essa modificação de forma eficaz, ou seja, publicar uma base de dados segura do ponto de vista da identificação e útil do ponto de vista dos utilizadores.

Após a identificação do tipo de variáveis é necessário implementar os métodos de CDE nas variáveis que possam conduzir à identificação de uma observação. Neste projeto, o objetivo principal consiste em explorar os métodos perturbativos, no entanto, faz-se uma breve descrição de outros métodos de CDE.

Existem três tipos de métodos CDE [24]:

- **Métodos Não Perturbativos:** São métodos que não alteram os dados, em vez disso produzem supressões ou reduções parciais, por exemplo, *Global Recoding, Local Supression*;
- **Métodos Perturbativos:** São métodos que provocam a alteração dos valores nas variáveis sujeitas a perturbação, por exemplo, *Post-Randomization Method, Shuffling*, Adição de Ruído e Micro-agregação;
- **Geração de dados sintéticos:** São técnicas geradoras de dados sintéticos que resultam geralmente numa base de dados de dimensão mais reduzida e que preservam algumas estatísticas ou relações entre variáveis presentes na base de dados original. Usualmente a aplicação destes métodos leva a um menor risco de identificação, no entanto é um processo mais complexo e de baixa relevância para o tema em estudo.

Para além desta distinção, também é possível dividir os métodos em probabilísticos e determinísticos:

- **Métodos Probabilísticos:** Baseiam-se em mecanismos probabilísticos ou em mecanismos geradores de números aleatórios;
- **Métodos Determinísticos:** Baseiam-se num certo algoritmo e produzem os mesmos resultados se aplicados repetidamente na mesma base de dados.

Estes métodos diferem consoante a natureza das variáveis.

4.1 Métodos Não Perturbativos

Na Tabela 1 estão apresentados os principais métodos não perturbativos de CDE, tipo de aplicação e a sua classificação como determinístico ou probabilístico.

Tabela 1: Métodos Não Perturbativos

Método	Variáveis Categóricas	Variáveis Contínuas	Determinístico	Probabilístico
Recoding	X	X	X	
Supressão Local	X		X	

Recoding

Recoding é um método determinístico que pode ser aplicado tanto a variáveis contínuas como a variáveis categóricas. Utiliza-se de forma a diminuir o número de categorias distintas de uma variável categórica ou o número de valores possíveis de uma variável contínua. Para variáveis categóricas, *Recoding* combina várias categorias, aumentando assim as *frequency counts* de cada combinação de variáveis chave e por outro lado, provocando uma diminuição do detalhe nos dados. Para variáveis numéricas, o método *Recoding* corresponde a discretizar a variável, ou seja, transformar a variável contínua numa variável categórica onde as categorias correspondem a intervalos de valores possíveis para a variável em estudo [24]. *Recoding* é aplicado a todas as observações da base de dados e não apenas às observações em risco de identificação [3].

Este processo dispõe de dois métodos alternativos:

- **Global Recoding**

  Combina as várias categorias de uma variável categórica ou então constrói intervalos de confiança para variáveis contínuas, reduzindo assim o número de categorias e o risco de identificação, especialmente para categorias com poucas observações.

  Global Recoding é normalmente o primeiro método a ser utilizado quando se pretende proteger uma base de dados [3].

- **Top and Bottom Coding**

  Este método pode ser aplicado apenas a variáveis categóricas ordinais ou numéricas, tendo em conta que os valores das variáveis chave necessitam de estar ordenados. *Top and Bottom Coding* é semelhante a *Global Recoding*, no entanto, em vez de recodificar todos os valores, apenas recodifica o valor inicial e/ou o valor final da variável numérica ou das categorias das variáveis chave [3].

Supressão Local

Após a aplicação do *Recoding* é possível que ainda existam combinações únicas de variáveis chave na base de dados perturbada, ou seja, $F_k = 1$. Quando tal acontece é aconselhada a aplicação do método de Supressão Local, com o objetivo de que todas as observações cumpram a condição de *K-anonymity*.

Uma aplicação deste método é a criação de valores em falta, ou seja, corresponde a substituir as combinações únicas de variáveis chave por valores em falta [24]. Este tipo de processo não é o mais eficaz, pois valores em falta criarão grande perda de informação no processo de proteção da base de microdados. Em alternativa, é possível a aplicação deste método a observações que contenham um risco individual superior a um dado valor limite fixado.

4.2 Métodos Perturbativos

Na Tabela 2 estão representados os principais métodos perturbativos de CDE, tipo de aplicação e a sua classificação como determinístico ou probabilístico.

Tabela 2: Métodos Perturbativos

Método	Variáveis Categóricas	Variáveis Contínuas	Determinístico	Probabilístico
Modelos lineares		X		X
Modelos não lineares		X		X
Microagregação	X	X		X
Re-amostragem		X		X
Arredondamento		X		X
PRAM	X			X
Generalização	X			X
Rank Swapping	X	X		X
Shuffling		X		X
DDP	X	X	X	

De seguida, apresentam-se os métodos mais detalhadamente por tipo de variável, categórica ou numérica.

4.2.1 Métodos Perturbativos para Variáveis Categóricas

Normalmente os métodos perturbativos são aplicados a variáveis numéricas sensíveis, no entanto, é possível a aplicação destes métodos a variáveis categóricas sensíveis ou a variáveis chave. Estes métodos quando aplicados a estas variáveis modificam as categorias da variável em estudo, isto é, provocam trocas entre categorias ou então a substituição de várias categorias em uma só.

Post Randomization Method (PRAM)

Caso exista um grande número de variáveis chave categóricas, normalmente superior a cinco, o método *Recoding* poderá não reduzir suficientemente o risco de identificação, ou então, o método de Supressão Local pode conduzir a resultados pouco eficazes quanto à utilidade da base de microdados perturbada e neste caso o método perturbativo PRAM é o método mais eficaz para a proteção dos dados [13]. No entanto, ao aplicar este método a variáveis chave categóricas, a medida de *K-anonymity* deixa de ter significado, pois as novas combinações não traduzem a realidade dos dados em estudo. A aplicação deste método realiza-se normalmente a variáveis não confidenciais e variáveis não chave.

PRAM é um método perturbativo probabilístico utilizado para a proteção das variáveis chave categóricas. Este método troca as categorias de uma variável categórica com base numa matriz de transição pré-definida, que especifica a probabilidade de uma categoria ser trocada por outra categoria.

Seja **P** a matriz de transição PRAM, então **P** é definida como:

$$\mathbf{P} = \begin{pmatrix} p_{11} & p_{12} & \cdots & p_{1m} \\ p_{21} & p_{22} & \cdots & p_{2m} \\ . & . & . & . \\ p_{m1} & p_{m2} & \cdots & p_{mm} \end{pmatrix}, \tag{4.1}$$

onde m representa o número de categorias na variável categórica sujeita a perturbação e p_{ij} é a probabilidade da categoria i e da categoria j trocarem de posição entre si. O valor de p_{ij}, quando $i = j$, representa a probabilidade de a categoria i permanecer inalterada.

PRAM protege os registos dos dados originais através da aplicação de uma perturbação, no entanto como o mecanismo de probabilidades é conhecido, matriz **P**, as características dos dados originais podem ser estimadas através dos dados perturbados. Este processo pode ser aplicado a cada registo separadamente, permitindo assim maior flexibilidade na especificação da matriz de transição **P** como uma função de parâmetros de acordo com os efeitos pretendidos [24].

Semantic Data Recoding

O método *Semantic Data Recoding*, também conhecido como Generalização, é um método perturbativo aplicado a variáveis categóricas que combina várias categorias em novas categorias.

Este método foi criado por Guillermo Navarro-Arribas e Vicenç Torra [18] e é capaz de manter a estrutura hierárquica da variável sujeita a perturbação. Este método baseia-se na substituição de todos os valores de variáveis chave de uma observação pelos valores das variáveis chave de outra observação. Desta forma, obtém-se a menor perda de informação possível.

Para a aplicação deste método criam-se, em cada interação, dois critérios que asseguram a seleção do conjunto de observações mais adequado a transformar:

1. A partir da base de dados original, selecionar um conjunto de K combinações de variáveis chave com o menor número de frequências;
2. Para cada combinação k_1 de K, calcula-se a combinação mais semelhante a k_1, k_2, na base de dados original.

O objetivo do primeiro critério, é de começar o processo com as observações que não cumpram a condição de *K-anonymity*, para um dado valor de k. O objetivo do segundo critério, é reduzir a perda de informação resultante de cada substituição. Para realizar a substituição, um operador de comparação semântica é definido. Como resultado da substituição, os valores das variáveis chave da combinação k_1 e da combinação k_2, assumem o mesmo valor.

Assim o algoritmo para a aplicação deste método é o seguinte:

1. Selecionar a combinação k_1 com o menor valor de *frequency count*, f_{k_1};
2. Selecionar a combinação k_2, mais próxima da combinação k_1, com o menor valor de *frequency count*;
3. Os valores originais da combinação k_1 são substituídos pelos valores da combinação k_2, aumentando assim a *frequency count* da combinação k_2;
4. Regressar ao passo 1.

Quando no passo 2 se obtém o valor mínimo estipulado para a condição de *K-anonymity*, o algoritmo é interrompido.

Vários testes efetuados a este método perturbativo permitem concluir que as variáveis perturbadas obtidas apresentam maiores similaridades do que quando é aplicado o método PRAM [18].

4.2.2 Métodos Perturbativos para Variáveis Numéricas

Na aplicação dos métodos perturbativos para variáveis numéricas é importante verificar se o processo de perturbação nos dados originais provocou diferenças significativas nas suas características, ou seja, se a matriz de covariâncias e o vetor das médias apresentam diferenças significativas entre os dados originais e os dados perturbados.

Métodos Perturbativos com base em Modelos Lineares

A perturbação de uma base de microdados através de modelos lineares baseia-se geralmente na adição ou subtração de um ruído aleatório ou estocástico aos valores originais. Desta forma, pretende-se proteger os dados de correspondência exata com ficheiros externos.

Modelo Aditivo de Ruído Independente

Este modelo pode ser descrito da seguinte forma [24]:

$$Y_p = X_p + \epsilon_p, \tag{4.2}$$

onde X_p representa o vetor das observações da variável p da base de dados original, Y_p é o vetor das observações da variável p na base de dados perturbada e ϵ_p (ruído branco) representa o vetor das observações do ruído que segue uma distribuição Normal, e são independentes entre si, isto é, a covariância, $Cov(\epsilon_l, \epsilon_k) = 0, \forall k \neq l$.

Em notação matricial,

$$Y = X + \epsilon, \tag{4.3}$$

onde $\epsilon \sim N(0, \Sigma_\epsilon)$ e $\Sigma_\epsilon = c \cdot diag(\sigma_1^2, \sigma_2^2, ..., \sigma_P^2)$, para um dado parâmetro $c > 0$ que define a magnitude de ruído adicionado, e cada σ_p^2 representa a variância de cada variável X_p, $p = 1, ..., P$.

Na aplicação deste modelo, o valor médio e a variância de cada variável são preservados, no entanto, as covariâncias e as correlações entre as variáveis sofrem alterações. O aumento ou diminuição de ruído é discutível, pois depende dos diversos cenários legais, da sensibilidade dos dados, do risco de identificação e da perda de informação pretendidos na divulgação dos dados.

Modelo Aditivo de Ruído Correlacionado

Este modelo apresenta vantagens em comparação ao modelo apresentado no ponto anterior, pois neste caso a matriz de covariâncias dos erros é proporcional à matriz de covariâncias das variáveis originais, ou seja $\epsilon \sim N(0, \Sigma_\epsilon = c\Sigma)$, em que Σ é a matriz de covariâncias de X. Assim, as covariâncias e as correlações existentes entre variáveis não se alteram [3].

O Modelo Aditivo de Ruído Correlacionado proposto por J. Kim [12], pode ser definido como:

$$Y = \beta_0 + \beta_1 X + \epsilon. \tag{4.4}$$

J. Kim [12] provou ainda que qualquer modelo de regressão aplicado a Y e a X conduz a resultados semelhantes. Este modelo provoca um risco de identificação inferior ao modelo aditivo apresentado anteriormente.

P. Tendick e N. Matloff [27] propuseram uma modificação a este modelo, de forma a maximizar a relação entre o risco de identificação e a perda de informação:

$$Y = (1 + c)^{-\frac{1}{2}}(X + \epsilon), \tag{4.5}$$

onde c representa a magnitude de ruído a adicionar. Com estas especificações a matriz de covariâncias de X e Y são idênticas, $\Sigma_Y = \Sigma_X$, provocando assim menor perda de informação.

Modelo Aditivo Geral para a Perturbação dos Dados (GADP)

Os modelos aditivos de ruído possuem diversas desvantagens relacionadas com a perda de informação nas bases de microdados perturbadas. Os modelos lineares apresentados não preservam certas relações entre as variáveis originais, e ainda, condicionam o responsável da base de dados a ter em consideração o equilíbrio entre o risco de identificação e a perda de informação. De seguida, propõe-se um novo modelo aditivo de perturbação capaz de resolver os problemas associados aos modelos apresentados até ao momento.

A variância conjunta das variáveis confidenciais (X), não confidenciais (S) e perturbadas (Y) é representada da seguinte forma:

$$\Sigma^G = \begin{bmatrix} \Sigma_{XX} & & \\ \Sigma_{XS} & \Sigma_{SS} & \\ \Sigma_{XY} & \Sigma_{SY} & \Sigma_{YY}, \end{bmatrix} \tag{4.6}$$

onde Σ_{XX} é a matriz de covariâncias das variáveis de X, Σ_{XS} é a matriz de covariâncias entre X e S, Σ_{SS} é a matriz de covariâncias de S, Σ_{SY} é a matriz de covariâncias entre S e Y, Σ_{XY} é a matriz de covariâncias entre X e Y e Σ_Y é a matriz de covariâncias de Y.

O modelo GADP pode ser especificado da seguinte forma [21]:

$$Y = \beta_0 + \beta_1 S + \epsilon \tag{4.7}$$

onde $\beta_0 = \mu_X - \Sigma_{XS}\Sigma_{SS}^{-1}\mu_S$, $\beta_1 = \Sigma_{XS}\Sigma_{SS}^{-1}$, $\epsilon = \Sigma_{XX} - \Sigma_{XS}\Sigma_{SS}^{-1}\Sigma_{SX}$, onde μ_X é o vetor de valores médios das variáveis de X e μ_S é o vetor de valores médios das variáveis de S.

Com estas especificações de parâmetros, Y apresenta o mesmo vetor de valores médios e a mesma matriz de covariâncias que X. A matriz de covariâncias entre Y e S é a mesma que a matriz de covariâncias entre X e S. Quando as variáveis de X e S apresentam uma distribuição multivariada normal, os valores de Y são gerados a partir da distribuição condicional de $X|S$.

Uma das grandes vantagens do método GADP é que representa a forma mais geral de adição de ruído, ou seja, todos os outros modelos aditivos de ruído podem ser definidos como casos especiais do método GADP.

K. Muralidhar e R. Sarathy [21] afirmam um das grandes vantagens deste processo como:

"*A perturbação é baseada não só nas variáveis confidenciais X, mas também nas variáveis não confidenciais S, o que permite, se necessário, estabelecer a relação entre Y e S.*"

Para concluir, quando as variáveis confidenciais sujeitas a perturbação possuem uma distribuição normal, o método GADP fornece utilidade máxima na informação presente na base de microdados perturbada e risco de identificação mínimo. É aconselhável a aplicação deste método apenas a bases de microdados de grandes dimensões, pois estudos realizados provaram que a perda de informação aumenta significativamente com a diminuição da dimensão da base de microdados.

Modelo Aditivo Exato para a Perturbação dos Dados (EGADP)

De forma a resolver o problema existente no método GADP, isto é, o aumento de perda de informação com a diminuição da dimensão da base de microdados, Burridge [4] criou o modelo *Information Preserving Statistical Obfuscation* (IPSO). Este modelo gera o ruído de forma que o vetor dos valores médios e matriz de covariâncias entre Y e S seja idêntica à matriz de covariâncias entre X e S, e o erro de perturbação seja zero, para qualquer dimensão da base de microdados.

A aplicação do modelo IPSO poderia resultar em riscos de identificação elevados. De forma a provocar resultados mais eficazes do ponto de vista do risco de identificação, Muralidhar e Sarathy [21] sugerem uma modificação a este modelo, o modelo EGADP. Seja ρ a correlação entre X e S. A proporção de variabilidade de X explicada por S é dada por $R^2_{X|S} = \rho^2$ e a proporção de variabilidade de X explicada por S e Y é $R^2_{X|S,Y} = \rho^2 + \frac{(1-\rho^2)^2}{(1-\rho^2)+\sigma^2}$, onde σ^2 representa a variância de ϵ.

Através da geração de um ruído ortogonal a X, este modelo assegura que $R^2_{X|S,Y} = R^2_{X|S}$. O procedimento para a implementação do modelo EGADP é o seguinte [20]:

1. Estimar um modelo de regressão para X a partir de S, da seguinte forma:

$$X = \hat{\beta}_0 + \hat{\beta}_1 S,$$

   em que $\hat{\beta}_0 = \bar{X} - \hat{\Sigma}_{XS}\hat{\Sigma}^{-1}_{SS}\bar{S}$, $\hat{\beta}_1 = \hat{\Sigma}_{XS}\hat{\Sigma}^{-1}_{SS}$, onde $\bar{X}$ é o vetor das médias das variáveis de X e $\bar{S}$ é o vetor das médias das variáveis de S; Calcular a matriz de covariâncias dos resíduos, $\hat{\Sigma}_\epsilon$;

2. Gerar uma matriz aleatória A, (N x K), a partir de uma distribuição normal multivariada estandardizada, onde K representa o número total de variáveis chave e N o número total de observações;

3. Criar um modelo de regressão para A a partir de S e de X. Sejam B os resíduos do modelo de regressão criado, de dimensão (N x K), onde o resíduo da observação i é representado por b_i. A matriz B é ortogonal a X e a S com vetor de médias igual a zero;

4. Calcular a matriz de covariâncias de B, $\hat{\Sigma}_B$;

5. Calcular a nova variável C onde a observação i é dada por: $c_i = \hat{\Sigma}_\epsilon^{\frac{1}{2}}\hat{\Sigma}_B^{-\frac{1}{2}}b_i$, $i = 1, 2, ..., N$;

6. Por fim, calcular $y_i = \hat{\beta}_0 + \hat{\beta}_1 s_i + c_i$, $i = 1, 2, ..., N$, onde y_i representa as observações das variáveis perturbadas e s_i é o vetor dos valores das variáveis de S na observação i.

Após a aplicação do método EGADP é possível afirmar que:

- $\mu_X = \mu_Y$, o vetor dos valores médios das variáveis originais é idêntico ao vetor dos valores médios das variáveis perturbadas;
- $\Sigma_{YY} = \Sigma_{XX}$, a matriz de covariâncias dos valores perturbados é idêntica à matriz de covariâncias dos valores originais;
- $\Sigma_{YS} = \Sigma_{XS}$, a matriz de covariâncias entre Y (variáveis confidenciais perturbadas) e S (variáveis originais não confidenciais) é idêntica à matriz de covariâncias entre X (variáveis originais confidenciais) e S (variáveis originais não confidenciais);
- $R^2_{X|S,Y} = R^2_{X|S}$, a proporção de variabilidade de X explicada por S e Y é idêntica à proporção de variabilidade de X explicada por S.

Assim, este modelo permite obter uma base de dados perturbada onde inferências baseadas em modelos lineares não sofrem alterações e o risco de identificação é o menor possível.

Métodos Perturbativos com base em Modelos Não Lineares

Os métodos perturbativos com base em modelos não lineares, são métodos capazes de incorporar modelos multiplicativos, ou então combinar mais que um processo de perturbação.

Modelo de Ruído Multiplicativo

Um dos maiores problemas com o Modelo de Ruído Aditivo é de que nos valores das variáveis originais próximos de zero provoca elevadas perturbações e nos valores mais afastados de zero provoca menor perturbação. De forma a combater este problema, foi criado o modelo de ruído multiplicativo.

Gera-se uma a matriz de perturbação Z, com média igual a 1 e desvio padrão $\sigma_Z > 0$. A matriz Y dos dados perturbados é obtida da seguinte forma [15]:

$$Y = X \bullet Z, \tag{4.8}$$

onde $\bullet$ refere-se ao produto de Hadamard, definido como:

$$y_{ip} = x_{ip} \cdot z_{ip}, \tag{4.9}$$

onde $i = 1, ..., N$, $p = 1, ..., P$.

Este modelo apresenta vantagens relativamente ao modelo anterior, pois na existência de valores nulos, o ruído multiplicativo não provoca alteração nesse valor, o que permite uma maior utilidade para utilizadores de determinadas bases de microdados, como por exemplo, em bases de dados financeiras, onde é importante não causar alterações em valores nulos devido à sua importância do ponto de vista da análise.

O ruído multiplicativo pode ainda ser aplicado como ruído multiplicativo logarítmico. Neste caso, é aplicado uma alteração logarítmica aos dados originais X,

$$W_p = ln(X_p), \tag{4.10}$$

onde X_p representa o vetor associado à variável original p [15].

O ruído multiplicativo pode agora ser visto como uma adição de ruído:

$$Y_p = W_p + \epsilon_p, \tag{4.11}$$

onde Y_p representa a variável perturbada p, W_p definido em (4.10) e ϵ_p é o ruído aleatório gerado a partir da distribuição exponencial.

Data Shuffling

Data Shuffling é um método criado por Sarathy e Muralidhar [23]. Este procedimento é bastante eficaz, pois resulta em riscos de identificação baixos e fornece a máxima utilidade na informação da base de microdados perturbada, visto que preserva todas as relações existentes entre as variáveis originais [23].

O método de *Data Shuffling* pode ser descrito pelos seguintes passos [21]:

1. Calcular a matriz de correlação de Spearman, R, para as variáveis originais confidenciais;
2. Ordenar as observações da matriz S de variáveis não confidenciais;
3. Gerar as novas variáveis das matrizes S^* e X^* da seguinte forma:

$$x_{ip}^{*} = \phi^{-1}(F_{Xp}(x_{ip})), p = 1, ..., P,$$
$$s_{iq}^{*} = \phi^{-1}(F_{Sq}(s_{iq})), q = 1, ..., Q;$$

onde F_{X_p} e F_{S_q} representam a função cumulativa de densidade das variáveis confidenciais e não confidenciais, X_p e S_q respetivamente e $\phi^{-1}(.)$ representa o inverso da distribuição normal estandardizada.

4. Calcular a matriz de correlação de Pearson, ρ, com base na matriz de correlação R. A densidade conjunta de X^* e S^* é descrita como uma distribuição multivariada normal estandardizada com matriz de correlação ρ. A relação entre a matriz de correlação ρ e R pode ser descrita da seguinte forma:

$$\rho_{ip} = 2sin(\frac{\pi r_{ip}}{6}), \tag{4.12}$$

onde r_{ip} representam os elementos de R.

5. Aplicar o modelo GADP nas matrizes X^* e S^* de forma a gerar Y, da seguinte forma:

$$Y = \rho_{X^*S^*}\rho_{S^*S^*}^{-1}S^* + \epsilon, \tag{4.13}$$

onde $\epsilon \sim MVN(0, \rho_{X^*S^*}\rho_{S^*S^*}^{-1}\rho_{S^*X^*})$

6. Obtém-se a base de microdados perturbada a partir da substituição do valor perturbado y_{ip} pelo valor original x_{ip} com o mesmo índice de ordenação.

Na aplicação deste modelo, os valores perturbados são os valores originais mas atribuídos a diferentes observações, e as distribuições univariadas das variáveis perturbadas não diferem significativamente das distribuições univariadas das variáveis originais [23]. Quanto ao risco de identificação este método garante o menor risco de identificação possível [20].

A aplicação deste procedimento em R é possível e é descrito no Capítulo 7.

Outros Métodos Perturbativos

De seguida apresentam-se alguns métodos perturbativos que não têm por base um modelo de regressão. São métodos que mascaram a informação confidencial presente na base de microdados original, permitindo assim a divulgação destes dados.

Random Orthogonal Matrix Masking (ROMM)

ROMM é um novo método de perturbação de adição de ruído para a proteção de atributos confidenciais em variáveis contínuas [6]. Usualmente ROMM é aplicado a bases de microdados e o seu procedimento pode ser descrito nos seguintes passos:

1. Geração de uma matriz aleatória ortogonal, T, a partir de uma distribuição G definida no grupo de $N * N$ matrizes ortogonais, que não provocam alterações no vetor 1_N , isto é, $T \cdot 1_N = 1_N$, onde 1_N é um vetor constituído por N observações com valor igual a 1;

2. Aplica-se a matriz ortogonal T às variáveis confidenciais, X, resultando na matriz Y dos dados perturbados, $Y = T \cdot X$;

3. Por fim, revela-se aos utilizadores:

    - As variáveis perturbadas, Y;
    - A informação de que os dados foram obtidos a partir de um operador ortogonal gerado aleatoriamente a partir da distribuição G;
    - A distribuição exata de G.

Este método apresenta diversas vantagens:

- Preserva os valores médios e as covariâncias;
- Preserva as propriedades multivariadas de variáveis que possuem uma distribuição normal;
- Preserva as estimativas dos parâmetros de modelos de regressão;
- Controla a magnitude da perturbação e desta forma, análises poderão ser feitas às variáveis perturbadas mesmo que as variáveis não sigam uma distribuição normal.

A preservação de valores médios e de covariâncias apresentam diversas vantangens tanto a nível prático como a nível teórico. A nível prático, modelos de regressão são preservados, sem que haja qualquer alteração nos seus parâmetros. A nível teórico a distribuição das variáveis originais com distribuição Normal é mantida inalterada. Por fim, o método ROMM permite o controlo da magnitude do ruído perturbativo, que é possível através da escolha exata e apropriada da distribuição G [6].

Microagregação

Microagregação é um método perturbativo normalmente aplicado a variáveis numéricas, no entanto, a sua aplicação pode ser alargada às variáveis categóricas. É um método natural para o cumprimento da condição de *K-anonymity*.

Em geral, este método começa por dividir as observações em partições, de seguida calcula uma medida estatística de cada partição (normalmente a média) e cada valor individual é substituído pela medida estatística calculada da partição a que esta pertence [24].

A eficácia deste método depende da escolha das partições, ou seja, obtém-se resultados eficazes quando as observações agrupadas são homogéneas, pois assim a medida estatística calculada não difere significativamente do valor original pela qual esta será substituída.

No caso univariado e com variáveis categóricas ordinais, a formação de partições realiza-se inicialmente com a ordenação dos valores, e de seguida a constroem-se G grupos com dimensão N_g, $g = 1, ..., G$, maximizando a homogeneidade dentro de cada partição.

A homogeneidade de cada grupo é medida através da Soma dos Quadrados dos Erros:

$$SSE = \sum_{g=1}^{G} \sum_{i=1}^{N_n} (x_{ig} - \bar{x}_g)^2, \tag{4.14}$$

onde x_{ig} é a observação i da partição g e $\bar{x}_g$ é a medida estatística calculada para a partição g. Quanto menor o valor da Soma dos Quadrados dos Erros, maior será a homogeneidade de cada partição. Para o caso multivariado,

é possível a realização de microagregação univariada a cada variável separadamente. Na prática, este processo provoca elevado risco de identificação e por isso, foi necessário a criação do método de microagregação multivariada.

Num caso multivariado, o processo baseia-se em dois passos: primeiro constroém-se grupos homogéneos e de seguida os valores de todas as observações de cada variável sujeita a perturbação são substituídos pelos representantes do respetivo grupo.

Este processo apresenta várias alternativas sobre como construir as partições e como fazer a substituição dos valores. Apresentam-se alguns métodos que permitem a obtenção de resultados eficazes.

- **Método do Ranking Individual:** Este processo substituí os valores originais pela sua medida estatística, coluna a coluna e de maneira independente. Inicialmente, a primeira coluna é ordenada por ordem cresceste e o índice de ordenação é memorizado, assim é possível uma re-ordenação dos valores, obtendo os valores originais a partir dos valores perturbados. De seguida, os primeiros k valores são substituídos pela sua medida estatística, os próximos k valores são também substituídos pela sua medida estatística, e repete-se este processo até à última observação da primeira variável. Por fim, a variável é desorganizada e os valores são atribuídos às respetivas observações, e executa-se este procedimento para todas as variáveis;

- **Microagregação com base em Análise de Componentes Principais:** Este método organiza os dados em função da primeira componente principal [24]. Inicialmente a primeira componente principal é estimada e com base nesta componente realizam-se as sucessivas partições. Esta alternativa é eficaz sempre que a primeira componente principal explica uma percentagem significativa da variância das variáveis em estudo, caso contrário, é recomendada a aplicação de outros métodos;

- **Microagregação por Máxima Distância ao Valor Médio (MDVM):** Este método agrupa as observações com base na distância euclidiana num espaço multivariado. O processo baseia-se nos seguintes passos:

  1. O centro dos dados é estimado utilizando os valores médios de cada coluna (variável) da matriz dos dados. Obtém-se assim o vetor centróide C que contém os valores médios de todas as variáveis;
  2. Seleciona-se uma observação x_{ip} à maior distância euclidiana de C;
  3. Constrói-se um grupo de k observações em torno de x_{ip}, formado por $k-1$ observações com a menor distância euclidiana de x_{ip};
  4. Seleciona-se uma outra observação x_{ip_1}, com a maior distância euclidiana de x_{ip};
  5. Dentro das observações restantes, selecionam-se $k-1$ observações mais próximas de x_{ip_1};
  6. Os passos anteriores são repetidos para as restantes observações. Quando existirem $2k-1$ ou menos observações para serem agrupadas, o processo é interrompido e a última partição é constituída pelas observações em falta.

- **Microagregação pela distância de Mahalanobis:** O método por MDVM foi alterado de forma a produzir resultados mais eficazes. Em vez da distância euclidiana passará a ser utilizada a distância de Mahalanobis, mas todo o processo é implementado da mesma forma.

Rank Swapping

Embora este método tenha sido concebido inicialmente apenas para variáveis ordinais, *Rank Swapping* pode ser aplicado a qualquer variável numérica [7]. Este método é baseado em alterações internas dos valores de uma certa variável confidencial, X_p, ao longo das observações.

Primeiro, os valores da variável X_p são ordenados por ordem ascendente. De seguida, cada valor ordenado de X_p é trocado por outro valor ordenado, aleatoriamente escolhido a uma distância d, de a modo que a ordem de dois valores trocados entre si não pode estar afastada mais que d por cento do número total de observações [7]. Caso a vizinhança da observação i na variável p (x_{ip}), contenha reduzida heterogeneidade, o valor x_{ip} pode ser substituído por um valor bastante semelhante. Portanto, para que o risco de identificação não seja demasiado elevado, é necessário existir um número significativo de categorias, no caso de variáveis categóricas, ou de pelo menos m observações num intervalo especificado, para o caso das variáveis numéricas. Se este método é aplicado num espaço multivariado a estrutura de correlação dos dados não é alterada.

No entanto, se um utilizador obtém acesso ao valor máximo ou ao valor mínimo da variável em estudo, a identificação de um sujeito após a aplicação de *Rank Swapping* é possível pois os valores das variáveis não são alterados, apenas trocados da sua ordem original.

Shuffling

Shuffling é um método semelhante ao *Rank Swapping*, no entanto utiliza modelos de regressão no processo de perturbação, ou seja, as variáveis não confidenciais funcionam como variáveis explicativas na determinação de novos valores das variáveis confidenciais sujeitas a perturbação.

A ideia presente neste método é, inicialmente organizam-se as observações com base nas variáveis originais. De seguida, constrói-se um modelo de regressão onde as variáveis a serem protegidas serão as variáveis respostas e o conjunto de variáveis correlacionadas com estas serão as variáveis explicativas. Este modelo gera então N observações para cada variável sujeita a perturbação. A observação predita para o indivíduo i na variável p representa-se por y_{ip}. Por fim, os valores gerados serão ordenados e a observação x_{ip} é substituída pela observação x_{i_1p}, $i_1 \neq i$, com a mesma ordem da observação y_{ip}. Desta forma, os dados perturbados apenas contêm valores da base de dados original.

Re-amostragem

O método de Re-amostragem foi inicialmente proposto para a proteção de dados tabulares, no entanto, a sua aplicação pode ser alargada para a proteção de microdados [11].

Seja X_p a variável original p num conjunto de dados e T amostras independentes, $X_p^1, ..., X_p^T$. Todas as amostras são ordenadas usando o mesmo critério de classificação. De seguida, é criada uma variável perturbada Z_p constituída por $\bar{x}_{11}^1, ..., \bar{x}_{NP}^T$, onde $\bar{x}_{ip}^t$ representa a média da amostra t à qual a observação i da variável p pertence, ou seja, os valores originais são substituídos pela média amostral de cada amostra.

Este método apresenta diversas desvantagens quanto à utilidade da informação na base de dados perturbada. O detalhe da base de dados original não é mantido na substituição dos valores originais pela média amostral. Como as amostras são formadas aleatoriamente, é possível obterem-se valores perturbados muito distantes dos valores originais, conduzindo a uma elevada perda de informação.

Arredondamento

Arredondamento é um método que consiste na substituição dos verdadeiros valores por valores arredondados. No caso multivariado, o arredondamento é realizado variável a variável.

Seja X_p uma variável contínua sujeita a perturbação, primeiro é necessário criar um conjunto de pontos de arredondamento definido como $p_1, ..., p_N$, onde cada $p_i = b \cdot i$, para $i = 1, ..., N$ e b um valor base definido pelo responsável da base de dados. Após o conjunto de pontos definido é preciso agora um conjunto de atração para cada p_i, definido como $[p_i - b/2; p_i + b/2]$, com $i = 2, ...N - 1$. Para p_1 e p_N esse conjunto é dado como $[0, p_1 + \frac{b}{2}]$ e $[p_N - \frac{b}{2}, X_{imax}]$, respetivamente, onde X_{imax} representa o maior valor possível da variável X_p. Assim a observação x_{ip} é substituída pelo valor arredondado p_i correspondente ao conjunto de atração onde x_{ip} está contido [7].

Uma das grandes desvantagens deste método é a perda de detalhe na base de dados perturbada, dificultando assim a possibilidade de análises eficazes por parte dos utilizadores.

Distorção dos Dados através da Distribuição de Probabilidade

É um método aplicável tanto a variáveis numéricas como a variáveis categóricas. Na sua aplicação serão necessários três passos [7]:

1. Identificação da função densidade (função de probabilidade) de cada variável confidencial/sensível e a estimação dos seus parâmetros. Através da aplicação do teste de Kolmogorov-Smirnov é possível identificar a função densidade (função de probabilidade) das variáveis em estudo;

2. Geração de um conjunto de dados perturbados para cada variável confidencial. Neste passo, são gerados novos valores através dos parâmetros e da função distribuição obtidos na primeira etapa;

3. Substituir os dados confidenciais pelos dados perturbados. Na substituição desses valores é necessário a ordenação dos dados perturbados e dos dados originais, substituindo-se cada observação original por uma observação perturbada da mesma ordem.

Este procedimento apenas poderá ser aplicado a uma variável de cada vez e caso sejam utilizadas funções de densidade multivariadas, o processo poderá conduzir a resultados pouco eficazes. Outra desvantagem é a possibilidade de os dados não possuírem uma distribuição bem definida, conduzindo assim à rejeição de todas as distribuições aplicadas no teste de Kolmogorov-Smirnov, o que provocaria um aumento no tempo de execução do procedimento e poderia conduzir a resultados com um nível elevado de perda de informação [7].

Mapeamento Inverso de Dados Perturbados

De seguida, apresenta-se um método a aplicar em bases de microdados perturbadas antes da sua divulgação. Este método realiza uma transformação simples nos valores das variáveis perturbadas através de uma permutação dos valores das variáveis originais, utilizando mapeamento inverso com base nos índices de ordenação [16]. Sejam $X_p = \{x_{1p}, ..., x_{Np}\}$ os valores da variável confidencial p da base de dados original e $Y_p = \{y_{1p}, ..., y_{Np}\}$ representa a versão perturbada dos valores da variável X_p. Não existe qualquer suposição sobre o método utilizado para obter Y_p, mas assume-se que os valores de X_p e Y_p podem ser ordenados de alguma forma. O conhecimento dos valores de X_p e Y_p permite obter a variável Z_p através de mapeamento inverso da seguinte forma:

- Ordenar os valores de Y_p, e a cada observação y_{ip} atribuiu-se o índice de ordenação i;
- Calcular $z_{ip} = x_{ip}$, onde x_{ip} representa o valor da variável X_p com índice de ordenação i;
- Realizar o passo anterior para as N observações, obtendo-se a variável $Z_p = \{z_{1p}, ..., z_{Np}\}$, onde cada z_{ip} representa o valor original correspondente ao índice de ordenação dos valores perturbados.

Desta forma, é claro de perceber que a variável Z_p apresenta exatamente os mesmos valores da variável original X_p, mas estes valores foram permutados de acordo com os índices de ordenação da variável perturbada Y_p. A extensão deste método para o caso multivariado não apresenta qualquer problema, basta realizar o procedimento para as P variáveis confidenciais.

A divulgação da variável perturbada por mapeamento inverso, Z_p, em vez da variável perturbada, Y_p, contém diversas vantagens:

- O mapeamento inverso não provoca qualquer alteração na estrutura de correlação das variáveis perturbadas, portanto as relações existentes entre as variáveis perturbadas permanecem inalteradas;
- Através de mapeamento inverso, as distribuições marginais das variáveis originais são preservadas. Assim, a perda de informação resultante do processo de mapeamento inverso não é maior que a perda de informação resultante do processo de perturbação;
- O risco de identificação pode ser medido através do coeficiente de correlação de Spearman entre X e Z, e quando maior o coeficiente maior o risco de identificação.
- A perda de informação pode ser medida através da diferença entre os coeficientes de correlação, isto é:

$$\rho_{X,S} - \rho_{Z,S}, \tag{4.15}$$

onde $\rho_{X,S}$ e $\rho_{Z,S}$ representam o coeficiente de correlação de Spearman entre X, S e Z, S, respetivamente.

4.2.3 O Paradigma de Permutação

O paradigma da permutação é uma recente contribuição na literatura de CDE e propõe uma função geral de equivalência baseada em permutações, capaz de descrever qualquer tipo de método de CDE. Com base em mapeamento inverso Domingo-Ferrer e Muralidhar [8] mostram que todo o processo de perturbação para uma base de microdados pode ser descrito como uma permutação que pode ser complementada com uma pequena adição de ruído.

Para ilustrar esta função, utiliza-se um exemplo [22]. Sejam $X = (X_1, X_2, X_3)$ três variáveis originais. A estas variáveis aplica-se um método de CDE de forma a obter valores perturbados $Y = (Y_1, Y_2, Y_3)$. De seguida, ordenam-se os valores das variáveis originais e das variáveis perturbadas, X e Y, como apresentado nas Tabelas 3 e 4.

Assim é sempre possível obter uma base de microdados Z que contém as variáveis X_1, X_2 e X_3 ordenadas de acordo com os índices de ordenação das observações de Y [8], como mostra a Tabela 4. Por fim, as variáveis perturbadas poderão ser reconstituídas através da adição de ruído com baixa magnitude, $E = (E_1, E_2, E_3)$, como é apresentado na Tabela 4. Desta forma Z possuí a mesma distribuição marginal que X.

O paradigma da permutação estabelece a permutação como o prícipio de qualquer método de perturbação, permitindo que este seja visto como o resultado de uma função de equivalência definida da seguinte forma [22]:

Para uma base de microdados X com N observações e P variáveis confidenciais, a sua versão perturbada Y pode ser sempre escrita, independentemente do método de CDE aplicado, como:

$$Y = Z + E \text{ em que } Z = (P_1 X_1, ..., P_P X_P), \tag{4.16}$$

onde Z é uma matriz com P variáveis e N observações, $P_1 = A_1^T D_1 A_1, ..., P_P = A_P^T D_P A_P$ representam um conjunto de P matrizes de permutação e E uma matriz de ruídos com baixa magnitude. $A_1, ..., A_P$ representam um conjunto de matriz que ordena as variáveis de ordem crescente, $A_1^T, ..., A_P^T$ um conjunto de matrizes que colocam a variável na ordem original, e $D_1, ..., D_P$ um conjunto de matrizes para a perturbação dos dados, de acordo com o método de CDE a utilizar.

Tabela 3: Exemplo do paradigma da permutação

Dados Originais X		
X_1	X_2	X_3
13	135	3707
20	52	826
2	123	-1317
15	165	2419
29	160	-1008
Indice de ordenação das variáveis		
X_1	X_2	X_3
4	3	1
2	5	3
5	4	5
3	1	2
1	2	4

Dados Perturbados Y		
Y_1	Y_2	Y_3
8	160	3248
20	57	822
-1	122	248
18	135	597
29	164	-1927
Indice de ordenação das variáveis		
Y_1	Y_2	Y_3
4	2	1
2	5	2
5	4	4
3	3	3
1	1	5

Tabela 4: Exemplo da aplicação de Mapeamento Inverso

Dados Originais X		
X_1	X_2	X_3
13	135	3707
20	52	826
2	123	-1317
15	165	2419
29	160	-1008
Dados Mapeados Inversamente Z		
Z_1	Z_2	Z_3
13	160	3707
20	52	2419
2	123	-1008
15	135	826
29	165	-1317

Ruído E		
E_1	E_2	E_3
-5	0	-459
0	5	-1597
-3	0	1256
0	-1	-229
0	-1	-610
Dados Perturbados $Y = Z + E$		
Y_1	Y_2	Y_3
8	160	3248
20	57	822
-1	122	248
18	135	597
29	164	-1927

Assim, os métodos de proteção de microdados podem ser vistos como uma função de equivalência para um conjunto de matrizes de permutação. Procendendo de variável em variável, primeiro os dados são permutados de forma a aparecer em ordem crescente, de seguida a chave de perturbação é inserida, e finalmente os dados são reordenados para a forma original.

4.3 Perturbação de Dados Longitudinais

A perturbação de dados longitudinais apresenta alguns cuidado adicionais, pois a informação individual é acumulada ao longo do tempo, permitindo assim mais cenários de identificação. Apesar da literatura de CDE oferecer uma vasta

variedade de ferramentas e processos para a proteção dos dados, poucos desses se focam na proteção de dados longitudinais. Neste capítulo, apresentam-se algumas métricas propostas por Ruiz [22] de forma a implementar os métodos existentes na proteção destes tipos de dados.

Os dados variam entre dados transversais, onde indivíduos são observados num ponto único de tempo, e dados temporais, onde uma entidade é observada ao longo de um espaço de tempo. No caso de base de microdados financeiras ou admnistrativas, os microdados possuem uma estrutura longitudinal, isto é, são bases de microdados que possuem várias observações para a mesma entidade. Os investigadores, principalmente os da área de economia, referem-se muitas vezes a estas estruturas de dados como dados em painel. Os dados longitudinais apresentam maior detalhe que os dados transversais, no entanto possuem um problema, caso as entidades não sejam observadas nos mesmos períodos de tempo, haverá uma grande quantidade de dados em falta, ou seja são dados não balanceados. Neste capítulo, considera-se que os dados são sempre balanceados, ou seja, em qualquer tempo t existem sempre N observações.

Como já referido a proteção destes dados apresenta alguns desafios. Em capítulos anteriores foi referido que o risco de identificação depende do conhecimento de um utilizador sobre as características de um indivíduo, e cenários de identificação podem ser traçados tendo em conta essa informação. Para os dados longitudinais, tal informação continua a permitir a possibilidade de um utilizador identificar um indivíduo. No entanto, a alteração da categoria ou valor de uma variável chave ao longo do tempo pode também permitir a identificação de um indivíduo. Por exemplo, caso um utilizador possua a informação de que um indivíduo mudou a sua situação de emprego, de desempregado no tempo t para empregado no tempo $t + k$, o conhecimento dessa alteração pode permitir a identificação do indivíduo em causa. Portanto, a situação de empregado pode ser considerada uma variável chave num estudo transversal e num estudo longitudinal, e a alteração da variável situação de emprego pode ser visto como uma variável chave num estudo longitudinal. Tal como as alterações em variáveis chave, alterações em variáveis confidenciais poderão ser vistas como variáveis chave também [22]. Assim, é percetível que a estrutura longitudinal dos dados aumenta a possibilidade de identificação de uma entidade. Quanto à utilidade da informação, como a principal caracteritisca destes dados é o detalhe presente nos mesmos, no processo de proteção este detalhe deverá ser mantido. O equilíbrio entre o risco de identificação e a perda de informação é ainda mais relevante para dados longitudinais, isto é, a informação presente na base de microdados é bastante detalhada, e ao mesmo tempo apresenta bastante informação confidencial.

4.3.1 Mapeamento Inverso de Dados Longitudinais

O responsável pela base de dados escolhe um método de CDE de forma a perturbar a base de dados original. O problema que se depara é de que as distribuições marginais são diferentes entre os dados originais e os dados perturbados. De forma a ultrapassar este problema o processo de mapeamento inverso assegura que tais distribuições são mantidas [16]. Neste Capítulo sugere-se um processo idêntico mas adaptado a dados longitudinais [22].

Primeiro começa-se por observar a relação entre duas variáveis durante o tempo e sobre o mesmo conjunto de observações. Uma variável observada em dois períodos de tempo t e $t + 1$ pode ser mapeada inversamente de tal forma a expressar a variável em $t + 1$ como uma função de si mesma em t. Este procedimento, geralmente, conduz a uma caracterização simples da informação essencial e dos riscos de identificação contidos nos dados longitudinais [22]. Este procedimento é equivalente a considerar o tempo como um método de perturbação, onde a variável em $t + 1$ é a versão perturbada da variável em t.

Seja $X_{p,t} = (x_{1,p,t}, ..., x_{N,p,t})$ e $X_{p,t+1} = (x_{1,p,t+1}, ..., x_{N,p,t+1})$ os valores da variável p no tempo t e

os valores da mesma variável em $t+1$. Esta variável pode assumir qualquer valor desde que seja possível a sua ordenação e N mantém-se cosntante ao longo do tempo, ou seja os dados são balanceados. O conhecimento dos valores de $X_{p,t}$ e $X_{p,t+1}$ permite exprimir uma função entre t e $t+1$ da seguinte forma [22]:

- Ordenar os valores da variável p em $t+1$, obtendo-se as observações $x_{k,p,t+1}$ com índice de ordenação k;
- Ordenar os valores da variável p em t, de acordo com as ordens em $t+1$, obtendo-se as observações $x_{k,p,t}$;
- Construir uma nova variável $Z_{p,t} = (z_{1,p,t}, ..., z_{N,p,t})$, onde cada $z_{i,p,t}$, corresponde ao valor ordenado $x_{k,p,t}$, com $i = 1, ..., N$ e $k = 1, ..., N$;

Neste procedimento, $Z_{p,t}$ corresponde à variável $X_{p,t}$ ordenada pelo índice de ordenação da variável $X_{p,t+1}$. $Z_{p,t}$ representa asssim a variável obtida pelo processo de mapeamento inverso, que expressa $X_{p,t+1}$ como uma permutação de $X_{p,t}$. Como os valores das variáveis podem alterar durante o tempo, particularmente no caso de variáveis numéricas, adiciona-se um ruído, $E_{p,t,t+1}$, que corresponde à diferença entre os valores de $X_{p,t+1}$ e $Z_{p,t}$. A adição de ruído ruído visa garantir que as distribuições marginais não se alteram pela aplicação desta metodologia. Desta forma, obtém-se a recomposição exata de $X_{p,t+1}$ como função de $X_{p,t}$. De seguida, é possível restringir a Equação (6.1) para o caso dos dados longitudinais da seguinte forma:

$$X_{p,t+1} = Z_{p,t} + E_{p,t,t+1}, \text{ em que } Z_{p,t} = Q_{T,p}X_{p,t} \text{ e } Q_{T,p} = C_{T,p}^{T}K_{T,p}C_{T,p}, \tag{4.17}$$

onde $Q_{T,p}$ representa a matriz de permutação que contém as alterações dos índices de ordenação ao longo do tempo, $C_{T,p}$ e $C_{T,p}^{T}$ provocam uma reordenanação na variável em estudo, e $K_{T,p}$ representa a chave temporal da variável p. As alterações ao longo do tempo serão sempre representadas na matriz $Q_{T,p}$, o que significa que a principal característica dos dados longitudinais pode ser representada pelas mesmas entidades para expressar qualquer método de perturbação.

Desta forma, $E_{p,t,t+1}$ caracteriza as alterações provocadas na distribuição da variável p ao longo do tempo, enquanto que $Q_{T,p}$ caracteriza os movimentos internos ao longo do tempo dos indivíduos na distribuição da variável p.

Pela Equação 4.16, é possível definir as versões perturbadas, por qualquer método de CDE, de $X_{p,t}$ e de $X_{p,t+1}$, representadas respetivamente por $Y_{p,t}$ e de $Y_{p,t+1}$ como:

$$Y_{p,t} = P_{p,t}X_{p,t} + E_{p,t}, \tag{4.18}$$

$$Y_{p,t+1} = P_{p,t+1}X_{p,t+1} + E_{p,t+1}, \tag{4.19}$$

onde $P_{p,t}$ e $P_{p,t+1}$ representam as matrizes que descrevem o processo de perturbação aplicado, de acordo com o paradigma de permutação, para a variável p nos tempos t e $t+1$, respetivamente, e $E_{p,t}$ e $E_{p,t+1}$ representam as matrizes de ruído com baixa magnitude.

Do ponto de vista da informação, é claro que a equação (4.17) terá de ser preservada para que a informação temporal dos dados longitudinais permaneça igual. Substituindo a equação (4.17) em (4.19), e de seguida utilizar a equação (4.18) para substituir $X_{p,t}$, obtém-se a seguinte expressão:

$$Y_{p,t+1} = P_{p,t+1}Q_{T,p}P_{p,t}^{T}Y_{p,t} + \left[P_{p,t+1}\left(E_{p,t,t+1} - Q_{T,p}P^{-1}E_{p,t}\right) + E_{p,t+1}\right], \tag{4.20}$$

Como resultado, se os dois processos de perturbação em t e em $t+1$ não alteram a informação temporal, deve se verificar, por comparação de (4.17) e (4.20), que:

$$P_{p,t+1}Q_{T,p}P_{p,t}^{T} = Q_{T,p}, \tag{4.21}$$

$$P_{p,t+1}\left(E_{p,t,t+1} - Q_{T,p}P^{T}E_{p,t}\right) = E_{p,t,t+1}, \tag{4.22}$$

As equações (4.21) e (4.22) descrevem como os dois processos de perturbação em t e em $t+1$ devem estar relacionados. Como ruídos de baixa magnitude são irrelevantes para descrever o funcionamento de um método de perturbação [22], pois o tais ruídos não provocam alterações nos índices de ordenação das observações, a equação (4.22) pode ser simplificada para:

$$P_{p,t+1}E_{p,t,t+1} = E_{p,t,t+1}. \tag{4.23}$$

Assim, de forma a verificar-se Equação (4.23), as matriz de permutação em $t+1$ tem ser a matriz identidade, e como consequência a matriz de permutção em t também teria de ser a matriz identidade. O que significa que nenhum método de perturbação foi aplicado. Para uma boa aplicação de um método perturbativo em dados longitudinais, existirá sempre perda de informação temporal nos dados, pois tais relações descritas não são possíveis na prática.

4.3.2 Perda de Informação e Risco de Identificação

Na perturbação de dados longitudinais o principal foco é a matriz de transposição $Q_{T,p}$, que descreve o efeito do tempo na variável p. Esta matriz contém a maioria da informação que deve ser protegida, sendo esta informação de elevado risco. Como resultado, qualquer alteração na matriz provoca uma diminuição no risco de identificação, e por outro lado aumenta a perda de informação. Neste contexto, Ruiz [22] apresenta uma medida capaz de calcular tanto o risco de identificação como a perda de informação numa base de microdados longitudinal, permitindo estabelecer o limiar desejado para ambas as medidas. Esta medida tem por base a equação (4.21) e calcula as diferenças presentes entre a matriz $P_{p,t+1}Q_{T,p}P_{p,t}^{T}$ e a matriz $Q_{T,p}$.

Sejam $r_{T,p}$ e $r_{A,p}$ os vetores com a informação sobre as trocas dos índices de ordenação das observações de $Q_{T,p}$ e de $P_{p,t+1}Q_{T,p}P_{p,t}^{T}$, respetivamente. $r_{T,A,p,i} = r_{T,p} - r_{A,p} = (r_{T,A,p,1}, ..., r_{T,A,p,N})$ representa o vetor das diferenças entre os N valores dos vetores $r_{T,p}$ e $r_{A,p}$ onde a variável p é observada em t e $t+1$.

$$J(\alpha) = \begin{cases} \left(\frac{1}{N}\sum_{i=1}^{N}|r_{T,A,p,i}|^{\alpha}\right)^{\frac{1}{\alpha}} & \text{para } \alpha \neq 0 \\ \prod_{i=1}^{N}|r_{T,A,p,i}|^{\frac{1}{N}} & \text{para } \alpha = 0 \end{cases} \tag{4.24}$$

.

O parâmetro α funciona como um limiar desejado para uma boa pertubação da base de microdados. Pelos testes realizadosm por Ruiz [22], quanto menor o valor de α, maior enfâse é dado às pequenas alterações de ordem. E quanto maior o valor de α maior efâse é dado a grandes alterações de ordem. Ou seja, se α tende para $-\infty$ então $J(\alpha)$ converge para o valor da menor alteração de ordem.

Esta estrutura forma uma classe de risco de identificação e de perda de informação na avalição de dados longitudinais perturbados. O principal objetivo consiste em medir a extensão da dissimilaridade introduzida na informação temporal pelo processo de perturbação [22]. Quando a perturbação dos dados é vista como uma permutação, apenas as alterações nas ordens interessam no cálculo das medidas, pois a informação presente nos

dados perturbados apresenta os mesmos valores que os dados originais, e como tal na medição da utilidade na informação apenas alterações nas ordens dos valores causam perda de informação.

4.4 Geração de Dados Sintéticos

A Geração de Dados sintéticos é uma solução para a divulgação de uma base de microdados sem que informação confidencial seja revelada [24]. Os dados são gerados de forma aleatória, preservando algumas estatísticas ou relações internas do conjunto de dados original [13]. Diversos processos e ferramentas foram desenvolvidas para este processo de divulgação. Existem três processos de maior importância:

- Reconstrução sintética;
- Optimização Combinatorial.
- Geração à base de modelos.

Nesta secção, o método mais importante e mais promissor será descrito, o método de geração à base de modelos, em particular a Imputação Múltipla.

Em comparação com os métodos tradicionais de CDE, os testes realizados com estes métodos mostram que os dados sintéticos contêm um risco de identificação inferior, no entanto, possuem menor utilidade de informação [24]. Estes métodos não tencionam substituir os métodos tradicionais e aconselha-se a sua aplicação apenas para casos em que o risco de identificação é demasiado elevado. O método à base de modelos de regressão é um processo flexível e diversificado. No entanto, ainda se encontra em desenvolvimento, várias alternativas e modificações foram realizadas de forma a provocar resultados mais equilibrados entre o risco de identificação e a perda de informação.

Imputação Múltipla

O processo de imputação múltipla pode ser definido da seguinte forma [1]. Considere-se uma base de microdados original W de dimensão n, provenientes de uma população de dimensão N, onde existem variáveis de suporte A, variáveis não confidenciais S e variáveis confidenciais X. As variáveis de suporte são observadas e disponíveis para todas as N observações, e S e X apenas estão disponíveis para as n observações da base de microdados W. O primeiro passo é construir, a partir de W, uma população imputada de N entidades. Esta população consiste nas n observações de W e M matrizes de (S, X) (o número de imputações múltiplas, normalmente entre 3 e 10), com as $N - n$ observações em falta. A variabilidade dos valores imputados assegura teoricamente que inferências válidas podem ser obtidas a partir da população gerada. Um modelo de previsão para (S, X) a partir de A é utilizado para a imputação múltipla de (S, X) na população. Assim que a população gerada esteja disponível, uma amostra Z de n^* observações pode ser obtida a partir dessa população, e a sua estrutura será idêntica à amostra retirada da população original de N observações. Realiza-se este procedimento M vezes de forma a criar M réplicas da matriz de (S, X). De forma a garantir que não é divulgada informação original, é possível limitar as amostras retiradas às $N - n$ observações geradas.

Em termos práticos, este método implica que os utilizadores tenham de efectuar a análise das m bases de microdados divulgadas para inferirem sobre a base de microdados original [20].

Na atualidade, o método de geração de dados sintéticos ainda se encontra em desenvolvimento e diversos métodos já foram criados, desde geração parcial de dados sintéticos a geração de uma base de microdados inteiramente sintética. Tais métodos não são referidos neste projeto, pois o objetivo desta dissertação não são os

métodos sintéticos. No entanto, apresenta-se um pequeno resumo das vantagens e desvantagens destes métodos. Os dados sintéticos são apelativos pelo facto de resolver o problema da identificação de um indivíduo, visto que as observações presentes na base de dados divulgada são os valores gerados e não os valores originais. No entanto, esta vantagem não é assim tão clara quanto parece, pois se por alguma razão os valores das variáveis não confidenciais de um indivíduo corresponder aos mesmos valores de uma outra base de dados pública, então a identificação é possível. Por outro lado, a utilidade de informação limitada é um problema, pois apenas as propriedades das estatísticas é que são preservadas a partir do modelo.

4.5 Comparação dos Métodos de CDE

A grande dificuldade na perturbação de uma base encontra-se em decidir qual o melhor método perturbativo a utilizar. Esta decisão depende da base de dados em estudo, e como tal não existe uma escolha exata, no entanto, teoricamente é possível comparar estes métodos de forma a perceber qual o(s) que funciona(m) melhor de acordo com a base de microdados em estudo.

4.5.1 Comparação de Modelos Lineares e Não Lineares de Ruído

De todos os modelos lineares descritos neste Capítulo, o modelo EGADP é o que fornece menor risco de identificação e ainda permite que todas as inferências realizadas na base de microdados perturbada sejam exatamente iguais à base de microdados original. No entanto, apresenta a desvantagem de que as distribuições marginais das variáveis perturbadas podem apresentar diferenças significativas das distribuições marginais das variáveis originais [20].

Quanto aos modelos não lineares, o modelo multiplicativo é o que provoca maior perda de informação. Por outro lado, o modelo *Data Shuffling* conduz a valores perturbados bastante semelhantes aos valores originais e com baixo risco de identificação. Desta forma, *Data Shuffling* é o que provoca resultados mais eficazes de todos os modelos não lineares.

É aconselhável a aplicação do modelo *Data Shuffling* quando a base de microdados:

- Possui uma dimensão relativamente grande;
- É utilizada principalmente para análises estatísticas não tradicionais, como por exemplo, *data mining*;
- Contém relações não lineares importantes.

Normalmente estas características estão presentes em bases de microdados com informação de negócios ou dados financeiros.

O modelo EGADP oferece uma vantagem significativa sobre o modelo *Data Shuffling*, que é o facto de independentemente da dimensão da base de microdados, inferências de modelos lineares nos dados perturbados conduzem aos mesmos resultados que nos dados originais. Em contraste, para base de dados de grandes dimensões, *Data Shuffling* oferece a mesma vantagem.

Assim a o Modelo EGADP é aconselhado a ser utilizado quando a base de microdados:

- Possui uma dimensão relativamente reduzida;
- É utilizada principalmente para análises estatísticas tradicionais e paramétricas;
- É utilizada para análises inferenciais.

Estas características estão presentes em bases de Microdados que correspondem a amostras utilizadas para inferir características populacionais.

Assim, a escolha entre um modelo linear e um modelo não linear, baseia-se essencialmente nas diferenças entre o Modelo EGADP e o modelo *Sata Shuffling*. Em termos de risco de identificação, ambos os modelos oferecem o menor risco possível. Quanto à perda de informação, ambos os modelos possuem vantagens e desvantagens.

Data Shuffling contém as seguintes vantagens sobre EGADP:

- Os valores originais das variáveis confidenciais não sofrem alterações;
- As relações não lineares são preservadas.

4.5.2 Comparação de Modelos com Métodos Perturbativos

Microagregação

O método perturbativo Microgregação modifica as distribuições marginais das variáveis sobre perturbação, as relações entre as variáveis confidenciais e as relações entre variáveis confidenciais e não confidenciais. Os testes realizados para este método mostram que o risco de identificação presente na base de Microdados perturbada apresenta níveis elevados de identificação. Quando as partições são de menor dimensão, a perda de informação diminui, no entanto, o risco de identificação aumenta. Quando a dimensão das partições é maior, o risco de identificação diminui e a perda de informação aumenta.

Comparando este método com os modelos *Data Shuffling* e EGADP, o método de Microagregação oferece resultados pouco eficazes, pois provoca aumentos no risco de identificação enquanto os modelos *Data Shuffling* e EGADP minimizam-no para o menor possível. Quanto à perda de informação, o método de Microagregação não preserva a validade de inferências como o modelo EGADP, ou não preserva distribuições marginais e relações não lineares das variáveis como o modelo*Data Shuffling*.

Rank Swapping

No método *Rank Swapping*, quando se define $d = 100$ (a ordem de duas observações trocadas entre si não pode estar afastada mais que d por cento do número total de observações), neste caso o processo é realizado aleatoriamente e desta forma, as relações existentes entre variáveis são destruídas, mas o risco de identificação é minimizado. Quando d é um valor pequeno, o risco de identificação aumenta significativamente, mas a perda de informação é mínima, pois a substituição será realizada por valores muitos próximos.

K. Muralidhar e R. Sarathy [17] realizaram uma comparação entre *Data Shuffling* e *Rank Swapping*. Dos testes realizados conclui-se que o risco de identificação resultante dos dois processos é idêntico quando o método *Rank Swapping* é realizado com $d = 1$, no entanto a perda de informação quando $d = 1$ é bastante superior à do modelo *Data Shuffling*. O mesmo acontece para valores baixos de d, *Data Shuffling* oferece a mesma utilidade de informação do método *Rank Swapping*, no entanto, o risco de identificação para o método *Rank Swapping* é bastante superior ao modelo *Data Shuffling*. Em suma, o modelo *Data Shuffling* apresenta sempre resultados mais eficazes do ponto de vista de identificação e da perda de informação.

ROMM

Grande parte das análises dos métodos perturbativos com base em modelos lineares focam-se nos efeitos provocados em regressões e nas covariâncias das variáveis originais. Através do modelo EGADP, os valores médios e as covariâncias das variáveis perturbadas e das variáveis originais são idênticos [6]. Para o método ROMM, estas

estimativas possuem algumas diferenças, no entanto, são estimativas consistentes nos seus erros. Por outro lado, o método ROMM oferece menor risco de identificação quando comparado com alguns modelos lineares. Tal não acontece com o modelo EGADP, pois este modelo oferece o menor risco de identificação possível.

Imputação Múltipla

A Imputação Múltipla é um processo de geração de dados sintéticos através de modelos de regressão. Quando o modelo aplicado no processo de imputação múltipla é linear, este método conduz a inferências válidas independentemente dos parâmetros utilizados. No entanto, K. Muralidhar e R. Sarathy [17] provaram que o modelo EGADP para além de oferecer inferências válidas, oferece um grau de confiança superior nas inferências quando comparado com o método de imputação múltipla. Outra vantagem é que o modelo EGADP não requer que múltiplas bases de microdados sejam analisadas e agregadas, e como tal a vantagem de imputação múltipla sobre EGADP não é clara e pode ser mais trabalhosa.

4.5.3 PRAM, Rank Swapping e Shuffling

A escolha entre estes métodos é baseada na estrutura que se pretende preservar dos dados originais. Nos casos em que existe um modelo de regressão significativo, quando aplicado nas variáveis originais, o método *Shuffling* fornece resultados satisfatórios. O método *Rank Swapping* funciona bem quando existem bastantes categorias ou valores muito heterogéneos nas variáveis em estudo. Já o método PRAM é preferido para perturbações com poucas variáveis categóricas onde o número de categorias é elevado. A vantagem deste método, como já foi referido, é a especificação da matriz de transição [3].

4.6 Conclusão

Neste Capítulo descreveram-se os diversos métodos de CDE e realizou-se uma comparação dos mesmos. Após a comparação dos métodos mais importantes, é percetível que a escolha recai maioritariamente sobre os modelos *Data Shuffling* e EGADP. No entanto, a escolha do método de CDE a utilizar não deve ser limitada a estas duas metodologias, pois as distribuições presentes na base de microdados podem influenciar a escolha do método a ser utilizado e resultados mais eficazes podem ser obtidos através de outros métodos. Na teoria, os modelos EGADP e *data shuffling* são os que oferecem simultaneamente o menor risco de identificação e a menor perda de informação. Na prática, a escolha depende do objetivo do responsável da base de microdados, ou seja, se o foco está em obter um baixo risco de identificação ou máxima utilidade nos dados divulgados. Assim, é aconselhável a aplicação de diferentes metodologias, separadamente, de forma a comparar os resultados obtidos. No final a base de microdados perturbada poderá ser resultado da aplicação de um ou mais métodos de CDE, dependendo dos objetivos do seu responsável.

5. Perda de Informação e Utilidade dos Dados

Após a aplicação dos métodos CDE é necessário uma avaliação da base de microdados perturbada, ou seja, avaliar a segurança e a perda de informação com a perturbação dos dados. Uma base de microdados perturbada deverá conter utilidade suficiente na sua informação, de tal forma, que permite os utilizadores retirar conclusões plausíveis através das análises na base de microdados perturbada. A relação entre o risco de identificação e a perda de informação é caracterizada por dois extremos: zero risco de identificação, no entanto a base de microdados perturbada apresenta baixa utilidade para os utilizadores; máximo risco de identificação, quando os dados são divulgados sem qualquer alteração e neste caso, existe zero informação perdida.

Em geral a perda de informação é medida através da análise das características das duas bases de dados, a base de dados original e a base de dados perturbada. Uma base de microdados perturbada apresenta utilidade na sua innformação caso sejam preservados [7]:

- Valores médios e covariâncias em pequenos subconjuntos;
- Valores marginais para algumas tabulações dos dados;
- Pelo menos uma característica da distribuição das variáveis perturbadas.

No entanto, uma base de microdados pode ser considerada válida sem cumprir estas condições, pois a perda de informação depende da escolha de um determinado nível de risco de identificação definido pelo responsável da base de microdados.

A base de microdados é constituída, em geral, por um conjunto de variáveis confidenciais (X) e não confidenciais (S), $W = \{X, S\}$. Idealmente, as características dos dados divulgados devem ser idênticas às dos dados originais, permitindo que qualquer análise realizada nos dados perturbados conduza exatamente aos mesmos resultados obtidos nos dados originais [20]. Matematicamente, a utilidade ideal dos dados pode ser definida como:

$$f(Y, S) = f(X, S)$$

onde $f(\cdot, \cdot)$ representa a função de densidade conjunta das variáveis confidenciais (X), não confidenciais (S) e perturbadas (Y). A avaliação dos diferentes métodos será realizada tendo por base esta relação. De seguida, apresentam-se algumas medidas úteis para avaliar a perda de informação.

5.1 Medidas de Perda de Informação para Variáveis Categóricas

Tal como os métodos de CDE, as medidas de perda de informação também diferem entre variáveis numéricas e variáveis categóricas. No caso das variáveis categóricas, a perda de informação é geralmente calculada através das diferenças nas categorias essencialmente das variáveis chave.

Comparação Direta de Variáveis Categóricas

Considerando X a matriz das variáveis originais e Y a matriz das variáveis perturbadas, ambas as matrizes de dimensão (NxQ), são compostas por Q variáveis categóricas sujeitas a perturbação e N observações. Antes de

realizar uma comparação direta entre estas matrizes é necessário a definição de distância para variáveis categóricas. Esta definição considera apenas a distância entre pares de categorias.

Para uma variável categórica X_q, $p = 1, ..., Q$, a única possibilidade é a comparação da igualdade, ou seja,

$$d_p(c, c') = \begin{cases} 0 & \text{se } c = c' \\ 1 & \text{se } c \neq c', \end{cases} \tag{5.1}$$

onde c representa a categoria da variável original e c' a categoria da mesma variável após perturbação.

Assim a medida para a perda de informação é obtida a partir do somatório das distâncias obtidas em (5.1) para todas as P variáveis categóricas sujeitas a perturbação. Esta medida corresponde ao número total de categorias que trocaram de posição durante o processo de perturbação, e é dada por:

$$PI = \sum_{p=1}^{P} d_p(c, c'). \tag{5.2}$$

Como será de se esperar quanto maior esta medida, maior serão as diferenças entre os dados originais e os dados perturbados.

Comparação de Tabelas de Contingência

Este método é apresentado como uma alternativa ao método apresentado anteriormente. Para que os dados apresentam utilidade, as tabelas de contingência deverão ser muito semelhantes.

Sejam $T^{(X)}$ e $T^{(Y)}$ as tabelas de contingência dos dados originais e dos dados perturbados, de dimensão C_1 x C_2 (número de categorias das variáveis c_1 e c_2), a distância entre as duas tabelas de contingência é definida por:

$$UT = \frac{1}{C_1 C_2} \sum_{h=1}^{C_1} \sum_{q=1}^{C_2} \left| T_{hq}^{(X)} - T_{hq}^{(Y)} \right|, \tag{5.3}$$

onde $T_{hq}^{(X)}$ representa o número de observações das variáveis originais que contêm simultaneamente a categoria h da variável c_1 e a categoria q da variável c_2. O mesmo acontece para $T_{hq}^{(Y)}$ mas neste caso para as variáveis perturbadas. Quanto maior o valor da medida calculada em (5.3), maior a perda de informação no processo de perturbação, e consequentemente menor a utilidade dos dados divulgados.

No entanto, como o número de tabelas de contingência depende do número de variáveis sujeitas a perturbação e do número de categorias para cada variável, uma versão normalizada é mais adequada em certos casos. Obtendo-se assim a medida:

$$UT1 = 100 \cdot \frac{1}{C_1 C_2} \sum_{h=1}^{C_1} \sum_{q=1}^{C_2} \left| \frac{T_{hq}^{(X)} - T_{hq}^{(Y)}}{T_{hq}^{(X)}} \right|. \tag{5.4}$$

Esta medida apenas considera a alteração relativa em cada célula da tabela de contingência, e é apresentada em percentagem.

Comparação de Valores em Falta

Uma outra medida da perda de informação é a comparação do número total de valores em falta nos dados originais e nos dados perturbados. Como já foi referido, existem métodos que na perturbação dos dados substituem os valores originais por valores em falta (NA) e obviamente este processo provoca uma redução significativa na utilidade dos dados perturbados [24].

Fixemos $\mathbf{R}^{(X)}$ e $\mathbf{R}^{(Y)}$, duas matrizes com a mesma dimensão de X e Y (NxQ). Um elemento de $\mathbf{R}^{(X)}$ assume o valor 1 quando a matriz das vaiáveis originais (X) apresenta um valor em falta nessa posição e assume valor 0 caso contrário. O mesmo acontece com $\mathbf{R}^{(Y)}$ relativamente à matriz das variáveis perturbadas (Y). Logo $\mathbf{R}^{(X)}$ e $\mathbf{R}^{(Y)}$ são matrizes compostas apenas por uns e zeros.

Constrói-se a matriz R com a mesma dimensão das matrizes anteriores, (NxP), e onde cada observação r_{ip} correspondente à observação i na variável p, da seguinte forma:

$$r_{ip} = \begin{cases} 0 & \text{se } r_{ip}^{(X)} = 0 \wedge r_{ip}^{(Y)} = 0 \\ 0 & \text{se } r_{ip}^{(X)} = 1 \wedge r_{ip}^{(Y)} = 1 \\ 1 & \text{se } r_{ip}^{(X)} = 0 \wedge r_{ip}^{(Y)} = 1 \\ 0 & \text{se } r_{ip}^{(X)} = 1 \wedge r_{ip}^{(Y)} = 0 \end{cases}. \tag{5.5}$$

Após a definição da matriz R com N observações e P variáveis, é possível obter o número total de valores em falta na base de dados após o processo de perturbação,

$$M = \sum_{i=1}^{N} \sum_{p=1}^{P} r_{ip}. \tag{5.6}$$

Assim, a medida M representa o número total de valores substituídos por valores em falta durante o processo de proteção.

Quanto maior o valor da medida M, maior a perda de informação existente na perturbação dos dados, e como tal menor a utilidade dos dados para os utilizadores.

Entropia

Entropia é uma medida teórica utilizada no cálculo da perda de informação, no entanto apenas pode ser usada em métodos de CDE onde processo de perturbação é modelado como ruído [7].

Como já foi visto no Capítulo 4, PRAM é um método que generaliza a Adição de Ruído, Microagregação e *Recoding*, portanto entropia será utilizada na avaliação da perda de informação apenas quando o método de perturbação aplicado nas variáveis em estudo é PRAM.

Seja X a matriz das variáveis originais sujeitas a perturbação pelo método PRAM, Y a matriz das variáveis perturbadas pelo método PRAM e $c_1, ..., c_l$ as categorias da variável perturbada Y_p . A entropia E_{Y_p} é definida como [24]:

$$E_{Y_p} = -\frac{1}{N} \sum_{c_l \in Y_p} f_{c_l} \cdot log\left(\frac{f_{c_l}}{N}\right), \tag{5.7}$$

onde f_{c_l} representa a frequência da categoria c_l na variável Y_p e N o número total de observações.

Esta medida quantifica a perda de informação pois avalia a perturbação causada na base de microdados perturbada, e caso este valor seja elevado, significa que existe elevada perda de informação, pois o processo de aplicação de ruído na perturbação foi intenso.

5.2 Medidas de Perda de Informação para Variáveis Numéricas

No caso das variáveis numéricas, as medidas de perda de informação são calculadas à *posteriori*, ou seja, são medidas que permitem a comparação entre as diversas estatísticas existentes nas duas bases de dados em estudo.

Considere-se uma base de dados com N indivíduos e P variáveis numéricas. Sejam X e Y, as matrizes das P variáveis contínuas numéricas e perturbadas, respetivamente, de dimensão (NxP). Apresentam-se a seguir algumas medidas úteis para a caracterização dos dados:

- **Matrizes de covariâncias:** V_X e V_Y;
- **Matrizes de correlação:** R_X e R_Y;
- **Comunalidades entre cada variável p e a primeira Componente Principal:** C_X e C_Y, isto é, é a percentagem de cada variável explicada pela componente principal PC_1;
- **Fator coeficiente:** F_X e F_Y. A matriz F_X contém os fatores que são multiplicados por cada variável de X obtendo assim a sua projeção sobre cada componente principal.

O objetivo dos métodos de medição da perda de informação para dados numéricos é avaliar as discrepâncias existentes entre as matrizes das variáveis originais e as matrizes das variáveis perturbadas. Existem três medidas que avaliam estas discrepâncias:

- **Erro Quadrático Médio (EQM):**

$$\frac{\sum_{i=1}^{N}\sum_{p=1}^{P}\left(x_{ip}-y_{ip}\right)^2}{NP}, \tag{5.8}$$

- **Erro Absoluto Médio (EAM):**

$$\frac{\sum_{i=1}^{N}\sum_{p=1}^{P}\left|x_{ip}-y_{ip}\right|}{NP}, \tag{5.9}$$

- **Variação Média:**

$$\sum_{i=1}^{N}\sum_{p=1}^{P}\frac{\frac{|x_{ip}-y_{ip}|}{|x_{ip}|}}{NP}, \tag{5.10}$$

onde P é o número total de variáveis numéricas confidenciais e N o número total de indivíduos. Estas medidas podem ser ainda aplicadas às matrizes de covariância ou de correlação, de forma a comparar as principais estatísticas entre os dados originais e os dados perturbados.

Assim é possível afirmar que o método mais eficaz na redução da perda de informação é o método que apresenta o menor valor nas três medidas apresentadas em (5.8),(5.9) e (5.10).

Medida IL1

A medida *Information Loss 1* (IL1) sugerida por J. Domingo-Ferrer [9] calcula a variação média entre os dados originais e os dados perturbados da seguinte forma:

$$IL1 = \frac{100}{PN} \sum_{i=1}^{N} \sum_{p=1}^{P} \frac{|x_{ip} - y_{ip}|}{|x_{ip}|}, \tag{5.11}$$

Caso a observação x_{ip} apresente valor nulo deverá ser substituída no denominador da fração pela observação perturbada y_{ip}. No caso de ambas as observações possuírem valor nulo, a variável p é excluída do cálculo. Para ultrapassar este problema criou-se a medida IL1s.

Medida IL1s

A medida IL1s pode ser interpretada como a distância escalar entre os dados originais e os dados perturbados [24] e é definida como:

$$IL1s = \frac{1}{PN} \sum_{p=1}^{P} \sum_{i=1}^{N} \frac{|x_{ip} - y_{ip}|}{\sqrt{2} S_p}, \tag{5.12}$$

onde S_p representa o desvio padrão da variável p nos dados originais.

Como esta medida é apresentada como uma distância escalar, quanto menor o valor desta distância, menor é a perda de informação no processo de perturbação.

Diferença dos Valores Próprios

Outra possibilidade de avaliação da perda de informação é através da comparação entre os valores próprios das variáveis originais e das variáveis perturbadas [3]. Neste caso, calculam-se as diferenças absolutas entre os valores próprios das matrizes de covariâncias das variáveis numéricas originais e perturbadas.

No Capítulo 6 é demonstrado como obter este valor usando a linguagem de programação R, e o resultado obtido é dado como a diferença dos valores próprios de cada matriz de covariâncias em percentagem. Quando a perda de informação é mínima, o resultado é 0, ou seja, não existem diferenças entre as matrizes de covariâncias.

Modelos de Regressão

Para além de comparações entre matrizes, modelos de regressão podem ser utilizados na avaliação das diferenças após a perturbação. Através da estimação de um modelo de regressão que contém as mesmas variáveis, nas duas bases de microdados, é possível comparar os dois modelos de forma a perceber se existem diferenças significativas nos parâmetros de cada modelo [3].

Esta medida pode ser definida como:

$$lm = \sum_{l=1}^{D} \left| \frac{\beta_l^x - \beta_l^y}{\beta_l^x} \right|, \tag{5.13}$$

onde β_l^x representa o parâmetro l do modelo obtido a partir dos dados originais e β_l^y representa o parâmetro l do mesmo modelo obtido a partir dos dados perturbados e D representa o número total de parâmetros do modelo.

Quanto menor a soma apresentada em (5.13) maior a utilidade da informação contida na base de microdados perturbada.

Coeficiente de Gini

O coeficiente de Gini é utilizado na medição da qualidade dos dados. É uma medida de dispersão, normalmente utilizada para medir a percentagem de desigualdade nas variáveis sensíveis. Este coeficiente compara as distribuições das variáveis em estudo [3].

O coeficiente Gini para uma variável X_p pode ser calculado da seguinte forma:

$$G_{Xp} = 100 \left[\frac{2 \sum_{i=1}^{N} \left(\omega_i x_{ip} \sum_{j=1}^{i} \omega_j \right) - \sum_{i=1}^{N} \omega_i^2 x_{ip}}{\left(\sum_{i=1}^{N} \omega_i \right) \sum_{i=1}^{N} \omega_i x_{ip}} - 1 \right], \tag{5.14}$$

onde x_i representa a observação original i da variável X_p e ω_i os pesos amostrais de cada observação.

O coeficiente de Gini pode ser obtido a partir da linguagem de programação R, como exemplificado no Capítulo 6, e é apresentado em percentagem. O coeficiente de Gini igual a 0 corresponde à igualdade total, ou seja, este apresenta o mesmo valor para todas as observações e o valor 100 corresponde a total desigualdade.

A medida de perda de informação pode ser definida da seguinte forma:

$$PI = \left| G_{X_p} - G_{Y_p} \right|, \tag{5.15}$$

onde G_{Xp} representa o coeficiente de Gini da variável original X_p e G_{Yp} o coeficiente de Gini da variável perturbada Y_p. Quanto maior esta diferença, menor a utilidade da informação da base de microdados perturbada, pois as desigualdades existentes na variável original deverão ser as mais similares possíveis às desigualdades presentes na variável perturbada.

6. Ambiente R

Neste projeto é utilizada a linguagem de programação R e neste Capítulo apresenta-se uma explicação das diversas funções existentes no *package* **sdcMicro**, fornecido pelo ambiente R para a aplicação dos métodos de CDE.

6.1 Ferramentas para CDE

Para além da linguagem de programação R existem outras ferramentas ou linguagens disponíveis para a aplicação dos métodos CDE, por exemplo, μ-Argus e C++. No entanto, estas opções são mais limitadas que o *package* **sdcMicro** do ambiente R, como se pode ver pela Figura 4 que ilustra as diferenças entre as três ferramentas [24]. O ambiente R fornece ainda o *package* **sdcTable** que contém diversas funções para a perturbação de dados tabulares. Para além do ambiente R existe ainda a ferramenta τ-argus que permite a perturbação de dados tabulares.

μ-argus

Através de um programa da União Europeia, μ-argus foi desenvolvido com o objetivo de perturbar bases de microdados. Em 1995, foi lançado a primeira versão da ferramenta pelo Departamento de Métodos Estatísticos na Holanda. Na atualidade, ainda se encontra a ser aprimorado por algumas instituições estatísticas. Esta ferramenta foi originalmente desenvolvida em visual basic até à versão 4.2 e agora é escrita em Java e pode ser obtida de forma gratuita no site do CASC (http://neon.vb.cbs.nl/casc/mu.html) [24].

Linguagem C++

Estudos por parte do *International Household Survey Network* (IHSN) permitiram o desenvolvimento de uma linguagem de programação para a perturbação de microdados. IHSN desenvolveu um código em C++ que permite perturbar e anonimizar bases de microdados com o objetivo de suportar a divulgação segura de dados confidenciais. Este código desenvolvido por IHSN é gratuito a todos os utilizadores, no entanto está totalmente presente e atualizado no *package* **sdcMicro** [24].

Method | Software	μ-**Argus** 4.2	**sdcMicro** > 4.3.0	C++
frequency counts	✔	✔	✔
individual risk (IR)	✔	✔	✔
IR on households	✔	✔	✔
l-diversity		✔	✔
suda2		✔	✔
global risk (GR)	✔	✔	✔
GR with log-lin mod.		✔	
recoding	✔	✔	(✔)
local suppression	(✔)	✔	(✔)
swapping	(✔)	✔	✔
pram	✔	✔	✔
adding correlated noise		✔	✔
microaggregation	✔	✔	✔
shuffling		✔	
utility measures	(✔)	✔	
GUI	(✔)		
CLI		(✔)	✔
reporting	✔	✔	
platform independent		✔	✔
free and open-source		✔	✔

Figura 4: Comparação entre as diferentes opções de aplicação de metodologias de CDE, retirado de [24]

O *package* **sdcMicro** não só permite aplicar um maior número de métodos, como também fornece uma linguagem simples e eficaz para um utilizador aplicar os diferentes métodos de CDE.

6.2 Package sdcMicro

A primeira versão do *package* **sdcMicro** foi lançada em 2017 e possuia apenas alguns métodos de CDE. Com o decorrer dos anos, este *package* sofreu diversas atualizações e na atualidade possui a maioria dos métodos de CDE sendo uma das ferramentas mais completas para proteção de bases de dados.

Este *package* apresenta muitas vantagens em comparação com outras ferramentas/linguagens:

- Utiliza objetos da classe S4 (um objeto da classe S4 permite a implementação de algoritmos complexos de forma eficaz);
- Utiliza uma linguagem amigável do ponto de vista do utilizador;
- A maioria da informação é atualizada automaticamente após a aplicação dos métodos;
- A possibilidade de voltar a diferentes fases do processo sem necessidade de cálculos adicionais.

Assim, o *package* **sdcMicro** fornece diversas funções capazes de aplicar as metodologias de CDE numa base de dados, e de seguida apresenta-se uma breve explicação das funções que se considera de maior importância na aplicação de métodos perturbativos a uma base de dados.

Aplicação Função sdcApp

Uma das grandes vantagens do *package* **sdcMicro** é que possui uma aplicação que permite executar os métodos de CDE e calcular as medidas necessárias, de forma a que utilizadores com pouca experiência em programação sejam capazes de aplicar métodos de CDE em bases de dados. Para que a aplicação desses métodos apresente resultados eficazes apenas é necessário o conhecimento dos conceitos da área de CDE.

Na Figura 5 apresenta-se a interface gráfica obtida ao executar a função **sdcApp**.

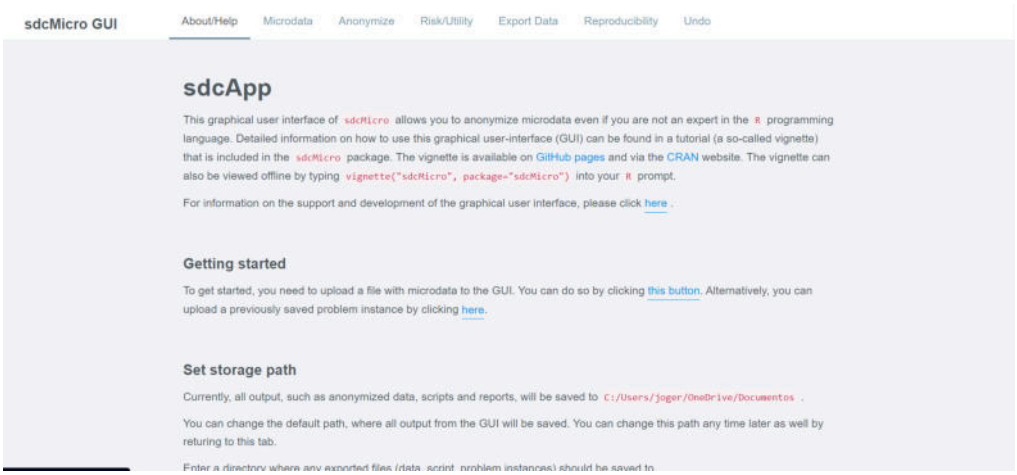


Figura 5: Interface gráfica da função **sdcApp**

Esta é uma possibilidade que o *package* oferece ao executar a função **sdcApp**, no entanto, reconhece-se maior interesse e flexibilidade na aplicação destes métodos através da linguagem de programação R.

6.2.1 Objeto CDE

Função CreateSdcObj

A função **createSdcObj** permite obter resultados de uma forma simples e é de grande importância pois é capaz de armazenar toda a informação necessária no processo de perturbação dos dados. No entanto, para a criação deste objeto é necessário o conhecimento de todos os tipos de variáveis que serão utilizadas no processo de perturbação.

No Código 1 apresenta-se a criação do objeto CDE em ambiente R, utilizando os seguintes argumentos:

- *X*: Base de dados original;
- *keyVars*: Variáveis chave;
- *numVars*: Variáveis numéricas sujeitas a perturbação;
- *pramVars*: Variáveis categóricas a que se pretende aplicar o método PRAM;
- *sensibleVar*: Variáveis que contêm informação confidencial;
- *weightVar*: Pesos amostrais;
- *hhId:* Variável que identifica o grupo, no caso de uma base de dados com estrutura hierárquica.

```
> sdc <- createSdcObj(X, keyVars, numVars, pramVars,
+                     sensibleVar, weightVar, hhId)
```

Código 1: Aplicação da função **createSdcObj**

Esta função cria um objeto em ambiente R e calcula as medidas de risco de identificação inicial. Após a aplicação dos métodos de CDE, o objeto armazena os resultados obtidos para as medidas de risco de identificação e de perda de informação da base de microdados perturbada.

Função slotNames

Com a função **slotNames** é possível visualizar a informação que o objeto CDE criado é capaz de armazenar quando se aplica os métodos de CDE. No Código 2 está apresentada essa informação, sendo que a maioria dos *slots* não contém qualquer informação antes de se aplicar um ou mais métodos de CDE.

```
> slotNames(sdc)
 [1] "origData"          "keyVars"           "pramVars"
 [4] "numVars"           "ghostVars"         "weightVar"
 [7] "hhId"              "strataVar"         "sensibleVar"
[10] "manipKeyVars"      "manipPramVars"     "manipNumVars"
[13] "manipGhostVars"    "manipStrataVar"    "originalRisk"
[16] "risk"              "utility"           "pram"
[19] "localSuppression"  "options"           "additionalResults"
[22] "set"               "prev"              "deletedVars"
```

Código 2: Aplicação da função **slotNames**

Os *slots* resultantes do Código 2 apresentam a seguinte informação:

- *origData*: Base de dados original sem qualquer alteração;
- *keyVars* e *numVars*: Indica as colunas pertencentes às variáveis chave e às variáveis numéricas às quais se aplicarão os métodos de CDE;
- *pramVars*: Indica as colunas das variáveis categóricas que se pretende perturbar pelo método PRAM;
- *ghostVars*: Indica, caso existam, as variáveis externas associadas a variáveis chave;
- *weigthVar*: Indica a coluna dos pesos amostrais;
- *hhId*: Indica a coluna a que pertence a variável de identificação do grupo;
- *strataVar*: Indica a coluna a que pertence a variável de estratificação;
- *sensibleVar*: Indica as colunas a que pertencem as variáveis sensíveis;
- *Slots* que começam com *manip*: Contêm as variáveis perturbadas após a aplicação dos métodos;
- *utility*: Contém a informação acerca das medidas de perda de informação;
- *additionalResults*: É capaz de armazenar medidas adicionais que se pretendam calcular;
- *pram*: Guarda a informação sobre o método PRAM, tal como a matriz de transição, as mudanças provocadas nas variáveis pelo método e ainda uma comparação entre a variável original e a variável perturbada;
- *originalRisk* e *Risk*: Contêm a informação quanto ao risco de identificação inicial e final. Caso nenhum método de CDE tenha sido aplicado, a informação guardada para o *slot originalRisk* é a mesma que para o *slot Risk*, que será atualizado quando se aplica um método de CDE.

  No Código 3 apresenta-se uma forma de como ter acesso à informação contida nestes *slots*, divididas em:

  - *global*: Risco de identificação global com as seguintes medidas:
    - *risk* e *hier_risk*: Risco de identificação global e risco de identificação hierárquico global da base de dados original, calculado a partir da média dos riscos de identificação individuais;
    - *risk_ER* e *hier_risk_ER*: Número esperado de identificações de observações e de grupos na base de dados original;
    - *risk_pct* e *hier_risk_pct*: Percentagem do risco de identificação global e hierárquico global;
    - *threshold*: Risco de identificação máximo tolerável para uma observação;
    - *max_risk*: Risco de identificação global máximo, definido como o rácio entre o número de observações superiores ao valor limite de risco de identificação individual (*threshold*) e o número total de observações;
  - *individual*: Risco de identificação individual com as seguintes medidas:
    - *risk*: Risco de identificação individual de cada observação;
    - *fk*: *Frequency counts* amostrais de cada observação, no caso da base de dados em estudo ser uma amostra;
    - *Fk*: *Frequency counts* populacionais de cada observação, ou estimativas das *frequency counts* populacionais, no caso da base de dados se tratar de uma amostra;

  - *hier_risk*: Risco hierárquico de cada observação;

```
> names(sdc@originalRisk)
[1] "global"       "individual"
> names(sdc@originalRisk$global)
[1] "risk"           "risk_ER"        "risk_pct"
[4] "threshold"      "max_risk"       "hier_risk_ER"
[7] "hier_risk"      "hier_risk_pct"
> colnames(sdc@originalRisk$individual)
[1] "risk"       "fk"         "Fk"         "hier_risk"
```

Código 3: Aplicação do*Slot* sdc@originalRisk

Função addGhostVars

É comum existirem variáveis conhecidas associadas a variáveis chave presentes na base de dados em estudo. Para uma maior eficácia na aplicação dos métodos, é possível adicionar estas variáveis ao objeto CDE criado. No Código 4 é apresentada a função **addGhostVars** com os seguintes argumentos [26]:

- *sdc*: Objeto CDE criado com o Código 1;
- *keyVar*: Variável chave da base de dados em estudo que está associada a uma variável externa;
- *ghostVar*: Variável externa associada à variável chave indicada.

```
> sdc <- addGhostVars(sdc, keyVar, ghostVar)
```

Código 4: Aplicação da função **addGhostVars**

No caso de se pretender adicionar mais que uma variável, teria de se repetir o processo.

6.2.2 Medição de Risco de Identificação

Função measure_risk

Esta função permite calcular o risco de identificação diretamente de uma base de dados, como é apresentado no Código 5. O risco de identificação é calculado a partir da função à *posteriori* das *Frequency Counts* utilizando os seguintes argumentos:

- *X*: Base de dados original;
- *keyVars*: Variáveis Chave;
- *w*: Variável que representa os pesos amostrais;
- *hhId*: Variável a que corresponde a identificação do grupo;
- *max_global_risk*: Risco de identificação global máximo para o qual se considera que uma observação é insegura.

```
> res <- measure_risk(X,
+                     keyVars ,
                    w, hhld ,max_global_risk)
```

Código 5: Aplicação da função **measure_risk**

A função **measure_risk** é útil caso se pretenda avaliar apenas o risco de identificação de uma base de dados sem se realizar qualquer tipo de perturbação, pois esta função fornece o risco individual e o risco global da base de dados.

Função suda2

Outra possibilidade para o cálculo do risco de identificação são os *DIS-SUDA scores* (Capítulo 3, Secção 3.1). O ambiente R fornece a função **suda2** que calcula os *DIS-SUDA scores* para todas as observações da base de dados. Este algoritmo como já foi referido anteriormente apresenta resultados bastante eficazes.

No Código 6 é apresentado a função **suda2** com os seguintes argumentos:

- *sdc*: Objeto CDE criado com o Código 1;
- *original_scores*: Por defeito este argumento é indicado como *FALSE*, o que significa que o cálculo dos *DIS-SUDA scores* é realizado através da procura por MSUs, como descrito no Capítulo 5.

```
> sdc <- suda2(sdc, original_scores=FALSE)
> names(sdc@risk)
[1] "global"     "individual" "numeric"    "suda2"
```

Código 6: Aplicação da função **suda2**

Pelo Código 6 é ainda visível que o objeto CDE criado guarda a informação dos *DIS-SUDA scores* diretamente no argumento do risco de identificação.

Função ldiversity

À semelhança do que é feito para o cálculo dos DIS-SUDA *scores*, também é possível calcular a medida *l-diversity* a partir do objeto CDE, que armazena essa informação juntamente com as restantes medidas de risco de identificação.

A função **ldiversity**, como é apresentado no Código 7, tem um único argumento que corresponde ao objeto CDE criado com o Código 1.

```
> sdc <- ldiversity(sdc)
> names(sdc@risk)
[1] "global"     "individual" "numeric"    "ldiversity"
```

Código 7: Aplicação da função **ldiversity**

É ainda possível aplicar esta função diretamente na base de microdados, e para tal os argumentos já sofrem alterações:

- *X*: Base de dados original;
- *keyvars*: Variáveis chave;

- *ldiv_index*: Variáveis sensíveis, ou seja, variáveis que necessitam de cumprir as condições de *l-diversity*.

```
> div <- ldiversity(X,
+                   keyVars,
+                   ldiv_index)
```

Código 8: Aplicação da função **ldiversity**

Os resultados obtidos com o Código 7 e com o Código 8 serão sempre os mesmos, no entanto, aconselha-se a abordagem apresentada no Código 7, pois desta forma a informação é armazenada num *slot* juntamente com o resto dos dados e as medidas importantes para a perturbação dos dados.

Função dRisk

A função **dRisk** calcula o risco de identificação das variáveis numéricas. É definida como a medida intervalar descrita no Capítulo 3, ou seja, o valor apresentado por esta função é a percentagem de valores perturbados que estão contidos nos intervalos definidos para as observações originais.

No Código 9 apresentam-se duas possibilidades de executar a função. A primeira, apresentada na linha 1, necessita da indicação da base de dados original (argumento *X*) e da base de dados perturbada, (argumento *Xm*). A segunda possibilidade recorre-se ao objeto CDE (argumento *sdc*), que permite guardar a informação acerca dessa medida no argumento *numeric*. A função **dRisk** possibilita ainda definir a percentagem do desvio padrão dos intervalos construídos em torno das observações originais (argumento *k* que varia entre 0 e 100%).

```
1 > dRisk(X, Xm, k)
2
3 > sdc <- dRisk(sdc, k)
4 > names(sdc@risk)
5 [1] "global"     "individual" "numeric"
```

Código 9: Aplicação da função **dRisk**

Função dRiskRMD

A função **dRiskRMD** calcula o risco de identificação através das distâncias robustas de Mahalanobis (RMD) entre os dados originais e os dados perturbados. É descrita como a medida de deteção de *outliers* (Capítulo 3, secção 3.2), e verifica se todas as observações perturbadas contêm pelo menos m observações na sua vizinhança.

Como é visível no Código 10, ao aplicar esta função num objeto CDE a informação é guardada no argumento *numericRMD*. Os argumentos utilizados nesta função são:

- *X*: Base de dados original;
- *Xm*: Base de dados perturbada;
- *sdc*: Objeto CDE criado com o Código 1.

```
1 > dRiskRMD(X, Xm)
2 > sdc <- dRiskRMD(sdc)
3 > names(sdc@risk$numericRMD)
4 [1] "risk1"       "risk2"       "wrisk1"      "wrisk2"
5 [5] "indexRisk1" "indexRisk2" "riskvec1"    "riskvec2"
```

Código 10: Aplicação da função **dRiskRMD**

Pelo Código 10 é percetível que a função retorna diversos valores, estes valores são representados da seguinte forma:

- *risk1*: Percentagem de observações inseguras;
- *risk2*: Versão estandardizada do valor de *risk1* com base nos pesos amostrais;
- *wrisk1*: Número de observações inseguras;
- *wrisk2*: Percentagem de valores considerados seguros, esta medida foi apresentada neste projeto como medida intervalar;
- *indexRisk1*: Observações que apresentam elevado risco de acordo com a medida *risk1*;
- *indexRisk2*: Observações que apresentam elevado risco de acordo com a medida *wrisk1*;
- *riskvec1*: Vetor com o risco de identificação de cada observação de acordo com as medidas *risk1*;
- *riskvec2*: Vetor com as observações que apresentam elevado risco de acordo com a medida *wrisk1*.

6.2.3 Aplicação de Métodos de CDE

Função CreateNewID

O primeiro passo na proteção de uma base de dados é a anonimização. O *package* **sdcMicro** fornece a função **CreateNewID** que permite a anonimização das variáveis identificadoras. No Código 11 apresenta-se a função que permite a anonimização das variáveis identificadoras, utilizando os seguintes argumentos:

- *sdc*: Objeto CDE criado com o Código 1;
- *newID*: Nome da variável a que se pretende atribuir as novas identificações;
- *withinVar*: Variável com a identificação dos grupos, caso se trate de uma base de dados hierárquica.

```
1 > sdc <- createNewID(sdc, newID, withinVar)
```

Código 11: Aplicação da função **createNewID**

Esta função substitui a variável identificadora por outra variável gerada aleatoriamente, ou seja, ao apresentar a base de dados perturbada não existe nenhuma variável original que seja identificadora de observações ou de grupos.

Função addNoise

A função **addNoise** contém diversas possibilidades para a aplicação do método perturbativo de adição de ruído. Esta função é apresentada no Código 12, com os seguintes argumentos:

- *obj*: Objeto CDE criado como no Código 1;
- *noise*: Magnitude percentual do ruído que se pretende adicionar. Caso não se defina a magnitude adicionada é de 150;

- *method*: Indicação do método de adição de ruído que se pretende executar, as possibilidades são:
  - *additive:* Método que adiciona ruído não correlacionado;
  - *ROOM:* Novo método de adição de ruído que apresenta resultados eficazes, pois preserva as propriedades da base de dados e permite maior controlo sobre a magnitude de ruído a adicionar. Este método é descrito com maior detalhe no Capítulo 4 deste projeto;
  - *restr:* Método de adição de ruído que tem em conta a dimensão da amostra;
  - *correlated* e *correlated2*: Duas formas de aplicar o método de adição de ruído ruído correlacionado;
  - *outdec:* Método que adiciona ruído apenas a observações *outliers*. Estas observações são identificadas através da distância de Mahalanobis.

No Código 12 apresenta-se a função para aplicação do método de adição de ruído. Com a função **print**, como está apresentado, é possível obter o risco de identificação e os valores para as medidas de perda de informação após a aplicação do método.

```
> sdc <- addNoise(obj, noise, method)
> print(sdc, "numrisk")
```

Código 12: Aplicação da função **addNoise**

Função argus_microaggregation

A função **argus_microaggregation** utiliza o código do *software* μ-*argus* e permite a execução do método perturbativo microagregação. Esta função apenas suporta base de dados com variáveis numéricas e após a aplicação deste método, a função apresenta os dados originais e os dados perturbados. No Código 13 está apresentado esta função com os seguintes argumentos:

- *X*: Base de dados apenas com variáveis numéricas;
- *K*: Dimensão de cada partição que se pretende formar;
- *useOptimal:* Indicação de microagregação univarida ou multivariada. Caso seja *TRUE*, corresponde a utilizar a microagregação em apenas uma variável, por defeito está definido como *FALSE*.

```
> pert <- argus_microaggregation(X, K, useOptimal = FALSE)
```

Código 13: Aplicação da função **argus_microaggregation**

Função microaggregation

Esta função permite também a aplicação do método microagregação, e comparando com a função descrita anteriormente, esta contém maior número de métodos para a formação de partições na aplicação do método perturbativo microagregação.

Como é apresentado no Código 14, a função pode ser executada num objeto CDE ou numa base de dados, desde que se indique as variáveis necessárias para a perturbação. Os argumentos necessários para esta função são:

- *sdc* ou *X*: Objeto CDE criado com o Código 1 ou base de dados original;

- *aggr*: Dimensão desejada para cada partição;
- *variables*: Variáveis que se pretende perturbar por microagregação;
- *method*: Método para a criação das partições, com as seguintes possibilidades:
  - *pca* - Análise em Componentes Principais;
  - *mdav* - Método que agrupa as observações de acordo com a sua distância euclidiana, este método é descrito com mais detalhe no Capítulo 4;
  - *rmd* - Método que constrói partições com base em distâncias multivariadas, é uma extensão do método *mdav* e é o método mais recomendado pelo ambiente R;
  - *onedims* - Método de ordenação individual, que permite que as propriedades das estatísticas univariadas sejam preservadas de forma mais eficaz. No entanto, as propriedades das estatísticas multivariadas são fortemente afetadas conforme descrito no Capítulo 4;
  - *simple* - Aplica o método microagregação sem ordenação das observações;
  - *clustpppca* - Método que constrói clusters no processo de criação de partições, geralmente apresenta resultados eficazes;
  - *pppca* - Método de projeção sobre as componentes principais. Este método também é descrito em detalhe no Capítulo 4;

```
1 > sdc <- microaggregation(sdc,aggr, method)
2
3 > micro <- microaggregation(X, variables, aggr, method)
```

Código 14: Aplicação da função **microaggregation**

Função argus_rankswap

Novamente o ambiente R utiliza o código do *software* μ-Argus, utilizando a função **argus_rankswap** para a aplicação do método *rank swapping*.

A função **argus_rankswap** apenas aceita base de dados com variáveis numéricas ou variáveis categóricas ordinais, no argumento *X*, e o argumento *P* define a distância máxima possível para a troca de observações, expressa como percentagem.

```
1 > pe <- argus_rankswap(X, P)
```

Código 15: Aplicação da função **argus_rankswap**

Função rankSwap

A função **rankSwap** permite também a aplicação do método *rank swapping* utilizando os seguintes argumentos:

- *TopPercent* e *BottomPercent*: Representam a percentagem dos valores superiores e dos valores inferiores que serão agrupados antes da aplicação do método *rank swapping*;

- *P:* Define a distância máxima possível para a troca de observações, expressa como a percentagem das observações totais. Dois valores são considerados elegíveis para troca caso as suas ordens de ordenação, i e j:

$$|i - j| \leq \frac{PN}{100}, \tag{6.1}$$

onde N é a dimensão da base de dados;

- *R0* : Fator de preservação de correlação, obtido como $R0$=$\frac{R_1}{R_2}$, onde R_1 representa a correlação original entre duas variáveis e R_2 a correlação entre as mesmas variáveis após a aplicação do método;

- $K0$: Fator de preservação dos valores médios, obtido por

$$|X_1 - X_2| < \frac{2K0X_1}{\sqrt{N}}, \tag{6.2}$$

onde X_1 e X_2 representam os valores médios de cada variável antes e após a aplicação do método, respetivamente, e N representa a dimensão da base de dados.

```
> sdc_rs <- rankSwap(sdc,TopPercent,
+                    BottomPercent ,
+                     K0, R0 ,P)
```

Código 16: Aplicação da função **rankSwap**

Função shuffle

O método *shuffling* pode ser aplicado através da função **shuffle** e vários argumentos são utilizados, como é apresentado no Código 17. A eficácia desta função depende do modelo que se utiliza, logo é necessário testar esse modelo, isto é, se o modelo criado é ou não estatísticamente significativo.

Os argumentos a utilizar para esta função são:

- *sdc*: Objeto CDE criado com o Código 1;

- *covmethod*: Método para a estimação da covariância, usando o coeficiente de Spearman ou de Pearson, estando por defeito definido o coeficiente de Spearman;

- *regmethod*: Método para a regressão multivariada, com as opções lm (*Linear model*) ou MM (*Robust linear Model*), estando por defeito definido o método lm;

- *gadp*: Caso seja *TRUE* executa o método de adição de ruído GADP, descrito com maior detalhe no Capítulo 5;

- *form*: Modelo de regressão que se pretende utilizar para a realização da substuituição.

```
> sdc <- shuffle(sdc,covmethod,
                 regmethod, gadp = TRUE, form)
```

Código 17: Aplicação da função **shuffle**

Função pram

A função **pram** permite a aplicação do método perturbativo PRAM a variáveis categóricas indicadas no argumento **pramVars** do objeto CDE ou então a uma base de dados. Esta função utiliza os seguintes argumentos:

- *sdc* ou *X*: Objeto CDE criado com o Código 1 ou base de dados original;
- *pd*: Matriz de transição ou valor mínimo para os valores diagonais da matriz de transição,isto é, a percentagem miníma para uma categoria permanecer inalterada;
- *alpha*: Representa a quantidade de perturbação que se pretende adicionar. Caso a matriz de transição seja especificada, este argumento é ignorado;
- *variables*: Variáveis categóricas sujeitas ao método perturbativo PRAM.

No Código 18 apresenta-se a aplicação desta função ao objeto CDE (linha 1) e diretamente a uma base de dados (linha 2).

```
1 > sdc_pram <- pram(sdc, pd, alpha)
2 > res2 <- pram(
3 +   X,
4 +   variables,
5 +   pd ,
6 +   alpha )
```

Código 18: Aplicação da função **pram**

A forma mais eficaz da aplicação do método PRAM é através do objeto CDE, visto que a informação é armazenada juntamente com as outras variáveis perturbadas. Para além dessa vantagem, ao aplicar o método a um objeto CDE é possível ter acesso à matriz de transição do processo, o que não acontece com a aplicação da função numa base de dados.

Função localSupp

A função **localSupp** permite a supressão de valores de uma variável chave, cuja observação possui um risco de identificação superior a um determinado valor. Desta forma, observações com risco de identificação elevado são eliminadas. Esta função utiliza os seguintes argumentos:

- sdc: Objetod CDE criado com o Código 1;
- keyvar: Variável chave onde se pretende realizar a supressão;
- threshold: Valor limite para o risco de identificação. Todas as observações com risco de identificação superior a este valor serão suprimidas na variável chave indicada.

No Código 19 é apresentado a aplicação da função em linguagem R a um objeto CDE. Após a aplicação da função, é possível apresentar o número total de observações suprimidas na variável (linha 2).

```
1 sdc <- localSupp(sdc, threshold = 0.15, keyVar)
2 sdc@localSuppression
```

Código 19: Aplicação da função **localSupp**

Função undolast

A função **undolast** oferece a possibilidade de desfazer a última função realizada no objeto CDE. Esta função é muito útil quando se verifica que um método não apresenta os resultados pretendidos pelo utilizador.

```
1 > undolast(sdc)
```

Código 20: Aplicação da função **undolast**

6.2.4 Avaliação de Perda de Informação

Função dUtility

O objeto CDE criado armazena a informação acerca da perda de informação e da utilidade dos dados. Esta informação é atualizada automaticamente após a aplicação de um método perturbativo ou então através da função **dUtility**. Pode ser executada como se apresenta no Código 21, com os seguintes argumentos:

- *X*: Base de dados original;
- *Xm*: Base de dados perturbada;
- *method*: Medida de avalição que se pretende obter. As opções são a diferença de valores próprios, IL1 ou IL1s.

Para um objeto CDE a informação acerca destas medidas pode ser obtida como apresentado na linha 2 do Código 21.

```
1 > sdc <- dUtility(sdc)
2 > sdc@utility
3 > dUtility(X, Xm , method)
```

Código 21: Aplicação da função **dUtility**

Função Gini - Package laeken

É possível armazenar o coeficiente Gini no argumento **additionalresults** do objeto CDE. Neste caso, utiliza-se o *package* **laeken**, que contém a função **gini**. Esta função contém diversos argumentos que podem ser utilizados para o cálculo do coeficiente.

No Código 22 apresenta-se a função e o procedimento necessário para armazenar a informação no argumento *additionalresults*:

- *inc*: Variável numérica que se pretende avaliar;
- *weights*: Pesos amostrais de cada observação;
- *breakdown*: Caso possua uma estrutura hierárquica, indica a variável identificadora dos grupos;
- *data*: Base de dados em estudo.

Na linha 3 do Código 22 é apresentado a forma de ter acesso ao valor calculado para o coeficiente Gini.

```
> sdc@additionalResults$gini <- gini(inc, weigths,
+                                    breakdown, data)$value
> sdc@additionalResults$gini
```

Código 22: Aplicação da função **gini**

6.2.5 Extração de Resultados

Função print

A função **print** permite obter as diversas propriedades e informações armazenadas no objeto CDE criado. No Código 23 é apresentado como executar esta função, os argumentos necessários para tal são:

- *sdc*: Objeto CDE criado como no Código 1;
- *"..."*: Argumentos guardados no objeto CDE, como apresentado no Código 2.

```
> print(sdc, "...")
```

Código 23: Aplicação da função **print**

Função extractManipData

Por fim, a função **extractManipData** permite obter a base de dados perturbada. No Código 24 está apresentada a função e utiliza os seguintes argumentos:

- *ignoreKeyVars*: Caso se pretenda as variáveis chave originais em vez das variáveis perturbadas, o argumento tem de ser definido como *TRUE*;
- *ignorePramVars*: Caso se pretenda as variáveis originais em vez das variáveis perturbadas pelo método PRAM, o argumento tem de ser definido como *TRUE*;
- *ignoreNumVars*: Caso se pretenda as variáveis numéricas originais em vez das variáveis numéricas perturbadas, o argumento tem de ser definido como *TRUE*;
- *ignoreStrataVar*: Caso se pretenda as variáveis de estratificação originais em vez das variáveis de estratificação perturbadas, o argumento tem de ser definido como *TRUE*;

Por defeito, todos estes argumentos estão definidos como *FALSE*, ou seja, permite obter uma base de dados que contém todas as variáveis perturbadas pelos métodos de CDE.

```
Anon <- extractManipData(sdc, ignoreKeyVars = F, ignorePramVars = F,
                         ignoreNumVars = F, ignoreStrataVar = F)
```

Código 24: Aplicação da função **extractManipData**

6.3 Exemplo EUSILCS

Nesta secção apresenta-se um exemplo de como utilizar algumas funções do *package* **sdcMicro**, em ambiente R, para a perturbação de uma base de dados.

A base de dados *EusilcS* (European Union Statistics on Income and Living Conditions) escolhida é constituída por 11723 observações e 18 variáveis, contida no *package* **simPop** [10]. Esta base de dados foi gerada a partir de dados reais que retratam o estilo de vida dos indivíduos que residem na Áustria no ano de 2006 e possui uma estrutura hierárquica, estando os indivíduos distribuídos por agregados familiares.

No Código 25 está apresentado a estrutura da base de dados em estudo.

```
> str(eusilcS)
'data.frame': 11725 obs. of  18 variables:
 $ db030     : int  1 1 2 3 4 4 4 5 5 5 ...
 $ hsize     : int  2 2 1 1 3 3 3 5 5 5 ...
 $ db040     : Factor w/ 9 levels "Burgenland","Carinthia",..: 4 4 7 5 7...
 $ age       : int  72 66 56 67 70 46 37 41 35 9 ...
 $ rb090     : Factor w/ 2 levels "male","female": 1 2 2 2 2 1 1 1 2 2...
 $ pl030     : Factor w/ 7 levels "1","2","3","4",..: 5 5 2 5 5 3 1 1 3 NA...
 $ pb220a    : Factor w/ 3 levels "AT","EU","Other": 1 1 1 1 1 1 3 1 1 NA...
 $ netIncome: num  22675 16999 19274 13319 14366 ...
 $ py010n    : num  0 0 19274 0 0 ...
 $ py050n    : num  0 0 0 0 0 ...
 $ py090n    : num  0 0 0 0 0 ...
 $ py100n    : num  22675 0 0 13319 14366 ...
 $ py110n    : num  0 0 0 0 0 0 0 0 0 NA ...
 $ py120n    : num  0 0 0 0 0 0 0 0 0 NA ...
 $ py130n    : num  0 16999 0 0 0 ...
 $ py140n    : num  0 0 0 0 0 0 0 0 0 NA ...
 $ db090     : num  7.82 7.82 8.79 8.11 7.51 ...
 $ rb050     : num  7.82 7.82 8.79 8.11 7.51 ...
```

Código 25: Estrutura da base de dados EusilcS

Após uma melhor perceção da estrutura da base de dados, de seguida, realiza-se uma análise sobre o risco de identificação da base de dados original.

Como é percetível no Código 26, o primeiro passo foi a criação do objeto CDE, como já referido anteriormente. A criação deste objeto permite o armazenamento de diversa informação e apresenta grande utilidade no processo de perturbação.

De seguida, utiliza-se a função **print** de forma a perceber quantas observações originais não satisfazem as condições de *K-anonymity*, e pelo resultado obtido no Código 26, cerca de 939 observações violam as condições de *2-Anonymity*, isto significa, que estas observações possuem combinações únicas de variáveis chave. É também apresentado o número de observações que não cumprem as condições de *3-Anonymity* e *5-Anonymity*.

Utilizando novamente a função **print**, é possível determinar o risco global original, como está apresentado na linha 15 do Código 26. Com esta função obtém-se o número esperado de identificações individuais, cerca de 544 observações, o risco de identificação global dado em percentagem, 5% aproximadamente. Como a base de dados possui uma estrutura hierárquica, é também apresentado o risco de identificação hierárquico global, cerca de 15%.

```
> sdc <- createSdcObj(eusilcS,
+                     keyVars = c("db040","pb220a","hsize","age"),
+                     numVars = c("netIncome", "py130n", "py100n","py090n"),
+                     pramVars = c("pl030"),
+                     sensibleVar = c("netIncome", "rb090"),
+                     weightVar="rb050", hhId = "db030")
> print(sdc)
Infos on 2/3-Anonymity:
Number of observations violating
  - 2-anonymity: 939 (8.009%)
  - 3-anonymity: 1605 (13.689%)
  - 5-anonymity: 2531 (21.586%)

---------------------------------------------------------------------
> print(sdc, "risk")
Risk measures:

Number of observations with higher risk than the main part of the data: 1476
Expected number of re-identifications: 544.28 (4.64 %)

Information on hierarchical risk:
Expected number of re-identifications: 1773.52 (15.13 %)
---------------------------------------------------------------------

> sdc <- ldiversity(sdc)
> sdc@risk$ldiversity
--------------------------

L-Diversity Measures
--------------------------

 netIncome_Distinct_Ldiversity rb090_Distinct_Ldiversity
 Min.   : 1.00                 Min.   :1.00
 1st Qu.: 1.00                 1st Qu.:1.00
 Median : 7.00                 Median :2.00
 Mean   :12.96                 Mean   :1.59
 3rd Qu.:21.00                 3rd Qu.:2.00
 Max.   :70.00                 Max.   :2.00
```

Código 26: Cálculo das medidas de risco de identificação

Para além da avaliação do risco de identificação, no Código 26 calcula-se a medida *l-diversity*, que é uma avaliação importante de se realizar, pois, caso não seja cumprido o valor pré-definido para esta medida, a variável sensível fica exposta à identificação do seu valor. Ao executar a função **ldiversity** obtém-se as principais propriedades da medida *l-diversity*. Neste caso, é percetível que a variável sensível **rb090** apresenta valores desta medida relativamente baixos, pois a média é cerca de 1,59 e o valor máximo é 2, indicando que esta variável contém um risco de identificação bastante elevado, pois para a mesma combinação de variáveis chave, a variável sensível apresenta no máximo duas categorias distintas. Por outro lado, a variável sensível **netIncome** possui um risco de identificação reduzido, pois existe uma maior variabilidade nos valores obtidos para a medida *l-diversity*.

Após algumas análises iniciais acerca do risco de identificação da base de dados em estudo, aplicam-se os

métodos perturbativos. Como já foi visto anteriormente, existem diversas funções e diversos métodos para a perturbação de uma base de dados e neste caso, opta-se por exemplificar com apenas dois métodos, o método PRAM e o método de Adição de Ruído. Os parâmetros escolhidos para estes métodos são meramente exemplificativos e como tal os resultados obtidos poderão não apresentar a eficácia desejada.

No Código 26 apresenta-se o método PRAM, define-se o argumento *pd* com o valor $0,2$, ou seja, cada categoria apresenta uma probabilidade de $0,2$ de permanecer inalterada, e aplica-se uma perturbação de $0,8$ na variável **pl030**.

Como é percetível pelo Código 27, ao executar a função **print** para o argumento *pram*, é apresentado a matriz de transição, o número de alterações realizadas (neste caso 3407 categorias trocaram entre si) e ainda a percentagem de alterações provocadas pelo método de PRAM, cerca de 29.06% para este caso. Comparações adicionais podem ser obtidas através do código *sdc@pram$comparison*.

```
> sdc_pram <- pram(sdc, pd = c(0.2) ,alpha = 0.8)
> print(sdc_pram, "pram")
Post-Randomization (PRAM):
Variable:pl030
--> final Transition-Matrix:
          1           2           3           4           5
1 0.6966678 0.07687171 0.031206416 0.039328104 0.07327371
2 0.3691773 0.32815387 0.041717432 0.052655256 0.09802322
3 0.3510305 0.09771249 0.389739998 0.025355153 0.05338635
4 0.2739294 0.07636756 0.015700045 0.543738030 0.03137379
5 0.1310640 0.03650865 0.008489156 0.008056874 0.78623510
6 0.3996625 0.11098933 0.036893304 0.043089231 0.08369762
7 0.3039635 0.08444819 0.026916733 0.030849089 0.06055860
            6          7
1 0.010544933 0.07210734
2 0.014063717 0.09620924
3 0.010949617 0.07182594
4 0.007918716 0.05097250
5 0.003950015 0.02569621
6 0.232580581 0.09308748
7 0.010353417 0.48291045
Changed observations:
  variable nrChanges percChanges
1    pl030       3407        29.06
```

Código 27: Aplicação do método PRAM

Após a aplicação do método PRAM às variáveis categóricas, aplica-se agora o método de Adição de Ruído às variáveis contínuas, como é apresentado no Código 28.

Para o exemplo em questão, aplica-se o método de adição de ruído não correlacionado (linha 1), adição de ruído tendo em conta a dimensão da base de dados (linha 11) e adição de ruído correlacionado (linha 21 e 31). Estes métodos são aplicados separadamente, de forma a perceber qual será o método mais indicado para a base de dados em estudo. Após a aplicação de cada método é possível ter acesso ao risco de identificação, que é fornecido na forma de um intervalo, e a duas medidas de perda de informação, a diferença de valores próprios e a medida IL1.

```
> sdc_add <- addNoise(obj = sdc_pram, noise = 160, method = "additive")
> print(sdc_add, "numrisk")
Numerical key variables: netIncome, py130n, py100n, py090n

Disclosure risk is currently between [0.00%; 18.79%]

Current Information Loss:
  - IL1: 2571579.24
  - Difference of Eigenvalues: 43.280%
-------------------------------------------------------------------------
> sdc_res <- addNoise(obj = sdc_pram, noise = 160, method = "restr")
> print(sdc_res, "numrisk")
Numerical key variables: netIncome, py130n, py100n, py090n

Disclosure risk is currently between [0.00%; 18.79%]

Current Information Loss:
  - IL1: 543549.10
  - Difference of Eigenvalues: 49.170%
-------------------------------------------------------------------------
> sdc_co <- addNoise(obj = sdc_pram, noise = 160, method = "correlated")
> print(sdc_co, "numrisk")
Numerical key variables: netIncome, py130n, py100n, py090n

Disclosure risk is currently between [0.00%; 18.79%]

Current Information Loss:
  - IL1: 1923152.50
  - Difference of Eigenvalues: 0.960%
-------------------------------------------------------------------------
> sdc_co2 <- addNoise(obj = sdc_pram, noise = 160, method = "correlated2")
> print(sdc_co2, "numrisk")
Numerical key variables: netIncome, py130n, py100n, py090n

Disclosure risk is currently between [0.00%; 20.42%]

Current Information Loss:
  - IL1: 153921.39
  - Difference of Eigenvalues: 0.360%
-------------------------------------------------------------------------
```

Código 28: Aplicação do método de Adição de Ruído

Através do Código 28 é possível fazer uma comparação rápida sobre qual o melhor método, isto é, o método que apresenta menor risco de identificação e menor perda de informação. Para o caso em estudo, é muito claro que a adição de ruído correlacionado apresenta os melhores resultados, e dos dois métodos possíveis escolhe-se o método *correlated2*, porque apesar de ter um risco de identificação maior, ($20,42\%$ em comparação com $18,79\%$), apresenta menor perda de informação ($0,360\%$ em comparação com $0,960\%$ para a medida de diferença de valores próprios e para a medida IL1 o método escolhido apresenta o valor de $153921,39$) e para o caso em estudo a prioridade está na utilidade da informação obtida.

O Código 28 retrata o risco de identificação das variáveis numéricas presentes na base de dados perturbada,

enquanto que o Código 29 apresenta o cálculo do risco de identificação aplicando o algoritmo SUDA, ou seja, o risco de identificação calculado a partir das *frequency counts*.

Como é visível pelo Código 28, onde se aplica a função para o cálculo dos DIS-SUDA *scores*, através do argumento *suda2* do objeto CDE obtém-se o número de observações presentes em cada intervalo de risco de identificação. Para a base de dados em estudo percebe-se que existem 10786 observações com risco de identificação igual a zero e cerca de 939 observações no intervalo de 0 a 0.1 do risco de identificação. Para além desta informação é apresentado ainda o argumento *contribution*, que indica o contributo de cada variável para o cálculo do risco total das observações.

Para além destes valores, são calculadas ainda três medidas de perda de informação, a partir da linha 32. É possível concluir que neste caso existe pouca perda de informação no processo de perturbação, pois a diferença dos valores próprios é mínima (0,0035) e tanto a medida IL1 como a medida IL1s não apresentam números elevados.

```
> sdc <- suda2(sdc_co2, original_scores=TRUE)
> sdc@risk$suda2

Dis suda scores table:
- - - - - - - - - - - -
    Interval Number of records
1       == 0              10786
2 (0.0, 0.1]                939
3 (0.1, 0.2]                  0
4 (0.2, 0.3]                  0
5 (0.3, 0.4]                  0
6 (0.4, 0.5]                  0
7 (0.5, 0.6]                  0
8 (0.6, 0.7]                  0
9      > 0.7                  0
- - - - - - - - - - - -
Attribute contribution:
- - - - - - - - - - - -
  variable contribution
1    pl030      70.96774
2   pb220a      51.01868
3    hsize      75.72156
4      age      98.72666
- - - - - - - - - - - -

> sdc_co2@utility
$il1
[1] 153921.4
$il1s
[1] 2161.582
$eigen
[1] 0.003561688
```

Código 29: Calculo dos SUDA *scores* e medidas de perda de informação

Por fim, no Código 30 apresenta-se o cálculo do risco de identificação intervalar, tanto pela distância euclidiana, função **dRisk**, como pela distância de Mahalanobis, função **dRiskRMD**. Para a primeira medida o resultado obtido foi 0.363, ou seja, cerca de 36% das observações perturbadas estão contidas no intervalo definido em torno das observações originais. Quanto à segunda medida, é percetível que nenhum valor perturbado está contido no intervalo

definido para as observações originais, o que significa que cada valor perturbado contém pelo menos m observações na sua vizinhança. Também se apresentam as observações iniciais do vetor indicador de observações inseguras (linha 11), que é zero para todas as observações e, portanto, não existem observações inseguras.

Para além da avaliação destes riscos de identificação, é necessário a avaliação do risco global, como é apresentado no Código 30, na linha 12, e é visível que o risco global da base de dados diminui de $4,00\%$ para $2,64\%$. Por outro lado, o risco hierárquico global diminui substancialmente de 15% para 5%. Estas diminuições permitem concluir que serão valores seguros do ponto de vista da identificação, já que se obtiveram valores baixos do ponto de vista da perda de informação. Assim, é possível afirmar que se realizou uma perturbação eficaz.

Concluindo-se que a base de dados cumpre os requisitos necessários e para a divulgação é necessário a extração da base de dados perturbada a partir do objeto CDE, como é apresentado no Código 29 na linha 37.

```
1  > sdc <- dRisk(sdc, k=0.1)
2  > sdc@risk$numeric
3  [1] 0.3630704
4  > sdc <- dRiskRMD(sdc)
5
6  > sdc@risk$numericRMD$risk1
7  [1] 0
8  > sdc@risk$numericRMD$wrisk1
9  [1] 0
10 > head(sdc@risk$numericRMD$riskvec1)
11 [1] 0 0 0 0 0 0
12 > sdc@risk$global
13 $risk
14 [1] 0.02642078
15
16 $risk_ER
17 [1] 244.2837
18
19 $risk_pct
20 [1] 2.642078
21
22 $threshold
23 [1] 0.01601718
24
25 $max_risk
26 [1] 0.01
27
28 $hier_risk_ER
29 [1] 773.518
30
31 $hier_risk
32 [1] 0.0512595
33
34 $hier_risk_pct
35 [1] 5.12595
36
37 BasePert <- extractManipData(sdc, ignoreKeyVars = F, ignorePramVars = F,
38                               ignoreNumVars = F, ignoreStrataVar = F)
```

Código 30: Cálculo do risco de identificação final

7. Caso de Estudo

O objetivo deste Capítulo é a aplicação e comparação de alguns métodos perturbativos na base de microdados PT2020 [2]. A base de microdados em estudo está anonimizada, isto é, não possui identificadores diretos e o número de identificação das entidades é uma combinação aleatória de digítos. Inicialmente descreve-se todas as variáveis presentes na base de microdados.

O primeiro passo foi escolher as variáveis chaves e as variáveis sensíveis da base de microdados a perturbar. Utilizando a linguagem de programação R avalia-se o risco de identificação da base de microdados original. De seguida, é aplicado os métodos de CDE a variáveis categóricas, avaliando-se as variáveis perturbadas quanto à perda de informação e ao risco de identificação. O mesmo procedimento é realizado para as variáveis numéricas, apresentado-se os resultados para diversos parâmetros a definir na perturbação. São também apresentados os resultados e as estatísticas principais das variáveis perturbadas, permitindo uma avaliação detalhada de cada método perturbativo apresentado. Por fim, procede-se às conclusões sobre qual o método perturbativo adequado à base de microdados PT2020.

7.1 Descrição da Base de Microdados

A base de microdados PT2020 resulta do Acordo de Parceria que Portugal propõe à Comissão Europeia, denominado Portugal 2020, que adota princípios de programação da Estratégia Europa 2020 e consagra a política de desenvolvimento económico, social, ambiental e territorial que estimulará o crescimento e a criação de emprego nos próximos anos em Portugal. Portugal 2020 define as intervenções, os investimentos e as prioridades de financiamento necessárias para promover no nosso país o crescimento inteligente, sustentável e inclusivo e o cumprimento das metas da Europa 2020 [19]. A base de microdados PT2020 é disponibilizada pelo BPLIM para efeitos de investigação sobre o protocolo com a Agência para o Desenvolvimento e Coesão e a Autoridade de Gestão do Programa Operacional Competividade e Internacionalização- COMPETE 2020. Os dados correspondem ao período de 2014 até maio de 2020, com 43333 observações e 60 variáveis [2].

Esta base de microdados é constituída pelas seguintes variáveis:

Identificadores Anonimizados

- **Código de operação anonimizado** (*codopa*): Identificador único da operação, esta variável encontra-se anonimizada;
- **Identificador da Firma** (*tina*): Número de identificação fiscal anonimizado da entidade.

Características do Projeto

- **Data da candidatura** (*datacand*): Mês e ano em que o beneficiário submeteu a candidatura de financiamento;
- **Fundo Europeu Estrutural e de Investimento** (*fundo*): Variável categórica com informação acerca do fundo europeu de investimento e estrutural;

Tabela 5: Variável *fundo*

Categoria	Significado
1	Fundo Europeu de Desenvolvimento Regional
2	Fundo Social Europeu

- **Sistema de Incentivos** (*instrumento*): Sistema de incentivo para qual o projeto foi submetido;

Tabela 6: Variável *instrumento*

Categoria	Significado
1	Formação Autónoma
2	Sistema de Incentivos à I&D Empresarial
3	Sistema de Incentivos à Inovação Empresarial
4	Sistema de Incentivos à Qualificação e Internacionalização de PME

- **Programa Operacional** *(pofinan):* Programa Operacional (PO) do projeto submetido;

Tabela 7: Variável *profinan*

Categoria	Significado
1	PO Alentejo
2	PO Algarve
3	PO Competitividade e Internacionalização
4	PO Centro
5	PO Lisboa
6	PO Norte

- **Aviso** *(aviso):* As candidaturas são submetidas numa chamada específica de propostas. A chamada inclui informações acerca dos prazos de submissão de candidaturas e as condições de financiamento. Esta variável está codificada de forma a proteger informação confidencial;

- **Ano do aviso de abertura de candidatura** *(ano_aac):* Ano da chamada para candidaturas;

- **Objetivo Temático** (*ot*): Variável categórica que representa o objetivo temático do projeto definido de acordo com o Artigo 9 de Regulamento (UE) No 1303/2013;

Tabela 8: Variável *ot*

Categoria	Significado
1	OT 1 - Reforçar a investigação, o desenvolvimento tecnológico e a inovação
2	OT 3 - Reforçar a competitividade das pequenas e médias empresas
3	OT 8 – Promover a sustentabilidade e a qualidade do emprego e apoiar a mobilidade laboral

- **Prioridade de Investimento** (*pi*): Variável categórica correspondente ao maior nível de desagregação da variável *ot*;

Tabela 9: Variável *pi*

Categoria	Significado
1	PI 1.2 - Promoção do investimento das empresas em inovação e investigação
2	PI 3.1 - Promoção do espírito empresarial
3	PI 3.2 - Desenvolvimento e aplicação de novos modelos empresariais para as PME
4	PI 3.3 - Apoio à criação e alargamento de capacidades avançadas de desenvolvimento de produtos e serviços
5	PI 8.5 - Adaptação de trabalhadores, empresas e empresários à mudança

- **Pontuação do Projeto** (*pontuação*): Pontuação atribuida ao projeto calculado de acordo com a fórmula anunciada na chamada de candidaturas;
- **Data de aprovação** (*data_aprov*): Data da decisão de financiamento do projeto;
- **Data da última decisão** (*dataultdec*): Data da última decisão efectuada no projeto;
- **Medida** (*medida*): Medida específica dentro dos respetivos sistemas de incentivo da operação;

Tabela 10: Variável *medida*

Categoria	Significado
1	ADAPTAR PME
2	Formação Autónoma
3	Formação-Ação para PME
4	I&D - Copromoção - COVID-19
5	I&D - Individuais - COVID-19
6	I&D - Infraestruturas de Ensaio e Otimização- COVID-19
7	I&DT - Copromoção
8	I&DT - Copromoção - RCI
9	I&DT - Demonstradores Copromoção
10	I&DT - Demonstradores Individuais
11	I&DT - Individuais
12	I&DT - Individuais - RCI
13	I&DT - Internacionalização
14	I&DT - Núcleos Copromoção
15	I&DT - Núcleos Individuais
16	I&DT - Programas Mobilizadores
17	I&DT - Propriedade Industrial
18	I&DT - Vales
19	Inovação - Produtiva - COVID-19
20	Inovação - Empreendedorismo
21	Inovação - Produtiva
22	Inovação - RCI
23	Inovação - Vales
24	QI PME - Conjuntos
25	QI PME - Individuais
26	QI PME - Vales

- **Organismo Intermédio** (*organismo*): Organismo público ou privado onde a entidade gerente opera e deve cumprir certas tarefas de acordo com os beneficiários que implementam as operações.

Tabela 11: Variável *organismo*

Categoria	Significado
1	AEP (Associação Empresarial de Portugal)
2	AICEP (Portugal Global - Agência para o Investimento e Comércio Externo de Portugal)
3	AIP-CCI (Associação Industrial Portuguesa - Câmara de Comércio e Indústria)
4	ANI (Agência Nacional de Inovação)
5	CAP (Confederação dos Agricultores de Portugal)
6	CCP (Confederação do Comércio e Serviços de Portugal)
7	CEC (Conselho Empresarial do Centro)
8	CTP (Confederação do Turismo de Portugal)
9	IAPMEI (Agência para a Competitividade e Inovação)
10	PO CI (Programa Operacional da Competitividade e Internacionalização)
11	TP (Turismo de Portugal)

- **Tipologia de Operação** (*to*): Variável categórica com informação acerca do tipo de operação, isto é, o agrupamento temático de ações dentro das categorias de uma prioridade do investimento e da tipologia de intervenção;

Tabela 13: Variável *to*

Categoria	Significado
196	SI Sistema de Incentivos à Investigação e Desenvolvimento Tecnológico - Proteção da propriedade intelectual e industrial
198	SI Sistema de Incentivos à Investigação e Desenvolvimento Tecnológico - Projetos de I&DT Empresas
199	SI Sistema de Incentivos à Investigação e Desenvolvimento Tecnológico - Projetos demonstradores
200	SI Sistema de Incentivos à Investigação e Desenvolvimento Tecnológico - Programas mobilizadores
201	SI Sistema de Incentivos à Investigação e Desenvolvimento Tecnológico - Núcleos de I&DT
202	SI Sistema de Incentivos à Investigação e Desenvolvimento Tecnológico - Internacionalização da I&D;
203	SI Sistema de Incentivos à Investigação e Desenvolvimento Tecnológico - Vale I&D
204	SI Sistema de Incentivos à Investigação e Desenvolvimento Tecnológico - regime contratual
206	SI Inovação empresarial e empreendedorismo - Inovação Produtiva Não PME
207	SI Inovação Empresarial e empreendedorismo - Inovação Produtiva Não PME - regime contratual
212	SI Inovação empresarial e empreendedorismo - Empreeendedorismo qualificado e criativo - Projeto individual
214	SI Inovação Empresarial e empreendedorismo - Vale Empreendedorismo
217	SI Qualificação e internacionalização das PME - Projeto individual
218	SI qualificação e internacionalização das PME - Vale Internacionalização
219	SI qualificação e internacionalização das PME - Projeto conjunto de internacionalização das PME (exceto formação-ação)
221	SI Inovação empresarial e empreendedorismo - Inovação Produtiva PME
222	SI Inovação empresarial e empreendedorismo - Inovação Produtiva PME - regime contratual
223	SI qualificação e internacionalização das PME - Vale Inovação
224	SI qualificação e internacionalização das PME - Projeto conjunto de qualificação das PME (exceto formação-ação)
249	Formação para a inovação empresarial

- **Projeto Financiado** (*apoiado*): Variável binária que assume valor 0 quando o projeto não é financiado e valor 1 caso contrário. Caso a entidade não possua valor significa que a resposta final ainda não foi reportada;
- **Tipologida da Intervenção** (*ti*): Variável categórica com o tipo de intervenção, que agrega objetivos especificos do mesmo tipo de prioridade de investimento.

Tabela 12: Variável *ti*

Categoria	Significado
1	TI 47 – Atividades de I&D empresarial
2	TI 49 – Investimento empresarial em inovação de não PME
3	TI 51 – Empreendedorismo qualificado e criativo
4	TI 52 – Internacionalização das PME
5	TI 53 – Qualificação e inovação das PME
6	TI 60 - Formação de empresários e trabalhadores das empresas
7	TI B7 – CRII – Atividades de I&D Empresarial
8	TI B8 - CRII - Investimento empresarial em inovação de não PME
9	TI B9 – CRII – Qualificação e inovação das PME

- **Domínio de Intervenção** *(dom_interv):* Domínio de intervenção de acordo com a Comissão Europeia (Regulamento (UE) No 1303/2013 do Parlamento Europeu e do Concelho, 17 Dezembro de 2013).

Tabela 14: Variável *dom_interv*

Categoria	Significado
1	Investimento produtivo genérico em pequenas e médias empresas («PME»)
2	Processos de investigação e inovação em grandes empresas
56	Investimento em infraestruturas, capacidades e equipamento em PME diretamente ligadas a atividades de investigação e de inovação
57	Investimento em infraestruturas, capacidades e equipamento em grandes empresas diretamente ligadas a atividades de investigação e de inovação
62	Transferência de tecnologia e cooperação entre universidades e empresas, sobretudo em benefício das PME
63	Apoio a grupos de empresas (clusters) e redes de empresas, sobretudo em benefício das PME
64	Processos de investigação e inovação nas PME (incluindo «vales», processos, conceção, serviços e inovação social)
66	Serviços avançados de apoio a PME e grupos de PME (incluindo serviços de gestão, marketing e design)
67	Desenvolvimento das atividades das PME, apoio ao empreendedorismo e incubação, incluindo apoio a empresas derivadas (spin-outs) e a novas empresas (spin-offs)
68	Eficiência energética e projetos de demonstração nas PME e medidas de apoio
74	Desenvolvimento e promoção de ativos comerciais turísticos em PME
75	Desenvolvimento e promoção de serviços comerciais turísticos em ou para PME
76	Desenvolvimento e promoção de ativos culturais e criativos em PME
77	Desenvolvimento e promoção de serviços culturais e criativos em ou para PME
106	Adaptação dos trabalhadores, das empresas e dos empresários à mudança

- **Projeto Ativo** (*nprojaprov*): Variável binária que assume valor 1 caso o projeto se encontre ativo e 0 caso contrário;

- **Data do termo de aceitação** (*datatemoaceit*): Data da formalização do contrato que estabelece as condições de financiamento;

- **Desistência/Anulação** (*dstanl*): Variável categórica que indica se o projeto foi anulado ou se a entidade desistiu do financiamento;

Tabela 15: Variável *dstanl*

Categoria	Significado
1	Anulação
2	Desistência

- **Fonte dos Dados** (*fonte*): Variável categórica com informação relativa à fonte dos dados;

Tabela 16: Variável *fonte*

Categoria	Significado
1	SGO (Sistema de informação do COMPETE 2020)
2	SIFSE (Sistema integrado do Fundo Social Europeu)

- **Estado da Candidatura registada no SIFSE** *(estadofse):* Classifica o estado da candidatura. Esta variável apenas é observada em entidades cuja fonte é a SIFSE;

Tabela 17: Variável *estadofse*

Categoria	Significado
1	A Aguardar Decisão
2	A Aguardar Decisão de Saldo Final
3	Aceite / Entidade Notificada
4	Aceite pela Entidade
5	Arquivada
6	Caducado
7	Com Análise Tec. Fin. de Saldo Final
8	Com Confirmação da Proposta de Decisão
9	Com Parecer do Responsável Saldo Final
10	Com Pedido Saldo Final Submetido
11	Com Saldo Final Aceite
12	Em Execução
13	Em Execução - A Aguardar Autorização para Emissão 1º adiant.
14	Em Execução - Autorização de Pagamento - 1º adiant. Emitido
15	Em Execução - Autorização de Pagamento - 1º adiant. Por Emitir
16	Extinta
17	Indeferida
18	Proposta para Arquivamento
19	Proposta para Caducidade
20	Proposta para Extinção
21	Proposta para Indeferimento

- **Projeto com acordo** *(comta*): Variável binária que assume valor 1 caso a operação funcione através de um acordo, onde os correspondentes se compromentem com o beneficiário para implementar as operações de acordo com as condições definidas, e valor 0 caso contrário;

- **Investimento associado à candidatura** (*investcand*): Despesas totais co-financiadas, não co-financiadas e as não elegíveis necessárias de forma a cumprir os objetivos estabelecidos. Esta variável é medida em euros;

- **Incentivo contratado** *(incentivocontrat):* Valor do incentivo financiado pelo *European Regional Development Fund* (ERDF) ou do Fundo Social Europeu (ESF). A variável é medida em euros;

- **Projeto Indústria 4.0** *(proj_i40):* Variável binária que assume valor 1 caso o projeto pertença à operação *Industry 4.0* e valor 0 caso contrário;

- **Classificação da atividade económica do projeto** *(cae_projeto):* Classificação da Atividade Económica (CAE) da operação definida dentro das atividades económicas do promotor. Reporta códigos de 5 digítos.

- **Transacionável** (*transacionavel*): Variável categórica do tipo de transações de mercadorias e serviços da principal atividade económica.

Tabela 18: Variável *transacionavel*

Categoria	Significado
1	Bens Transacionáveis
2	Serviços Internacionalizáveis
3	Serviços Não Mercantis
4	Serviços Não Transacionáveis

- **Intensidade tecnológica e de conhecimento** (*intensidadetecnologica*): Variável categórica com o nível de tecnologia e conhecimento da principal atividade económica da operação.

Tabela 19: Variável *intensidadetecnologica*

Categoria	Significado
1	Alta intensidade tecnológica
2	Baixa intensidade tecnológica
3	Média-alta intensidade tecnológica
4	Média-baixa intensidade tecnológica
5	Outros serviços com forte intensidade de conhecimento
6	Outros serviços com fraca intensidade de conhecimento
7	Serviços de alta tecnologia com forte intensidade de conhecimento
8	Serviços financeiros com forte intensidade de conhecimento
9	Serviços de mercado com forte intensidade de conhecimento
10	Serviços de mercado com fraca intensidade de conhecimento

- **Tecnologias de informação e comunicação** (*tic*): Variável categórica que indica a classificação das tecnologias de informação e comunicação da principal atividade económica da operação.

Tabela 20: Variável *tic*

Categoria	Significado
1	Fabricação TIC
2	Serviços Intangíveis TIC
3	Serviços relacionados a bens TIC

Variáveis no momento de decisão

- **Data de início do projeto** (*dec_datainiproj*): Data prevista para o início do projeto, na data de decisão. A data de início do projeto corresponde à data de início do investimento;
- **Data de término do projeto** (*dec_datafimproj*): Data prevista para o término do projeto;
- **Investimento aprovado** (*dec_investaprov*): Montante total em euros decidido no momento de aprovação do projeto;
- **Investimento aprovado elegível** (*dec_investeleg*): Montante total de despesas aprovadas a serem consideradas pela comunidade co-financiada. A variável é expressa em euros;
- **Despesas públicas aprovadas** (*dec_despesapublica*): Montante total de despesas públicas, em euros, aprovadas no momento de decisão do projeto;
- **Incentivo aprovado** (*dec_incentivoaprov*): Montante total de incentivos aprovados, em euros, no momento de aprovação do projeto;
- **Incentivo reembolsável** (*dec_incentivoreemb*): Empréstimo garantido ao beneficiário, que se sujeita a um reembolso de acordo com o prazo estabelecido na operação. A variável é expressa em euros;
- **Incentivos não reembolsáveis reembolsavel** (*dec_incentivonaoreemb*): Montante de pagamentos não reembolsáveis ao beneficiário de acordo com o cumprimento de objetivos. A variável é expressa em euros;

Variáveis no momento de encerramento

- **Data de aprovação do encerramento** (*dataenc*): Data de encerramento do projeto;
- **Projeto com encerramento** (*comenc*): Variável binária que assume valor 1 caso o projeto se encontre terminado e valor 0 caso contrário;
- **Data do começo do projeto** (*enc_datainiproj*): Data do começo do projeto no momento de encerramento;
- **Data de encerramento do projeto** (*enc_datafimproj*): Data do encerramento da operação no momento de encerramento;
- **Investimento aprovado** (*enc_investaprov*): Montante total de investimento aprovado no momento de encerramento do projeto. A variável é expressa em euros;
- **Investimento elegível aprovado** (*enc_investaprov*): Montante total de despesas elegíveis aprovadas e consideradas para a comunidade co-financiada no momento de encerramento do projeto. A variável é expressa em euros;
- **Despesas públicas aprovadas** (*enc_despesapublica*): Montante total de despesas públicas aprovadas no momento de encerramento do projeto. A variável é expressa em euros;
- **Incentivos aprovados** (*enc_incentivoaprov*): Montante total de incentivos aprovados no momento de encerramento do projeto. A variável é expressa em euros;
- **Incentivos reembonsáveis** (*enc_incentivoreemb*): Montante dos empréstimos fornecidos ao beneficiário, livre de juros de acordo com os prazos estabelecidos no contrato. A variável é expressa em euros e reportada no momento de encerramento do projeto;

- **Incentivos não reembonsáveis** (*enc_incentivonaoreemb*): Incentivos pagos ao beneficiários, de acordo com o cumprimento de objetivos no fecho da operação. A variável é expressa em euros;

Variáveis no momento de execução

- **Projeto com execução** (*comexecucao*): Variável binária que assume valor 1 caso a operação esteja a ser implementada e 0 caso contrário;
- **Investimento elegível executado** (*exec_investeleg*): Montante total de investimento implementado até ao momento. A variável é expressa em euros;
- **Despesas públicas executadas** (*exec_despesapublica*): Montante total de despesas públicas que foram implementadas até ao momento. A variável é expressa em euros;
- **Financiamento executado** (*exec_incentivofundo*): Montante total de financiamento implementado. A variável é expressa em euros;
- **Outras fontes de financiamento executados** (*exec_outrasfontesdinanc*): Montante total de financiamento implementado por outras fontes para além do Fundo Europeu de Investimento. A variável é expressa em euros;

Pagamentos

- **Projeto com pagamentos** (*compagamentos*): Variável *dummy* que assume valor 1 caso o projeto tenha recebido pagamentos e 0 caso contrário;
- **Pagamentos Totais** *(totpagam_realizado)*: Montante de pagamentos realizados, expressos em euros;
- **Pagamentos adiantados certificáveis** (*pagadiantcertif*): Montante de pagamentos adiantados ao beneficiário, expressos em euros;
- **Pagamentos adiantados não certificáveis** (*pagadiantnaocertif*): Montante de pagamentos adiantados ao beneficiário que não cumprem os requisitos de despesas submetidas na comissão europeia. Esta variável é expressa em euros;

7.2 Escolha das variáveis chave

Tendo em conta que as duas variáveis identificadoras presentes na base de microdados se encontram anonimizadas, o próximo passo é a determinação das variáveis chave. Este processo é de grande importância pois as variáveis, escolhidas com base em cenários de identificação, indicarão o risco de identificação individual e global da base de microdados. As variáveis chave são escolhidas tendo em conta as características que indiretamente possam permitir a identificação de uma observação por si só, ou por cruzamento com outras variáveis de bases de microdados externas.

Como primeira opção considera-se o seguinte conjunto de variáveis:

Opção 1: *aviso, compagamentos, pofinan, medida, apoiado, dom_interv, cae_projeto, to*,

que correspondem na sua maioria a variáveis disponíveis em informação pública. Como é visível pela Tabela 21, cerca de 46% das observações não cumprem a condição da medida *2-anonymity*, ou seja, são combinações únicas, o que corresponde a um elevado risco de identificação. Após uma análise a este conjunto de variáveis, percebe-se que a variável *cae_projeto* é a variável que apresenta maior detalhe e como tal é responsável pela existência de grande parte das combinações únicas obtidas. Assim, reduz-se a variável *cae_projeto* de 5 digítos para 2 dígitos (*div*, que corresponde à divisão da CAE) e obtém-se a segunda opção:

Opção 2: *aviso, compagamentos, pofinan, medida, apoiado, dom_interv, div, to*.

Novamente pela Tabela 21 é visível que o número de observações que não cumprem a condição da medida *2-anonymity* ainda é elevado (cerca de 20%) e cerca de 47% das observações da base de microdados não cumprem a condição de *5-anonymity*, ou seja, as combinações de variáveis chave são partilhadas apenas por 4 ou menos observações. Como estes valores ainda são elevados, o que indica que o risco de identificação também é elevado, é necessário novas alterações nas variáveis de forma a reduzir o risco de identificação.

Como tentativa de redução do risco de identificação, altera-se a variável *div* para a variável *seccao* que corresponde à secção da CAE, obtendo-se a terceira opção:

Opção 3: *aviso, compagamentos, pofinan, medida, apoiado, dom_interv, seccao, to*.

Como é visível pela Tabela 21, os valores da medida *k-anonymity* reduziram significativamente. Estes valores são satisfatórios, no entanto, ao efetuar-se a mudança da variável *cae_projeto* a perda de informação aumenta significativamente. Como o foco do responsável da base de microdados é a utilidade da informação, a mudança da variável *cae_projeto* provoca resultados não satisfatórios do ponto de vista da perda de informação. Como resultado, e pelo elevado nível de detalhe presente nas variáveis *cae_projeto* e *to*, opta-se pela remoção das mesmas na base de microdados divulgada e apenas serão disponibilizadas aos utilizadores mediante pedido.

Tabela 21: Observações que não cumprem as condições da medida *k-anonymity*

Opção	2-anonymity	3-anonymity	5-anonymity
1	45.667%	61.983%	75.707%
2	19.532%	31.752%	47.091%
3	6.231%	11.589%	20.333%

Após a remoção das duas variáveis na base de microdados, obtém-se a seguinte combinação de variáveis chave, opção final: *aviso, compagamentos, pofinan, medida, apoiado, dom_interv*. Neste momento os valores da medida *k-anonymity* não são elevados, como é visível pelos resultados no Código 31, o que leva a concluir que esta combinação é a combinação a utilizar para a perturbação da base de microdados.

A escolha destas variáveis chave deve-se ao facto de serem variáveis potencialmente identificadoras de um projeto e pelo facto de estarem disponíveis em documentos externos, nomeadamente *online*. Para além destas variáveis ainda existem outras que estão presentes em documentos externos, no entanto, após uma análise dos riscos individuais e globais conclui-se que tais variáveis não contribuem significativamente para o aumento dos riscos de identificação e não serão tidas em conta como variáveis chave. A única variável não disponível em documentos externos é a variável *compagamentos*, no entanto, escolhe-se como variável chave devido às suas características, uma vez que, apenas projetos com valor 1 possuem valores nas variáveis pagamentos, aumentando a probabilidade de identificação de uma observação. No Código 31 é apresentada a criação do objeto sdcf na linguagem R e os resultados para a medida de *k-anonymity*. Os valores reduziram significativamente em comparação com as combinações anteriores e para estas variáveis chaves existem apenas 243 combinações únicas.

```
> sens <- c("dec_investaprov","dec_investeleg",
 "dec_incentivoaprov", "investcand","totpagam_realizado")
> VarC <- c("apoiado", "aviso", "pofinan", "medida",
"compagamentos", "dom_interv")
> sdcf <- createSdcObj(df, keyVars = VarC, sensibleVar=sens)
> print(sdcf)
Infos on 2/3-Anonymity:
Number of observations violating
  - 2-anonymity: 243 (0.561%)
  - 3-anonymity: 593 (1.368%)
  - 5-anonymity: 1354 (3.125%)
---------------------------------------------------------------------
```

Código 31: Criação do objeto SDC e cálculo do risco de identificação original

Quanto às variáveis sensíveis, a informação confidencial da base de microdados em estudo está presente nas variáveis quantitativas *dec_investaprov, dec_investeleg, dec_incentivoaprov, investcand, totpagam_realizado*, que estão essencialmente relacionadas com os montantes de financiamento. A estas variáveis serão aplicados os métodos de CDE de forma a garantir a confidencialidade das mesmas.

7.3 Avaliação do Risco de Identificação

De seguida, apresentam-se os resultados da medida DIS-SUDA (Tabela 22) e do risco de identificação individual (Figura 6).

Tabela 22: Riscos de identificação individuais através dos *SUDA-scores*

DIS-SUDA score	Nº de observações
0	42965
]0;0.1]	366
]0.1;0.2]	2
>0.2	0

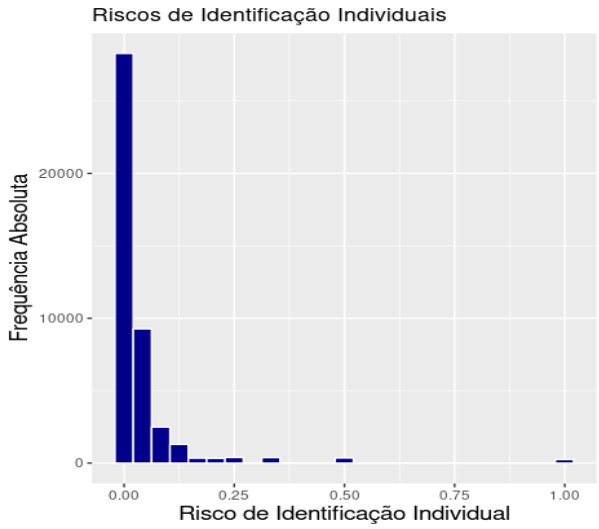


Figura 6: Histograma dos riscos de identificação individuais

Na Tabela 22 apresentam-se os resultados da medida *DIS-SUDA* (Ver Capítulo 3, Secção 1) da combinação final de variáveis chave, e pelos resultados apresentados é possível perceber que os riscos de identificação individuais calculados usando esta medida são significativamente baixos (apenas duas observações apresentam um risco superior a 0.1).

Na Figura 6 está apresentada a distribuição dos riscos individuais calculados a partir das *frequency counts* das combinações de variáveis chave. Para esta medida, grande parte das observações apresentam risco de identificação

individual inferior a 0.25. Como o cálculo destes riscos tem por base as *frequency counts* das combinações é possível afirmar que grande parte das combinações de variáveis chave contém pelo menos 4 observações, o que garante o mínimo de 3 observações geralmente estipulado.

```
> print(sdcf, "risk")
Risk measures:
Number of observations with higher risk than the main part of the data: 3305
Expected number of re-identifications: 1668.48 (3.85 %)
```

Código 32: Cáculo do risco de indentificação global

No Código 32 observa-se que o risco de identificação global é de 3.85%, indicando que aproximadamente 1669 observações podem ser identificadas. É ainda apresentado o número de observações que possui risco individual superior à maior parte da base de dados, isto é, observações que cumprem duas condições:

$$r_i > mediana(R) + 2 * mediana(|R - mediana(R)|) \wedge r_i > 0.1, \tag{7.1}$$

onde R representa o vetor dos riscos de identificação individuais e r_i o risco de identificação da observação i. No caso da base de microdados em estudo, cerca de 3305 observações possuem risco individual superior a estas duas condições, ou seja, cerca de 8% das observações cumprem as duas condições descritas. De seguida, realiza-se a perturbação das variáveis categóricas e como esta base de microdados possui um risco de identificação global significativamente reduzido, apenas se realiza a perturbação de variáveis chave com o objetivo de eliminar as combinações únicas.

7.4 Aplicação dos Métodos Perturbativos em Variáveis Categóricas

Após a avaliação do risco de identificação na base de microdados original, perturbam-se as variáveis categóricas de forma a diminuir a probabilidade de identificação de uma dada observação.

7.4.1 Supressão Local

Para a combinação de variáveis chave considerada, existem cerca de 243 combinações únicas e cerca de 593 combinações que não cumprem a condição da medida *3-anonymity* (Código 31). De forma a reduzir a perda de informação resultante deste processo, o foco será apenas eliminar as combinações únicas suprimindo as observações com risco individual superior a um determinado limite. Este limite é estabelecido de acordo com os riscos de identificação calculados e de acordo com as características da variável em causa. Assim, para diminuir a perda de informação resultante, aplica-se uma maior supressão a variáveis com menor importância, pois tais variáveis apresentam valores em falta.

No Código 33 está apresentado o número de supressões realizadas em cada variável. As variáveis *dom_interv* e *pofinan* possuem um elevado número de valores em falta, por isso, opta-se por estabelecer um limite de 0.4 para o risco de identificação individual, garantindo assim que estes riscos nunca serão superiores a 40% para essas variáveis. Por outro lado, para as variáveis *medida* e *pofinan* definiu-se o limite de 0.7, o que leva a um menor número de observações suprimidas.

```
> sdcf <- localSupp(sdcf, threshold=0.4, keyVar="apoiado")
> sdcf@localSuppression
```

```
$supps
   apoiado  aviso  pofinan  medida  compagamentos  dom_interv
1:     557      0        0       0              0           0
> sdcf <- localSupp(sdcf, threshold=0.7, keyVar="medida")
> sdcf@localSuppression
$supps
   apoiado  aviso  pofinan  medida  compagamentos  dom_interv
1:       0      0        0     201              0           0
> sdcf <- localSupp(sdcf, threshold=0.4, keyVar="dom_interv")
> sdcf@localSuppression
$supps
   apoiado  aviso  pofinan  medida  compagamentos  dom_interv
1:       0      0        0       0              0         404
> sdcf <- localSupp(sdcf, threshold=0.7, keyVar="pofinan")
> sdcf@localSuppression
$supps
   apoiado  aviso  pofinan  medida  compagamentos  dom_interv
1:       0      0      121       0              0           0
```

Código 33: Aplicação de Supressão Local

Para perceber as alterações provocadas nas variáveis em estudo, apresentam-se nas Figuras 7 e 8 os valores em falta de cada variável chave, antes e após a supressão das observações. Por análise dos gráficos é percetível que houve um aumento de apenas 0.5% de valores em falta no conjunto das variáveis chave, o que representa um valor significativamente reduzido.

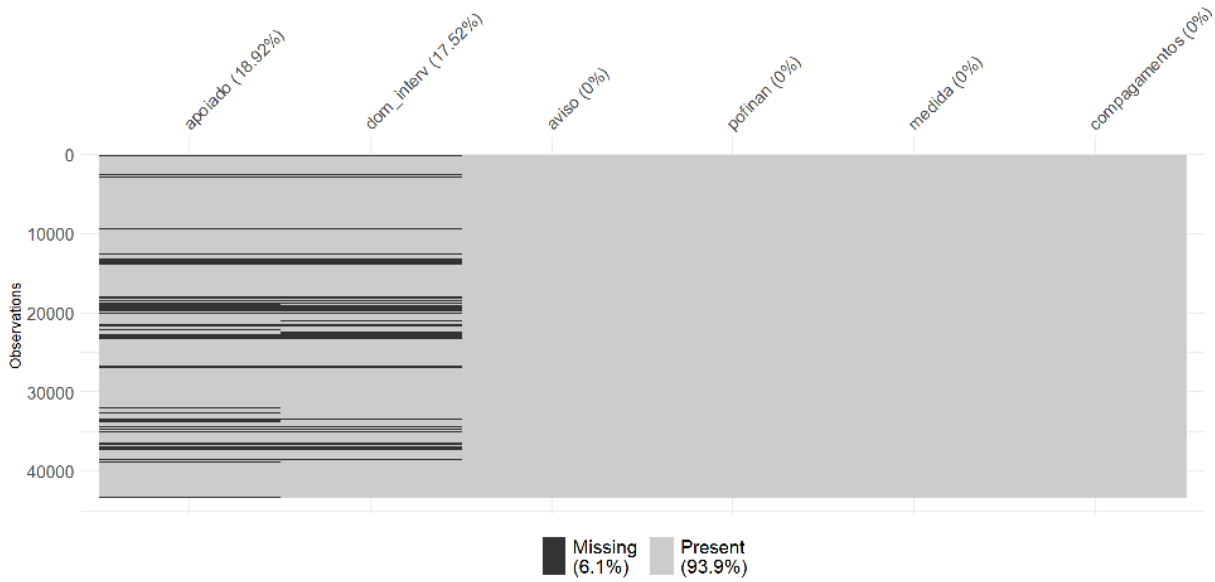


Figura 7: Gráfico de valores em falta antes da supressão

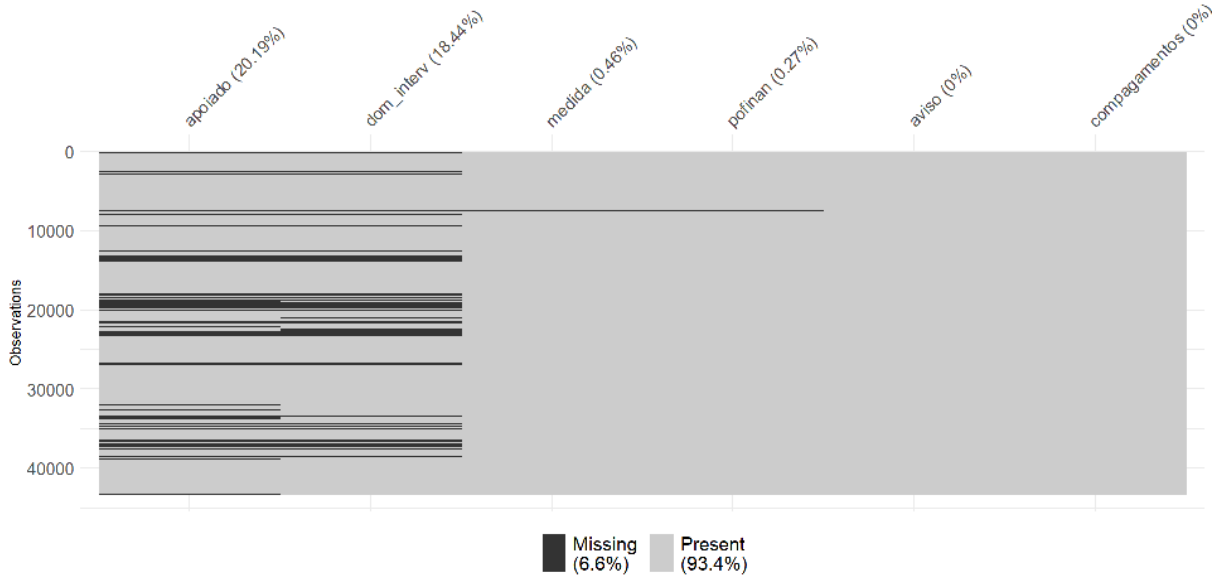


Figura 8: Gráfico de valores em falta após a supressão

A variável *apoiado* foi a que sofreu maior alteração, apresentava originalmente 18.92% de valores em falta e após a supressão das observações apresenta 20.19% de valores em falta. Já a variável *dom_interv* sofreu um aumento de apenas 0.92% de valores em falta. Quanto às variáveis que originalmente não apresentavam valores em falta, *medida* e *pofinan*, as suas versões perturbadas apresentam, respetivamente, 0.46% e 0.27% de valores em falta. Assim, por análise do Código 33 e das Figuras 7 e 8, é possível concluir que o método de supressão aplicado não provoca diferenças significativas na base de microdados.

No Código 34 é apresentada a medida de *k-anonymity* para a base de microdados perturbada, onde é visível que agora existem apenas 3 combinações únicas. É ainda apresentada a medida *DIS-SUDA*, e através dos resultados obtidos para estas medidas é possível concluir que a base de dados possui um nível baixo de risco de identificação, apesar de que existem 3 combinações únicas e 93 combinações que não cumprem a condição de *3-anonymity*.

```
> print(sdcf)
Infos on 2/3-Anonymity:

Number of observations violating
  - 2-anonymity: 3 (0.007%) | in original data: 243 (0.561%)
  - 3-anonymity: 93 (0.215%) | in original data: 593 (1.368%)
  - 5-anonymity: 744 (1.717%) | in original data: 1354 (3.125%)
---------------------------------------------------------------------
> sdcf <- suda2(sdcf)
> sdcf@risk$suda2

Dis suda scores table:
- - - - - - - - - - -
    Interval Number of records
1       == 0             43242
2 (0.0, 0.1]                91
3       >0.1                 0
```

Código 34: Medida *k-anonymity e DIS-SUDA scores*

Como o foco na perturbação é a utilidade dos dados, aplica-se uma última perturbação nas variáveis chave de forma a eliminar apenas as combinações únicas. Para que a perda de informação seja reduzida, aplica-se o método *Semantic Data Recoding* a estas 3 combinações únicas.

O primeiro passo é perceber quais as observações que contêm as combinações únicas de variáveis chave. Na Tabela 23 estão apresentadas as observações e as respetivas categorias das variáveis chave.

Tabela 23: Combinações únicas de variáveis chave

Observação	*apoiado*	*aviso*	*pofinan*	*medida*	*compagamentos*	*dom_interv*
28325	1	111	6	21	1	1
35338	1	107	5	15	1	56
37599	1	5	6	15	1	56

Com este método, eliminam-se as combinações únicas substituindo-as pelas combinações de variáveis chave com 2 ou 3 observações, isto é, substitui-se a categoria 111 por 122 na variável *aviso* da observação 28325, substitui-se a categoria 107 por 119 na variável *aviso* da observação 35338 e substitui-se a categoria 5 por 119 na variável *aviso* da observação 37599. Desta forma, substitui-se combinações únicas por combinações mais próximas, ou seja, que apenas diferem numa das variáveis, garantindo a menor perda de informação possível.

7.4.2 PRAM

Aplica-se o método PRAM às variáveis *fundo, estadofse* e *proj_i40*. A variável *fundo* está disponível em documentos externos, no entanto, não contribui significativamente para o aumento do risco de identificação quando é escolhida como variável chave. A variável *estadofse* possui valores apenas para projetos cuja fonte de informação é SIFSE e portanto aplica-se o método PRAM com a variável de estratificação *fonte* de forma a garantir a consistência nas diferentes fontes. A variável *proj_i40*, não sendo uma variável chave, pelo facto de se referir a uma característica muito especifica, contém informação sensível, e pode aumentar significativamente a probabilidade de identificação quando cruzada com outras variáveis na base de microdados.

As Tabelas 24, 25 e 26 representam as tabelas de contingência das variáveis a perturbar. Como é percetível na combinação das variáveis duas a duas, isto é, existem várias células com valores nulos, que correspondem a combinações de categorias que não se observam na base de dados. Assim, quando se aplica o método PRAM o ideal é que as tabelas de contingência das variáveis perturbadas preservem as células de valor nulo que se observam nas variáveis originais.

Tabela 24: Tabela de contigência das variáveis *proj_i40* e *pestadosfe*

proj_i40/ estadosfe	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
0	2	1	2	32	11	2	6	7	1	3	139	3	1	456	1	11	89	8	1	9	52
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0

Tabela 25: Tabela de contigência das variáveis *fundo* e *estadosfe*

fundo/ estadosfe	1	2	3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20	21
1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
2	2	1	2	32	11	2	6	7	1	3	139	3	1	456	1	11	89	8	1	9	52

Tabela 26: Tabela de contigência das variáveis *fundo* e *proj_i40*

fundo/ *proj_i40*	0	1
1	36692	2398
2	4243	0

De forma a que as tabelas de contingência das variáveis perturbadas apresentem as mesmas características das variáveis originais, a matriz de transição da variável *fundo* é definida como $\begin{bmatrix} 1 & 0 \\ 0.01 & 0.99 \end{bmatrix}$, ou seja, a probabilidade da categoria 1 permanecer inalterada é de 100% e a probabilidade da categoria 2 permanecer inalterada é de 99%. Para variável *proj_i40* a matriz de transição é definida como $\begin{bmatrix} 0.95 & 0.05 \\ 0 & 1 \end{bmatrix}$, ou seja, a probabilidade da categoria 0 permanecer inalterada é de 95% e a probabilidade da categoria 1 permanecer inalterada é de 100%. Já para a variável *estadofse* não se define matriz de transição, apenas a probabilidade de uma categoria permanecer inalterada, cerca de 75%. Após alguns testes, concluiu-se que as condições impostas preservam as células de valor nulo observadas nas variáveis originais, mantendo a consistência entre as variáveis, originando menor perda de informação. No Código 35 é apresentado a aplicação do método PRAM.

```
> mat <- matrix(c(1,0.01,0,0.99),ncol=2)
> rownames(mat) <- colnames(mat) <- c("1","2")
> sdcf <- pram(sdcf, variables="fundo", pd=mat)
> sdcf <- pram(sdcf, variables="estadofse", pd=c(0.75),
strata_variables = c("fonte"))
> mat1 <- matrix(c(1,0.05,0,0.95),ncol=2)
> rownames(mat) <- colnames(mat) <- c("0","1")
> sdcf <- pram(sdcf, variables="proj_i40", pd=mat1)
> print(sdc, "pram")
Changed observations:
   variable nrChanges percChanges
1     fundo        40        0.09
2 estadofse       103        0.24
3  proj_i40        98        0.23
```

Código 35: Aplicação do método PRAM

Como já referido, a variável *estadofse* apenas possui valores de acordo com a variável *fonte*, sendo necessário definir a variável de estratificação, o que permite que todas as observações com valores em falta desta variável (98.1%) não serão tidos em conta na aplicação do método PRAM. No Código 35 é ainda possível ver as alterações gerais que as variáveis sofreram, apresentando as alterações em percentagem e em valor absoluto, 40 alterações para a variável *fundo*, 103 alterações para a variável *estadofse* e 98 alterações para a variável *proj_i40*. Estas alterações são significativamente reduzidas, sendo o maior valor cerca de 0.24% das observações totais da base de micodados.

Na Figura 9 são apresentados os gráficos de barras das variáveis antes e após a aplicação do método PRAM apenas das categorias que são sujeitas a alteração e é percetível as alterações em cada categoria, que suportam os resultados obtidos no código. Como já foi referido a variável *estadofse* possui 98.1% de valores em falta, sendo o número de valores em falta antes e após a aplicação do método PRAM, exatamente o mesmo. Pelos resultados obtidos no Código 35 e da Figura 8, é possível concluir que a utilidade da informação neste caso é bastante elevada visto que o número de observações nas categorias perturbadas não diferem significativamente das categorias originais.

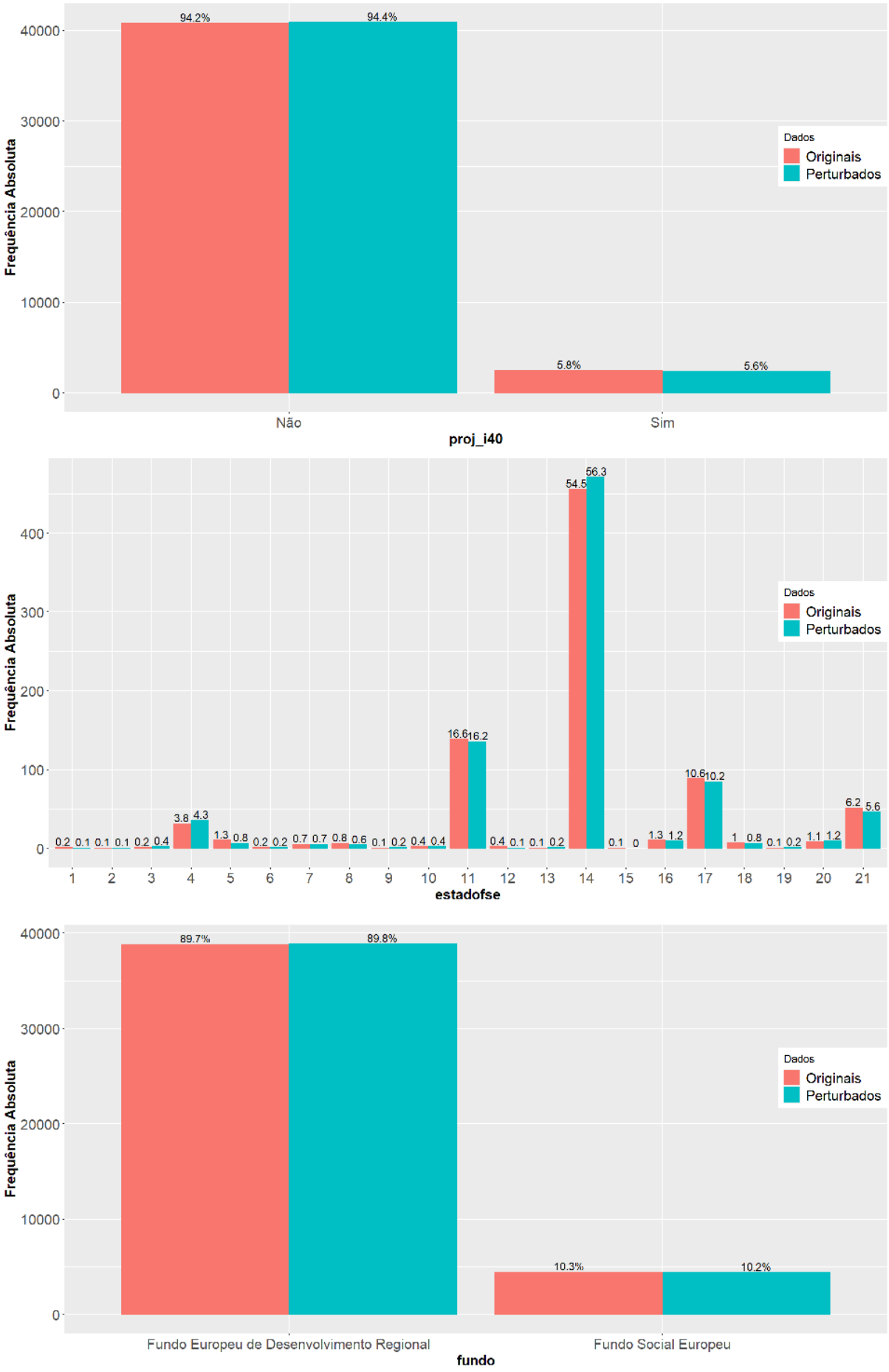


Figura 9: Gráficos de Barras das variáveis antes e após o método PRAM

7.4.3 Avaliação dos Métodos

Para concluir sobre os resultados da perturbação das variáveis categóricas, avaliam-se os métodos aplicados através do risco de identificação e da perda de informação resultante. No Código 36 é visível que o número de combinações únicas de variáveis chave é zero e o risco global da base de microdados é de apenas 3%. Neste caso, o número de observações que possuem um risco de identificação individual elevado, face à grande maioria das observações, diminui de 3305 para 2809, ou seja, cerca de 6.5% das observações. Antes da perturbação estes valores já eram relativamente baixos e após a perturbação ainda se tornaram mais reduzidos. Portanto, é possível concluir que os resultados são satisfatórios em termos de risco de identificação de uma dada observação. Através da medida *SUDA-scores* (Código 34) é possível concluir que o risco individual mais elevado é de cerca de 10%.

```
> print(sdcf)
Infos on 2/3-Anonymity:
Number of observations violating
  - 2-anonymity: 0 (0.000%) | in original data: 243 (0.561%)
  - 3-anonymity: 90 (0.208%) | in original data: 593 (1.368%)
  - 5-anonymity: 739 (1.705%) | in original data: 1354 (3.125%)
---------------------------------------------------------------------
> print(sdcf, "risk")
Risk measures:
Number of observations with higher risk than the main part of the data:
    in modified data:2809 | in original data: 3305
Expected number of re-identifications:
    in modified data:1300.84 (3.00 %) | in original data: 1668.48 (3.85 %)
```

Código 36: Cálculo dos riscos de Indentificação da base de microdados perturbada

Na Figura 10 é visível o histograma dos riscos de identificação individuais na base de microdados perturbada. O maior risco de identificação individual é cerca de 50% e as observações apresentam uma grande concentração para valores inferiores a 25%. A base de microdados possui assim um nível baixo de risco de identificação e que as diversas combinações formadas pelas variáveis chave são partilhadas por diversas observações, não existindo combinações únicas na base de microdados perturbada. Concluí-se então que as variáveis categóricas apresentam um nível seguro para divulgação.

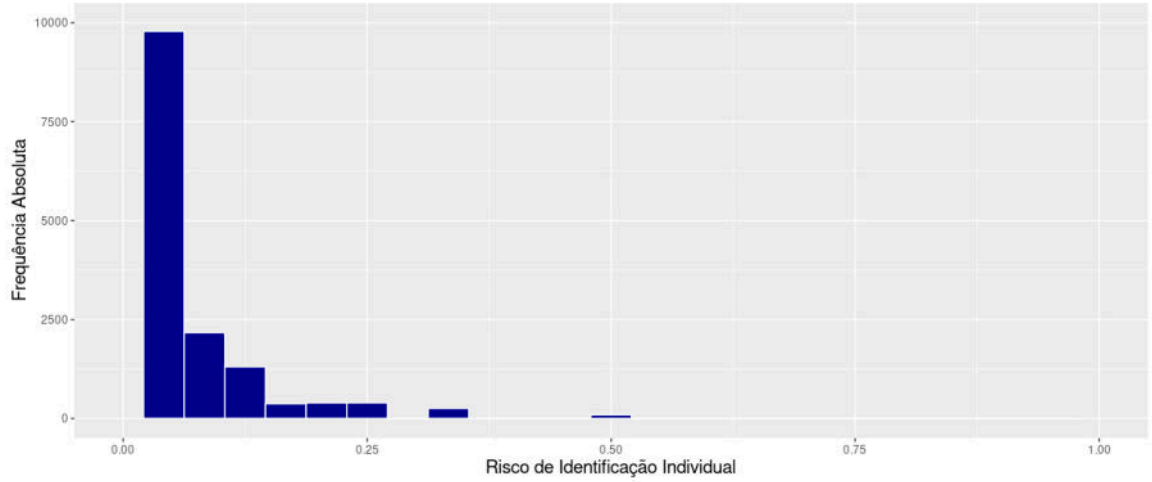


Figura 10: Histograma dos riscos de identificação individuais

Quanto à perda de informação, ao aplicar o método de supressão local existe um aumento de valores em falta. A medida de comparação dos valores em falta, apresentada no Capítulo 5, apresenta o seguinte valor:

$$M = 1267$$

ou seja, cerca de 3% das observações foram suprimidas. Com base na Figura 7, e na medida M, é possível afirmar que a utilidade da informação é alta após a aplicação do método de supressão local.

No Código 37 apresentam-se as alterações provocadas em cada categoria aplicando o método PRAM. Este código é suportado pela Figura 8 onde estão representados os gráficos de barras de cada variável. Como já foi referido, não existem grandes alterações nas categorias das variáveis após a aplicação do método PRAM. A maior diferença de observações numa categoria é na categoria 1 da variável *proj_i40*, cerca de 98 observações.

```
> sdcf@pram$comparison
$fundo
                            fundo     1    2 NA
1:            Original Frequencies 38856 4477  0
2: Frequencies after Perturbation 38896 4437  0
$estadofse
             estadofse 1 2 3  4  5 6 7 8 9 10  11 12 13  14 15 16 17 18
1:           Original  2 1 2 32 11 2 6 7 1  3 139  3  1 456  1 11 89  8
2: After Perturbation 1 1 3 36 7  2 6 5 2  3 136  1  2 471  0 10 85  7
   19 20 21    NA
1:  1  9 52 42496
2:  2 10 47 42496
$proj_i40
                         proj_i40     0    1 NA
1:            Original Frequencies 40810 2523  0
2: Frequencies after Perturbation 40908 2425  0
```

Código 37: Alterações provocadas pelo método PRAM

Através da medida de comparação das tabelas de contingência apresentada no Capítulo 5, calcula-se a perda de informação nestas variáveis. A Tabela 27 apresenta os valores da medida UT1, que mede a distância em percentagem entre os valores das tabelas de contingência considerando as variáveis duas a duas. Como é percetível o valor mais elevado é entre as variáveis *fundo*/*estadosfe* e *proj_i40*/*estadosfe* com cerca de 19% nos dois casos. Como a variável *estadosfe* apresenta 98.1% de valores em falta, as tabelas de contingência (Tabelas 24 e 25) contêm várias categorias com um número reduzido de observações (inferior a 3). Assim, qualquer alteração numa categoria da variável *estadofse* provoca um grande aumento na medida UT1, pois para o cálculo desta medida tem-se em conta o número de alterações e o número de observações presente nas tabelas originais, e quanto menor o número de observações presentes numa célula da tabela, maior será o impacto das alterações nessa célula. Como já se esperava os valores na diagonal da Tabela 27 são os valores mais baixos, devido ao pequeno número de alterações que as variáveis sofreram individualmente.

Assim, é possível concluir que na aplicação do método PRAM, com as matrizes de transição definidas, obteve-se a menor perda de informação possível, existindo alterações suficientes para que estas variáveis não conduzam à identificação de uma observação.

Tabela 27: Valores da medida UT1

Variáveis	*fundo*	*proj_i40*	*estadofse*
fundo	0.25%	1.29%	18.93%
proj_i40	—	1.03%	18.97%
estadofse	—	—	1.44%

7.5 Métodos para Variáveis Numéricas

Após a perturbação das variáveis categóricas, realiza-se a perturbação das variáveis numéricas sensíveis. Nesta secção aplica-se a maioria dos métodos apresentados no Capítulo 5 e para cada método é apresentado o gráfico da perda de informação (representada como a diferença em percentagem dos valores próprios das matrizes de covariâncias em percentagem) *versus* o risco máximo de identificação (medida de *record linkage*), para diferentes parâmetros do método em estudo.

7.5.1 Modelos Lineares de Ruído

Modelo Aditivo de Ruído Independente

Na Figura 11 apresenta-se o gráfico para o modelo de Ruído Independente, onde é visível diversos pontos que representam diferentes magnitudes de ruído. Para este método, o ruído é expresso em percentagem e é proporcional ao desvio padrão das variáveis sensíveis originais.

Como era de esperar, com o aumento do ruído, existe maior perda de informação e menor risco de identificação. Os valores para a perda de informação são relativamente reduzidos, sendo o valor mais elevado inferior a 50%. Como o foco nesta perturbação continua a ser a utilidade dos dados, fixa-se o limite para o risco de identificação de 0.5, tentando-se assim obter menor perda de informação. Para este modelo, pode-se observar pelo gráfico, que há quatro valores muito próximos, com risco de identificação inferior a 0.5, optando-se por um ruído de 5% que garante um risco de identificação inferior a 0.2 e uma perda de informação de apenas 5%. Para este valor em particular, o desvio padrão do ruído é 0.05 vezes o desvio padrão das variáveis sensíveis originais.

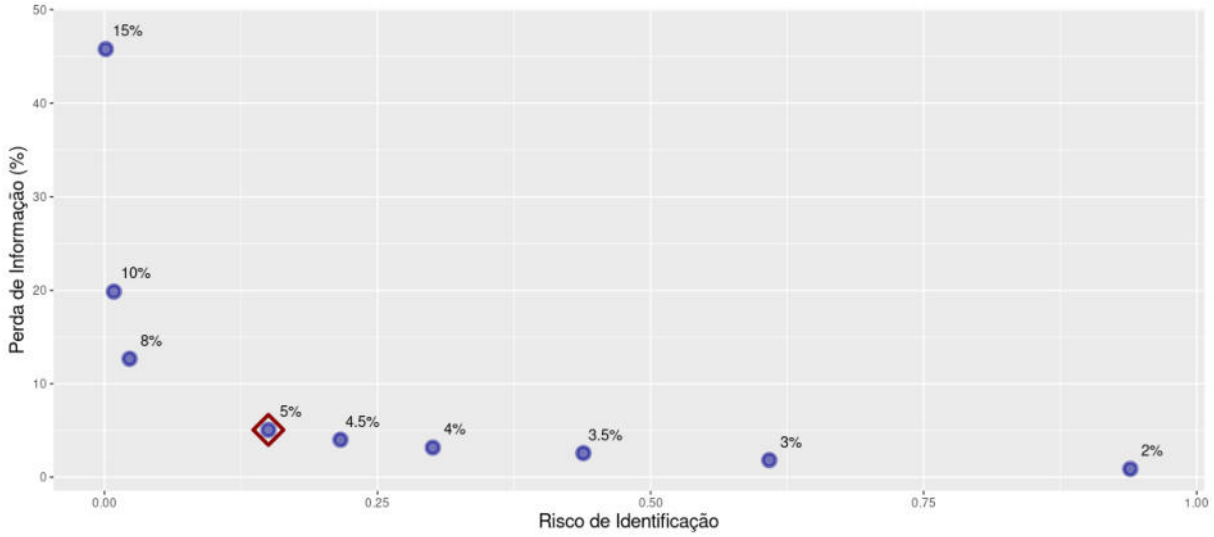


Figura 11: Perda de informação *vs* Risco de Identificação (Modelo de Ruído Independente)

Nos Códigos 38 está representado o código para a aplicação deste modelo, com ruído de 5%. Obtém-se os

valores da perda de informação de IL1= 2.62 x 10^9 e da diferença de valores próprios= 4.8% e do risco de identificação que não ultrapassa os 15%, $[0\%; 14.6\%]$.

```
> sdc_ind <- addNoise(sdcf, noise= 5, method = "additive" )
> print(sdc_ind, "numrisk")
Numerical key variables: dec_investaprov, dec_investeleg,
dec_incentivoaprov, investcand, totpagam_realizado

Disclosure risk is currently between [0.00%; 14.60%]

Current Information Loss:
  - IL1: 2620000000
  - Difference of Eigenvalues: 4.800%
-------------------------------------------------------------------------
```

Código 38: Aplicação do Modelo de Ruído Independente

Modelo Aditivo de Ruído Correlacionado

No caso do modelo de Ruído Correlacionado, a Figura 12 apresenta o gráfico da perda de informação vs o risco de identificação. Como neste modelo as diferentes magnitudes de ruído provocam valores muito próximos entre si (Gráfico da direita), para uma melhor análise apresenta-se um gráfico com limites mais reduzidos para os eixos horizontal e vertical (Gráfico da esquerda). Neste caso, os valores apresentados para a magnitude de ruído representam o valor c (Capítulo 4, Secção 2.2), ou seja, a matriz de covariâncias do ruído é obtida a partir da matriz de covariâncias das variáveis originais como $\Sigma_\epsilon = c\Sigma$. Pela Figura 12, é visível que o risco de identificação em qualquer um dos ruídos adicionados é muito reduzido (aproximadamente 0 em todas as magnitudes de ruído).

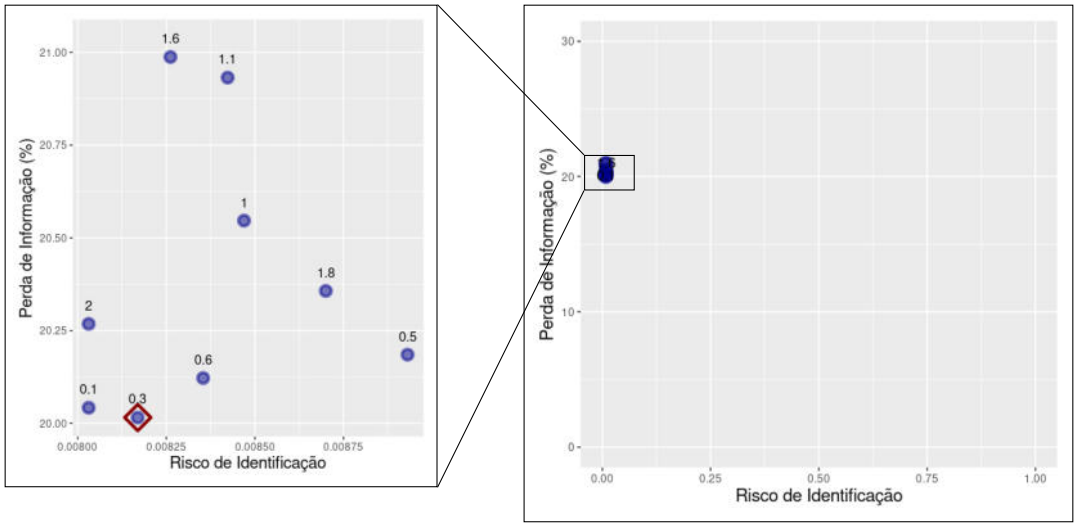


Figura 12: Perda de informação *vs* Risco de Identificação (Modelo de Ruído Correlacionado)

Através dos gráficos apresentados, é possível concluir que independentemente do ruído escolhido, o risco de identificação é reduzido, assim a escolha do nível de ruído depende da perda de informação, e como é visível no pelo gráfico da esquerda, os ruídos contêm valores muito semelhantes de perda de informação, sendo a variação máxima de 1%. Assim, o ruído que apresenta menor diferença de valores próprios é o que tem $c = 0.3$, sendo este o modelo escolhido.

No Código 39 é apresentado a aplicação do modelo e os resultados das medidas de risco de identificação e de perda de informação. Pelo código é também possível ver que o risco de identificação neste caso é muito reduzido,

$[0\%; 0.73\%]$, no entanto, a perda de informação sofre um aumento em comparação com o modelo anterior, mas ainda assim com resultados satisfatórios.

```
> sdc_corr <- addNoise(sdcf, noise= 0.3, method = "correlated2" )
> print(sdc_corr, "numrisk")
Numerical key variables: dec_investaprov, dec_investeleg,
dec_incentivoaprov, investcand, totpagam_realizado

Disclosure risk is currently between [0.00%; 0.73%]

Current Information Loss:
  - IL1: 10253690272.57
  - Difference of Eigenvalues: 20.930%
-----------------------------------------------------------------------
```

Código 39: Aplicação do modelo de Ruído Correlacionado

Modelo EGADP

Em linguagem R, na aplicação deste modelo, não é possível definir a quantidade de ruído que se pretende adicionar, sendo apenas possível aplicar o modelo através da função **shuffle**, como é apresentado no Código 40. Para a aplicação desta função, primeiro é necessário testar o modelo com todas as variáveis chaves (*apoiado, aviso, pofinan, medida, compagamentos e dom_interv*). No Código 40, realiza-se o teste ANOVA para comparar os diferentes modelos possíveis.

```
> anova(mod,mod1)
Analysis of Variance Table
Model 1: dec_investaprov + dec_investeleg + dec_incentivoaprov + investcand
    +  totpagam_realizado ~ apoiado + pofinan + dom_interv
Model 2: dec_investaprov + dec_investeleg + dec_incentivoaprov + investcand
+ totpagam_realizado ~ pofinan + compagamentos + dom_interv + medida +
    apoiado
  Res.Df         RSS Df  Sum of Sq     F    Pr(>F)
1  34288 1.5036e+18
2  34287 1.5032e+18  1 3.9857e+14 9.091 0.002571 **
---
Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1

> anova(mod,mod1)
Analysis of Variance Table

Model 1: dec_investaprov + dec_investeleg + dec_incentivoaprov + investcand
    + totpagam_realizado ~ apoiado + pofinan + aviso + compagamentos
    + dom_interv
Model 2: dec_investaprov + dec_investeleg + dec_incentivoaprov + investcand
    + totpagam_realizado ~ apoiado + aviso + pofinan + medida +
    compagamentos + dom_interv
  Res.Df         RSS    Df   Sum of Sq       F     Pr(>F)
1  34169 9.7460e+17
2  34287 1.5032e+18 118 5.2861e+17 157.06 < 2.2e-16 ***
```

Código 40: Criação dos modelos de regressão

No modelo completo, apenas a variável *compagamentos* aparenta não ser significativa. Assim, no Código 40, inicialmente realiza-se a comparação de dois modelos, o modelo completo (mod1) e o modelo sem a variável *medida* (mod), obtendo-se um p-valor de aproximadamente zero, então rejeita-se a hipótese nula:

$$H_0\text{: } \beta_{compagamentos} = 0 \text{ vs } H_1\text{: } \beta_{compagamentos} \neq 0$$

concluindo-se que é preferível o modelo com a variável *compagamentos*. Realiza-se esta comparação para os diversos modelos possíveis, obtendo-se o modelo final apresentado no Código 41 (mod1), ou seja, o modelo com as variáveis chave *compagamentos*, *pofinan*, *apoiado dom_interv* e *medida*, e assim este modelo é o modelo escolhido.

Após definir o modelo de regressão é possível aplicar o modelo EGADP. Neste caso, aplica-se o modelo à base de microdados, como é demonstrado no Código 41, onde se apresentam ainda os valores das três medidas de perda de informação e da medida de risco de identificação. Com os resultados obtidos é possível concluir que este modelo provoca um risco de identificação nulo, no entanto provoca elevada perda de informação quando comparado com os resultados dos modelos anteriores. Assim, a diferença dos valores próprios, em percentagem, é 355.23%, ou seja, um valor bastante mais elevado do que seria desejável.

```
> originais <- df[,sens]
> mod <- formula(mod1)
> egadp <- shuffle(df, gadp = TRUE, mod)
> dUtility(originais, egadp$egapd, "IL1")
[1] 9031562107
> dUtility(originais, egadp$egapd, "IL1s")
[1] 53302.99
> dUtility(originais, egadp$egapd, "eigen")
[1] 3.049706
> dRisk(originais, egadp$egapd)
[1] 0
```

Código 41: Cálculo das medidas de utilidade e de risco de identificação do Modelo EGADP

7.5.2 Modelos Não Lineares de Ruído

Modelo de Ruído Multiplicativo

No caso do modelo de Ruído Multiplicativo, multiplica-se as variáveis sensíveis por um ruído com diferentes variâncias. Desta forma, é possível perceber a influência da variância do ruído nos dados perturbados. Na Figura 13 é apresentado o gráfico para este modelo e cada ponto representa a variância de um ruído, como seria de esperar, existe um aumento de perda de informação com o aumento da variância, sendo, a perda de informação resultante bastante elevada, tal como o risco de identificação. Apesar dos resultados obtidos por este modelo não serem satisfatórios (elevada perda de informação e elevado risco de identificação), opta-se pelo ruído com variância de 0.5 por ser o que garante alguma redução no risco de identificação sem aumentar demasiado a perda de informação.

No Código 42 apresenta-se o risco de identificação e a perda de informação do modelo de ruído multiplicativo, quando o ruído apresenta uma variância de 0.5. Neste caso, não se apresenta a aplicação do método, pois a base de microdados perturbada é obtida através da multiplicação das variáveis originais por um ruído aleatório. E como já foi referido, os resultados obtidos para estas medidas não são muito satisfatórios, a diferença de valores próprios é de 345.51% e o risco de identificação entre $[0\%; 58.5\%]$. Este modelo apresenta elevada perda de informação para um risco de identificação ainda bastante elevado, quando comparado com os modelos anteriores.

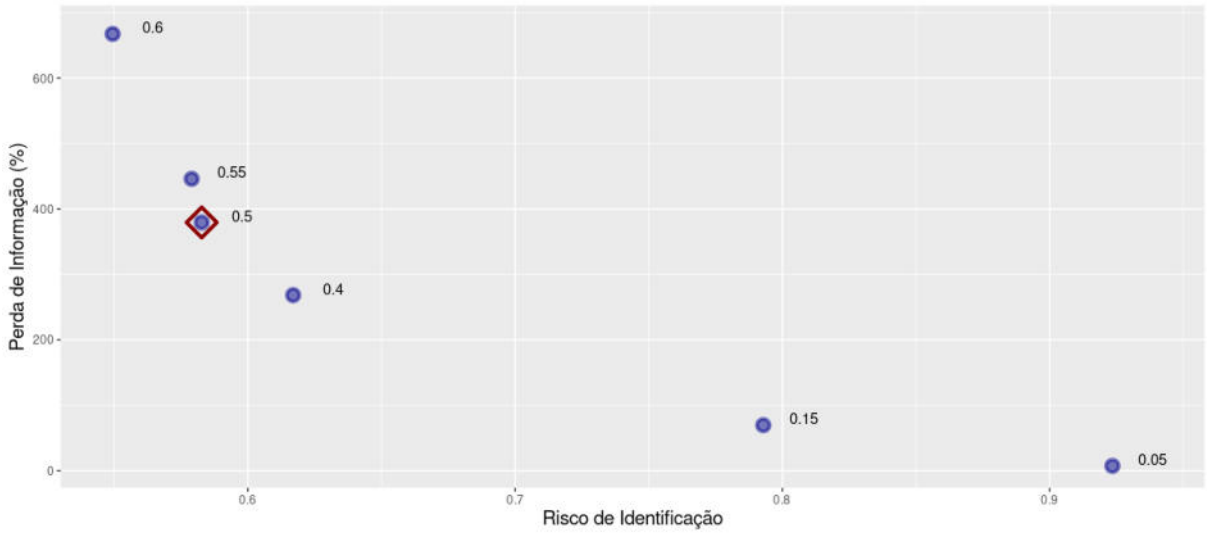


Figura 13: Perda de informação *vs* Risco de Identificação (Modelo de Ruído Multiplicativo)

```
> 100*dUtility(originais,multi, "eigen")
[1] 345.5097
> dUtility(originais,multi, "IL1")
[1] 5123140
> dUtility(originais,multi, "IL1s")
[1] 11487.21
> dRisk(o1, Y)
[1] 0.5854199
```

Código 42: Cálculo das medidas de utilidade e de risco de identificação do modelo de ruído multiplicativo

Data Shuffling

Para o modelo *Data Shuffling* não existem ruídos ou parâmetros a definir, apenas é necessário criar um modelo com base nas variáveis sensíveis e nas variáveis chave. A partir do teste de ANOVA (Código 40) concluiu-se que o modelo apresentado no Código 40 é o modelo a utilizar.

No Código 43 apresenta-se a aplicação do modelo bem como as medidas para a perda de informação e para o risco de identificação. Como é visível, este modelo provoca um risco de identificação reduzido, não ultrapassando os 19.22%, no entanto, a perda de informação apresenta um valor mais elevado do que seria desejável.

```
> mod2 <- lm(dec_investaprov + dec_investeleg + dec_incentivoaprov + investcand +
totpagam_realizado ~ apoiado + medida +pofinan + compagamentos + dom_interv,data=df)

> sdc_shuf <- shuffle(sdcf, gadp = TRUE, mod2)
> print(sdc_shuf, "numrisk")
Numerical key variables: dec_investaprov, dec_investeleg,
dec_incentivoaprov, investcand, totpagam_realizado
Disclosure risk: modified data: [0.00%; 18.22%]
Current Information Loss in modified data (0.00% in original data):
  IL1: 920177082
  Difference of Eigenvalues: 770.927%
```

Código 43: Aplicação do modelo *Data Shuffling*

7.5.3 Outros Métodos Perturbativos

Microagregação pela distância de Mahalanobis

Na Figura 14 está apresentado o gráfico da perda de informação *versus* o risco de identificação para este método. Neste caso, são apresentados diversos pontos correspondentes aos diferentes valores de k, ou seja, a quantas observações se pretende agrupar para o cálculo do valor médio. Como é de se esperar, quanto maior o número de observações num grupo, mais heterogéneos se tornam os grupos e portanto, maior será a perda de informação e menor o risco de identificação. Neste caso, os valores de perda de informação são relativamente elevados (não sendo dos métodos com maior perda de informação), já o risco de identificação apresenta valores elevados, e como tal, é necessário escolher $k = 1000$ de forma a obter um risco de identificação não muito superior a 0.5 e ao mesmo tempo uma perda de informação não muito elevada.

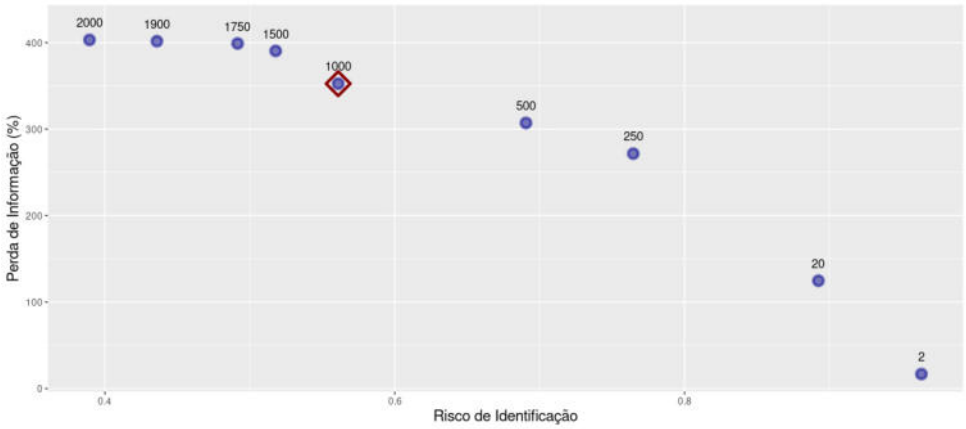


Figura 14: Perda de informação *vs* Risco de Identificação (Microagregação pela distância de Mahalanobis)

No Código 44 é apresentada a aplicação do método bem como os resultados da perda de informação e do risco de identificação. O risco máximo de identificação é aproximadamente 56.10% e a diferença de valores próprios é cerca de 352.56%, que são ainda valores relativamente elevados.

```
> sdc_micro_maha <- microaggregation(sdcf, aggr = 1000, method = "rmd")
> print(sdc_micro_maha, "numrisk")
Numerical key variables: dec_investaprov, dec_investeleg,
dec_incentivoaprov, investcand, totpagam_realizado

Disclosure risk is currently between [0.00%; 56.10%]

Current Information Loss:
  - IL1: 109777281.03
  - Difference of Eigenvalues: 352.560%
-------------------------------------------------------------------------
```

Código 44: Aplicação de Microagregação pela distância de Mahalonobis

Microagregação por Máxima Distância ao Valor Médio

Para este método estão apresentados, na Figura 15, os resultados obtidos para o risco de identificação e para a perda de informação com diferentes níveis de agrupamento. Os valores de perda de informação e do risco de identificação são muito próximos do método anterior e opta-se novamente por $k = 1000$. Apesar de este

modelo apresentar um risco de identificação superior a 0.6, apresenta uma vantagem em relação ao modelo com $k = 2000$, já que existe uma diferença relativamente elevada na perda de informação entre os dois modelos, cerca de 50%.

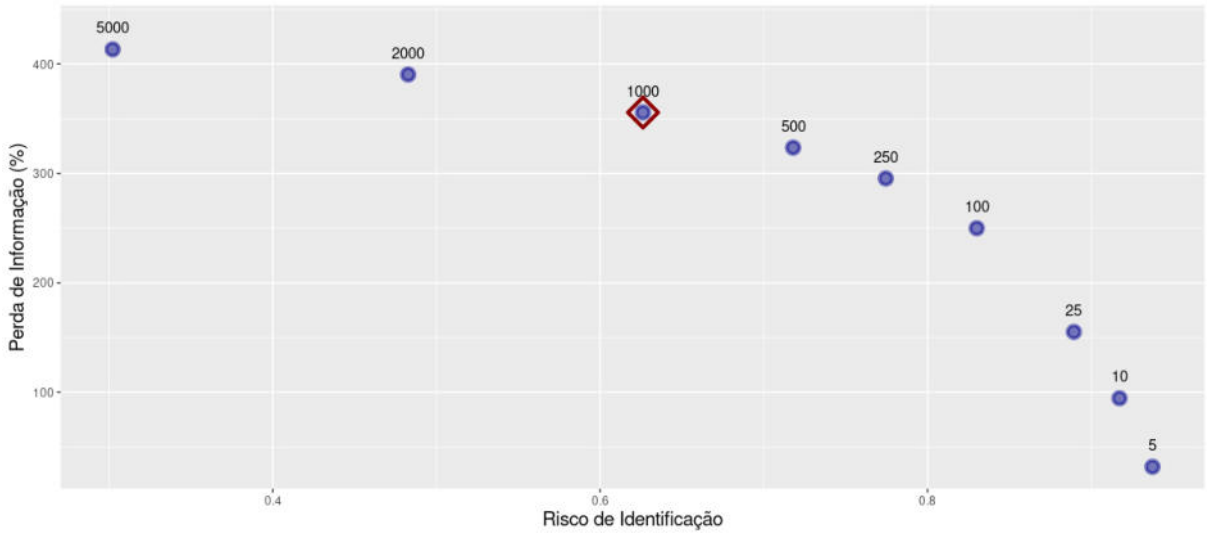


Figura 15: Perda de informação *vs* Risco de Identificação (Microagregação por Máxima Distância ao Valor Médio)

No Código 45 é visível a aplicação do método, com $k = 1000$, com o qual se obtém um risco de identificação superior ao do método anterior, $[0\%; 62.62\%]$. Os resultados obtidos para a perda de informação não são elevados.

```
> sdc_micro_mdav <- microaggregation(sdcf, aggr = 1000, method = "mdav")

> print(sdc_micro_mdav, "numrisk")
Numerical key variables: dec_investaprov, dec_investeleg,
dec_incentivoaprov, investcand, totpagam_realizado

Disclosure risk is currently between [0.00%; 62.62%]

Current Information Loss:
  - IL1: 108596716.12
  - Difference of Eigenvalues: 355.720%
-------------------------------------------------------------------------
```

Código 45: Aplicação de Microagregação por Máxima Distância ao Valor Médio

Microagregação com base em Análise de Componentes Principais

Na Figura 16 é apresentado o gráfico do risco de identificação e de perda de informação para este método, com os diversos valores de k, que representa o número de observações existentes em cada grupo. Pelo gráfico, é possível afirmar que para valores baixos de k, existe um nível de risco de identificação mais reduzido do que em outros métodos de microagregação. Assim, escolhe-se $k = 5$, sendo este valor o que garante um risco de identificação mais próximo de 0.5, obtendo-se assim a menor perda de informação para o risco máximo de identificação desejável.

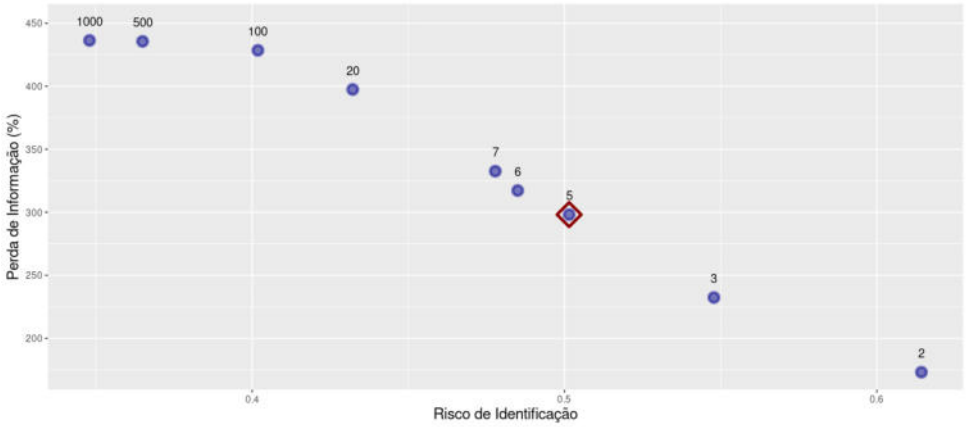


Figura 16: Perda de informação *vs* Risco de Identificação (Microagregação com base em Análise de Componentes Principais)

No Código 46 demonstra-se como aplicar este método e apresentam-se as medidas de perda de informação e de risco de identificação. Como já foi referido, neste método o valor de k é bastante inferior, relativamente aos outros métodos de microagregação, provocando menor perda de informação ($IL1 = 7.26 \text{ x} 10^6$, Diferença de valores próprios= 298.150%) e menor risco de identificação $[0\%; 50.15\%]$.

```
> sdc_micro_pc <- microaggregation(sdcf, aggr = 5, method = "pca")
> print(sdc_micro_pc, "numrisk")
Disclosure risk is currently between [0.00%; 50.15%]
Current Information Loss:
  - IL1: 7275559.10
  - Difference of Eigenvalues: 298.150%
```

Código 46: Aplicação de Microagregação com base em componentes principais

Microagregação pelo método do Ranking Individual

Na Figura 17 está apresentado o gráfico de perda de informação *versus* risco de identificação, e é possível concluir que este método não apresenta um claro padrão de aumento ou diminuição das medidas de CDE para os diferentes valores de k, que representa o número de observações em cada grupo. No entanto, para qualquer valor de k este método apresenta valores relativamente baixos de perda de informação, mas valores elevados para o risco de identificação, optando-se por $k = 4000$. Assim, obtém-se um risco de identificação aceitável e uma perda de informação reduzida.

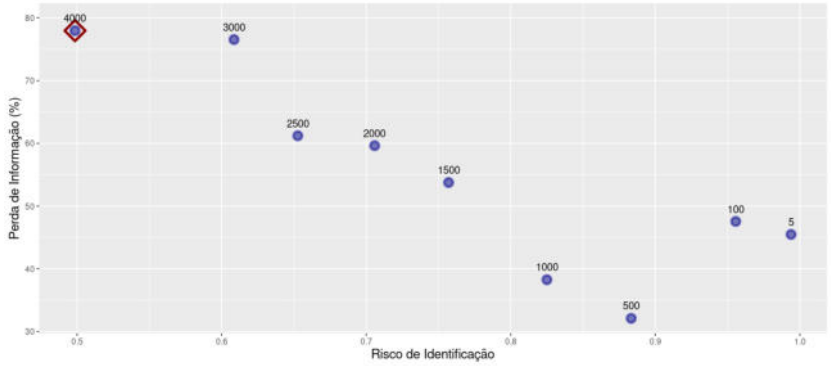


Figura 17: Perda de informação *vs* Risco de Identificação (Microagregação pelo método do Ranking Individual)

No Código 47 é possível ver a aplicação do método e os resultados das medidas de perda de informação e de risco de identificação. Como é percetível, existe um elevado risco de identificação para uma baixa perda de informação ($IL1 = 237921362.39$ e a diferença dos valores próprios é cerca de 77.96%, o risco de identificação é de $[0\%; 49.85\%]$).

```
> sdc_micro_ind  <- microaggregation(sdcf, aggr = 4000, method = "onedims")
> print(sdc_micro_ind, "numrisk")
Numerical key variables: dec_investaprov, dec_investeleg,
dec_incentivoaprov, investcand, totpagam_realizado

Disclosure risk is currently between [0.00%; 49.85%]

Current Information Loss:
  - IL1: 237921362.39
  - Difference of Eigenvalues: 77.960%
-----------------------------------------------------------------------
```

Código 47: Aplicação de Microagregação pelo método de *ranking* individual

Rank Swapping

O método *Rank Swapping* é aplicado usando diferentes valores para o parâmetro P (Capítulo 5), ou seja, a distância máxima entre duas observações elegíveis para troca, definida em percentagem. Na Figura 18 está apresentado o gráfico com os diferentes valores de P. Neste caso, o valor que apresenta valores desejáveis para o risco de identificação é o de $P = 10\%$, ou seja, 10% de distância máxima entre duas observações elegíveis para troca. Por outro lado, este método apresenta elevada perda de informação independentemente do valor de P.

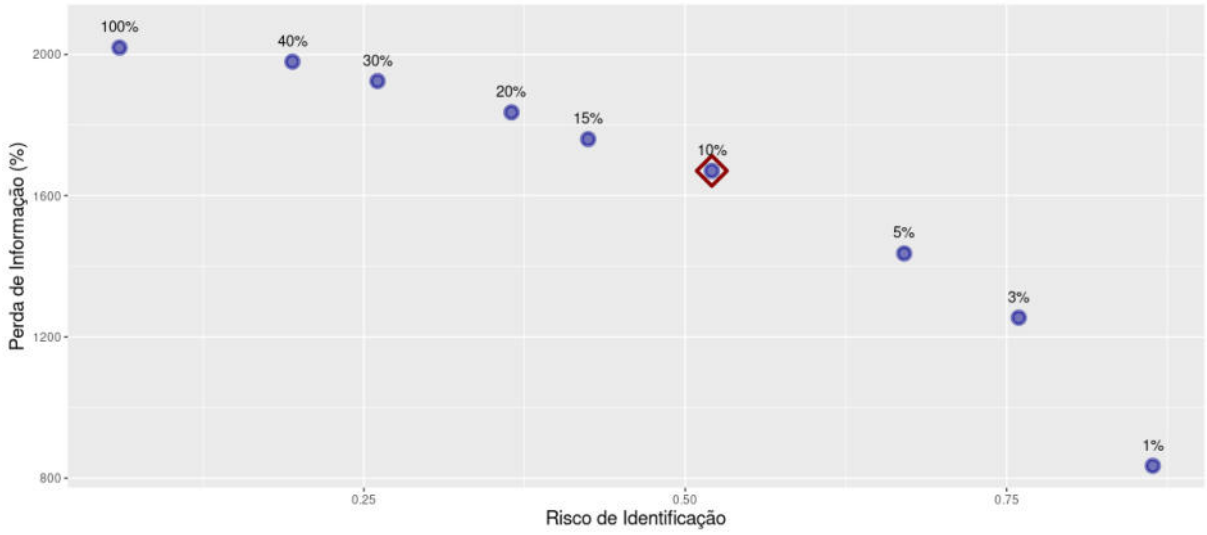


Figura 18: Perda de informação *vs* Risco de Identificação (Método de *Rank Swapping*)

Nos Códigos 48 e 49 apresentam-se os resultados e a aplicação deste método. Como é visível pelos resultados, este método provoca um nível razoável de risco de identificação $[0\%; 51.94\%]$, no entanto, apresenta elevada perda de informação (1678.070%), o que não é desejável.

```
> sdc_rank_sw <- rankSwap(sdcf, TopPercent = 0, BottomPercent = 0, P=0.1)
```

Código 48: Aplicação do método *Rank Swapping*

```
> print(sdc_rank_sw, "numrisk")
Numerical key variables: dec_investaprov, dec_investeleg,
dec_incentivoaprov, investcand, totpagam_realizado

Disclosure risk is currently between [0.00%; 51.94%]

Current Information Loss:
  - IL1: 244131934.55
  - Difference of Eigenvalues: 1678.070%
-----------------------------------------------------------------------
```

Código 49: Aplicação do método *Rank Swapping*

Mapeamento Inverso

Por fim, aplicam-se o método de mapeamento inverso (M.I.) em alguns dos modelos aplicados anteriormente. Como já foi referido no Capítulo 4, é aconselhável a divulgação de uma base de microdados perturbada após o mapeamento inverso. E como tal realiza-se a aplicação deste método de forma a perceber as diferenças existentes face às bases de microdados antes da aplicação deste método. Os resultados estão apresentados na Tabela 28, juntamente com todos os resultados dos métodos aplicados.

Na aplicação deste modelo verifica-se que, independentemente do método aplicado anteriormente, as distribuições marginais de cada variável são iguais às originais, isto é, o valor médio, os quartis e a mediana são exatamente iguais. Como era de esperar, os valores das medidas de CDE dependem do método aplicado anteriormente. Em certos casos observam-se aumentos elevados e noutros casos existe uma diminuição acentuada da perda de informação e do risco de identificação. Nem todos os métodos apresentam valores aceitáveis com a aplicação do mapeamento inverso. Uma análise mais detalhada e comparativa dos resultados, é realizada na próxima secção.

7.5.4 Avaliação dos Métodos

Para concretizar a perturbação das variáveis sensíveis é importante avaliar os diferentes métodos e escolher o que se considera melhor em termos de risco de identificação e de perda de informação. Esta escolha é realizada através de diversas análises e comparações entre as variáveis originais e as variáveis perturbadas. No Capítulo 4, é apresentada uma comparação teórica destes métodos, mas como é referido, os resultados dependem da base de microdados em estudo e das variáveis escolhidas para perturbação. De seguida, comparam-se os diferentes métodos quando aplicados à base de microdados em estudo.

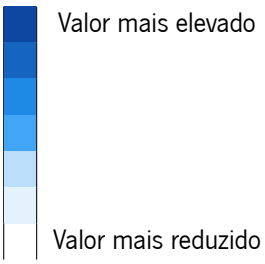
Na Tabela 28 apresentam-se os resultados para o risco máximo de identificação (pela Medida Intervalar) e as medidas de perda de informação (IL1, IL1s, diferença de valores próprios, EAM e EQM). Esta tabela apresenta um padrão de cores para facilitar a comparação entre as ordens de grandeza das medidas nos diferentes métodos.

É possível afirmar que os modelos Aditivos de Ruído Independente e de Ruído Correlacionado são os métodos de CDE que apresentam os valores mais equilibrados (elevada utilidade dos dados e reduzido risco de identificação). O modelo Aditivo de Ruído Independente apresenta os valores mais baixos em sete das nove medidas de perda de informação apresentadas, e o modelo Aditivo de Ruído Correlacionado apresenta valores muito próximos aos das medidas do modelo de Ruído Independente. No caso do modelo Aditivo de Ruído Correlacionado, o risco de identificação é muito próximo de 0%, enquanto que o modelo Aditivo de Ruído Independente apresenta um risco de identificação de 14.60%. Para além destes dois modelos, o modelo de Ruído Multiplicativo e a Microagregação em

Componentes Principais apresentam valores relativamente reduzidos nas medidas de perda de informação, o que provoca aumentos no risco de identificação, cerca de 58.54% e 50.15% respetivamente. Em geral, a diminuição do risco de identificação faz aumentar a perda de informação, sendo por isso aconselhável analisar as diferenças entre as variáveis originais e perturbadas.

Tabela 28: Medidas de perda de informação e risco de identificação

Medidas/ Métodos	M.I.	IL1	IL1s	Valores Próprios	EQM	EQM Σ	EQM ρ	EAM	EAM Σ	EAM ρ	Risco (Med. Int.)
Modelo Aditivo de Ruído Independente	NÃO	2.62 x 10^{9}	6.12 x 10^{3}	0.048	1.29 x 10^{10}	8.06 x 10^{15}	2.28 x 10^{-10}	7.34 x 10^{4}	3.46 x 10^{5}	1.00 x 10^{-7}	14.60%
	SIM	3.70 x 10^{8}	3.0 x 10^{4}	16.543	8.75 x 10^{12}	2.68 x 10^{20}	2.76 x 10^{-5}	3.34 x 10^{7}	1.21 x 10^{8}	4.81 x 10^{-5}	21.50 %
Modelo Aditivo de Ruído Correlacionado	NÃO	1.03 x 10^{10}	1.22 x 10^{4}	0.209	5.15 x 10^{10}	4.20 x 10^{16}	5.67 x 10^{-9}	1.47 x 10^{5}	6.12 x 10^{6}	6.00 x 10^{-7}	0.73%
	SIM	1.93 x 10^{7}	1.00 x 10^{5}	14.410	8.87 x 10^{12}	5.36 x 10^{20}	2.65 x 10^{-5}	4.89 x 10^{5}	1.83 x 10^{8}	4.69 x 10^{-5}	6.30%
Modelo EGADP	NÃO	9.03 x 10^{9}	5.33 x 10^{4}	3.050	4.56 x 10^{12}	1.53 x 10^{21}	9.63 x 10^{-6}	4.56 x 10^{5}	2.17 x 10^{8}	2.33 x 10^{-5}	0.00 %
	SIM	1.40 x 10^{10}	5.20 x 10^{4}	10.644	1.03 x 10^{13}	1.61 x 10^{20}	1.17 x 10^{-5}	6.01 x 10^{5}	8.72 x 10^{7}	3.04 x 10^{-5}	11.90 %
Modelo de Ruído Multiplicativo	NÃO	5.12 x 10^{6}	1.15 x 10^{4}	3.455	9.39 x 10^{10}	1.96 x 10^{19}	0.00	1.39 x 10^{5}	2.82 x 10^{7}	0.00	58.54 %
	SIM	2.20 x 10^{8}	1.00 x 10^{5}	30.941	8.88 x 10^{12}	5.20 x 10^{20}	2.50 x 10^{-5}	4.90 x 10^{8}	1.75 x 10^{8}	4.34 x 10^{-5}	7.00 %
Data Shuffling	NÃO	9.20 x 10^{8}	4.34 x 10^{4}	7.709	9.60 x 10^{12}	1.14 x 10^{20}	1.47 x 10^{-5}	5.28 x 10^{5}	8.11 x 10^{7}	3.36 x 10^{-5}	18.22 %
Rank Swapping	NÃO	2.44 x 10^{8}	3.01 x 10^{4}	16.780	8.75 x 10^{12}	2.68 x 10^{20}	2.76 x 10^{-5}	3.34 x 10^{5}	1.22 x 10^{8}	4.81 x 10^{-5}	51.94 %
Microagregação Mahalanobis	NÃO	1.10 x 10^{8}	1.89 x 10^{4}	3.526	3.80 x 10^{12}	1.00 x 10^{21}	0.00	8.32 x 10^{8}	1.87 x 10^{5}	0.00	56.10 %
	SIM	1.10 x 10^{8}	1.00 x 10^{5}	3.526	8.89 x 10^{12}	4.93 x 10^{20}	2.07 x 10^{-5}	4.91 x 10^{5}	1.62 x 10^{8}	3.70 x 10^{-5}	6.60 %
Microagregação MDVM	NÃO	1.09 x 10^{8}	1.53 x 10^{4}	3.557	3.81 x 10^{12}	1.09 x 10^{21}	0.00	1.74 x 10^{5}	1.63 x 10^{8}	0.00	62.62 %
	SIM	1.90 x 10^{7}	1.00 x 10^{4}	16.455	8.91 x 10^{12}	5.03 x 10^{20}	1.48 x 10^{-5}	4.91 x 10^{5}	1.61 x 10^{8}	2.99 x 10^{-5}	6.86 %
Microagregação em Componentes Principais	NÃO	7.28 x 10^{6}	1.34 x 10^{4}	2.982	2.19 x 10^{12}	3.98 x 10^{20}	0.00	1.68 x 10^{5}	6.24 x 10^{7}	0.00	50.15 %
	SIM	1.30 x 10^{9}	5.20 x 10^{4}	1.288	1.03 x 10^{13}	2.11 x 10^{19}	1.40 x 10^{-6}	6.06 x 10^{5}	1.49 x 10^{8}	9.70 x 10^{-6}	21.40 %
Microagregação de *ranking* Individual	NÃO	2.38 x 10^{8}	1.80 x 10^{4}	0.780	4.56 x 10^{12}	1.56 x 10^{21}	2.47 x 10^{-5}	2.72 x 10^{5}	2.34 x 10^{8}	1.17 x 10^{-5}	49.85%
	SIM	9.90 x 10^{8}	5.20 x 10^{4}	10.930	1.04 x 10^{13}	1.12 x 10^{20}	1.12 x 10^{-5}	6.00 x 10^{5}	1.68 x 10^{8}	2.92 x 10^{-5}	24.30 %



Através da Tabela 28 é possível concluir que existe um método que claramente apresenta os menores valores na maioria das medidas, o Modelo Aditivo de Ruído Independente. Existem ainda outros métodos que se destacam com valores reduzidos nas medidas, os Modelos Aditivo de Ruído Correlacionado e de Ruído Multiplicativo. O modelo EGADP, apresenta um risco de identificação igual a zero, no entanto, contém medidas de perda de informação com valores relativamente elevados. A mesma conclusão é retirada do modelo de *Data Shuffling*, que apresenta valores mais elevados de perda de informação comparativamente aos modelos de Ruído. Os métodos de Microagregação preservam as relações entre as variáveis, visto que os valores das medidas de EQM e EAM dos coeficientes de correlação apresentam valor zero, sendo o método de Microagregação em Componentes Principais o que apresenta valores mais reduzidos para estas medidas. Quanto aos métodos de Mapeamento Inverso (M.I.), estes apresentam resultados razoáveis. Em certos casos, existe uma diminuição nos valores das medidas relativamente ao método aplicado anteriormente (por exemplo, no método de Microagregação de Mahalanobis), mas em outros casos, a aplicação do método provoca um aumento (por exemplo, no método de Ruído Multiplicativo). Assim, verifica-se que a aplicação do método de Mapeamento Inverso não oferece vantagem quando aplicado aos métodos de Ruído Multiplicativo, de Microagregação de Mahalonobis e de Ruído Aditivo Correlacionado.

Da análise da Tabela 28, podemos ainda concluir que não existem grandes diferenças nos valores das medidas dos diferentes métodos. No risco de identificação é visível uma maior variação de acordo com o método, sendo o valor mais elevado 62.62%, que se verifica no método de Microagregação por Máxima Distância ao Valor Médio. Neste caso, em particular, a aplicação do mapeamento inverso permite reduzir consideravelmente o risco de identificação.

Nas Tabelas 29, 30, 31, 32 e 33 apresentam-se as estatísticas descritivas principais das variáveis perturbadas, com uma escala de cores para distinguir o grau de afastamento face às estatísticas descritivas das variáceis originais.

A análise destas tabelas é de grande importância pois apesar de um método apresentar valores baixos nas medidas de CDE, pode não manter certas características da base de microdados que são importantes, como por exemplo, a existência de valores sempre positivos ou nulos.

Na Tabela 29 estão apresentadas as estatísticas descritivas principais da variável *dec_investaprov*. O método de Mapeamento Inverso preserva as estatísticas descritivas principais da variável, não preservando a variância. Pela Tabela 29 destacam-se os métodos de *Data Shuffling* e *Rank Swapping*, que preservam todas as estatísticas da variável. Por outro lado, o modelo EGADP apresenta diversos valores diferentes dos originais, preservando apenas a variância e o valor médio. De referir que os modelos Aditivos de Ruído Independente e de Ruído Correlacionado apresentam valores negativos, o que não é desejável para a base de microdados em estudo, por outro lado, existem valores semelhantes para as estatísticas: valor médio, variância, 3º quartil e valor máximo. Assim, embora estes dois modelos apresentem medidas de CDE bastante reduzidas, as variáveis perturbadas apresentam valores negativos, o que pode ser uma desvantagem, visto que as variáveis sensíveis originais representam montantes de valores sempre iguais ou superiores a zero. Quanto aos métodos de microagregação, estes apresentam valores relativamente idênticos na maior parte das estatísticas, com exceção do método de Microagregação em Componentes Principais, onde a variável perturbada apresenta discrepâncias relativamente elevadas na maioria das estatísticas da variável. Por outro lado, o modelo de Ruído Multiplicativo apresenta valores relativamente diferentes da variável original. A Tabela 29 mostra que a variância desta variável não é preservada por muitos métodos, no entanto, o valor médio e o valor mínimo são preservados na maioria dos casos.

Os resultados da Tabela 30 referem-se às estatísticas descritivas principais da variável *dec_investeleg*. Novamente, o método Mapeamento Inverso preserva todas as estatísticas, com exceção da variância, enquanto que os métodos *Data Shuffling* e *Rank Swapping* preservam todas as estatísticas. O modelo EGADP apresenta

valores relativamente diferentes dos valores originais, preservando apenas o valor médio e a variância. Verifica-se novamente que o modelo Aditivo de Ruído Independente e o modelo Aditivo de Ruído Correlacionado apresentam valores negativos e valores relativamente próximos dos originais para a maioria das estatísticas. Neste caso, o modelo de Ruído Multiplicativo apresenta valores próximos dos originais para a maioria das estatísticas, enquanto que os métodos de microagregação apresentam resultados mais distantes em comparação com a variável anterior, existindo grandes discrepâncias no valor máximo e na variância em todos os métodos de microagregação. A Tabela 30 mostra que o valor máximo e a variância são as estatísticas preservadas por menos métodos, enquanto que o valor mínimo, o valor médio e a mediana são os valores preservados pela maioria dos métodos.

Tabela 29: Estatísticas principais da variável *dec_investaprov*

Medidas/ Métodos	M.I.	Valor Mínimo	1° Quartil	Mediana	Valor Médio	3° Quartil	Valor Máximo	Variância
Originais	–	0	2878	69813	541879	419794	120476104	4.754 x 10^{12}
Modelo Aditivo de Ruído Independente	NÃO	−430744	1548	133902	542600	436896	120457922	1.153 x 10^{12}
	SIM	0	2878	69813	541879	419794	120476104	1.142 x 10^{11}
Modelo Aditivo de Ruído Correlacionado	NÃO	−892316	−26897	178184	540110	488595	120504449	4.754 x 10^{12}
	SIM	0	2878	69813	541879	419794	120476104	1.142 x 10^{11}
Modelo EGADP	NÃO	−527508	10435	362284	541879	853137	9050457	1.677 x 10^{12}
	SIM	0	2878	69813	541879	419794	120476104	1.142 x 10^{11}
Modelo de Ruído Multiplicativo	NÃO	0	0	0	150579	50950	70601784	6.842 x 10^{11}
	SIM	0	2878	69813	541879	419794	120476104	1.142 x 10^{11}
Data Shuffling	NÃO	0	2878	69813	541879	419794	120476104	4.754 x 10^{12}
Rank Swapping	NÃO	0	2878	69813	541879	419794	120476104	4.754 x 10^{12}
Microagregação Mahalanobis	NÃO	0	7089	77830	541879	468436	8395101	1.855 x 10^{12}
	SIM	0	2878	69813	541879	419794	120476104	1.142 x 10^{11}
Microagregação MDVM	NÃO	0	4445	100652	541879	424498	8808226	1.98 x 10^{12}
	SIM	0	2878	69813	541879	419794	120476104	1.142 x 10^{11}
Microagregação em Componentes Principais	NÃO	0	11518	105205	541879	458915	68773677	3.830 x 10^{12}
	SIM	0	2878	69813	541879	419794	120476104	1.142 x 10^{11}
Microagregação de *ranking* Individual	NÃO	0	4627	54201	541879	363838	4594127	1.447 x 10^{12}
	SIM	0	2878	69813	541879	419794	120476104	1.142 x 10^{11}

Valor igual
Valor da mesma ordem de grandeza
Valor de ordem de grandeza diferente

Tabela 30: Estatísticas principais da variável *dec_investeleg*

Medidas/ Métodos	M.I.	Valor Mínimo	1° Quartil	Mediana	Valor Médio	3° Quartil	Valor Máximo	Variância
Originais	–	0	0	10000	307151	160720	120476104	2.829 x 10^{12}
Modelo Aditivo de Ruído Independente	NÃO	−328086	−25357	50543	307418	182771	120536574	2.839 x 10^{12}
	SIM	0	0	10000	307151	160720	120476104	8.807 x 10^{10}
Modelo Aditivo de Ruído Correlacionado	NÃO	−715614	−62770	78149	305908	256604	120015326	2.831 x 10^{12}
	SIM	0	0	10000	307151	160720	120476104	8.807 x 10^{10}
Modelo EGADP	NÃO	−371106	−92848	241081	307151	549420	7823249	7.669 x 10^{11}
	SIM	0	0	10000	307151	160720	120476104	8.807 x 10^{10}
Modelo de Ruído Multiplicativo	NÃO	0	1001	41129	545954	386775	133839253	5.586 x 10^{12}
	SIM	0	0	10000	307151	160720	120476104	8.807 x 10^{10}
Data Shuffling	NÃO	0	0	10000	307151	160720	120476104	2.829 x 10^{12}
Rank Swapping	NÃO	0	0	10000	307151	160720	120476104	2.829 x 10^{12}
Microagregação Mahalanobis	NÃO	0	219	9803	307151	190592	4605576	6.845 x 10^{11}
	SIM	0	0	10000	307151	160720	120476104	8.807 x 10^{10}
Microagregação MDVM	NÃO	0	190	19339	307151	198421	6879746	1.111 x 10^{12}
	SIM	0	0	10000	307151	160720	120476104	8.807 x 10^{10}
Microagregação em Componentes Principais	NÃO	0	4500	20000	307151	166086	62524456	2.504 x 10^{12}
	SIM	0	0	10000	307151	160720	120476104	8.807 x 10^{10}
Microagregação de *ranking* Individual	NÃO	0	0	8266	307151	113893	3029292	6.391 x 10^{11}
	SIM	0	0	10000	307151	160720	120476104	8.807 x 10^{10}

Valor igual
Valor da mesma ordem de grandeza
Valor de ordem de grandeza diferente

Tabela 31: Estatísticas principais da variável *dec_incentivoapov*

Medidas/ Métodos	M.I.	Valor Mínimo	1° Quartil	Mediana	Valor Médio	3° Quartil	Valor Máximo	Variância
Originais	–	0	0	0	149843	68217	42166636	4.950 x 10^{11}
Modelo Aditivo de Ruído Independente	NÃO	−152460	−11176	20463	149945	77781	42174724	4.962 x 10^{11}
	SIM	0	0	0	149843	68217	42166636	4.950 x 10^{11}
Modelo Aditivo de Ruído Correlacionado	NÃO	−281007	−26008	32564	149876	112049	41983856	4.952 x 10^{11}
	SIM	0	0	0	149843	68217	42166636	4.950 x 10^{11}
Modelo EGADP	NÃO	−142919	−50886	107420	149843	371184	2176334	2.398 x 10^{11}
	SIM	0	0	0	149843	68217	42166636	4.950 x 10^{11}
Modelo de Ruído Multiplicativo	NÃO	0	0	6307	309519	126394	172583282	3.479 x10^{12}
	SIM	0	0	0	149843	68217	42166636	4.950 x 10^{11}
Data Shuffling	NÃO	0	0	0	149843	68217	42166636	4.950 x 10^{11}
Rank Swapping	NÃO	0	0	0	149843	68217	42166636	4.950 x 10^{11}
Microagregação Mahalanobis	NÃO	0	14.8	3839.4	149843	83081.5	1914965	1.551 x 10^{11}
	SIM	0	0	0	149843	68217	42166636	4.950 x 10^{11}
Microagregação MDVM	NÃO	0	32	2245	149843	95563	3250884	2.557 x 10^{13}
	SIM	0	0	0	149843	68217	42166636	4.950 x 10^{11}
Microagregação em Componentes Principais	NÃO	0	0	8981	149843	75358	19093011	4.373 x 10^{11}
	SIM	0	0	0	149843	68217	42166636	4.950 x 10^{11}
Microagregação de *ranking* Individual	NÃO	0	0	0	149843	44524	1493263	1.559 x 10^{11}
	SIM	0	0	0	149843	68217	42166636	4.950 x 10^{11}

Valor igual
Valor da mesma ordem de grandeza
Valor de ordem de grandeza diferente

As estatísticas descritivas principais da variável *dec_incentivoaprov* são apresentadas na Tabela 31. Neste caso, o Mapeamento Inverso preserva todas as estatísticas, tal como os métodos de *Data Shuffling* e *Rank Swapping*. O modelo de Ruído Multiplicativo apresenta resultados relativamente diferentes dos originais, preservando apenas o valor médio. Os modelos Aditivos de Ruído Correlacionado e de Ruído Independente

apresentam novamente valores negativos, possuindo valores próximos dos originais nas estatísticas do valor médio, 3° quartil, valor máximo e variância. Quanto ao modelo EGADP, este apresenta grandes diferenças relativamente à variável original, preservando apenas o valor médio e a variância. Os métodos de microagregação preservam o valor médio, como era de esperar, e neste caso a Microagregação em Componentes Principais é o método que apresenta os resultados mais desejáveis. Pela Tabela 31 conclui-se que para esta variável a variância e o valor médio são preservados pela maioria dos métodos, já o valor máximo e a mediana são os valores preservados por menos métodos.

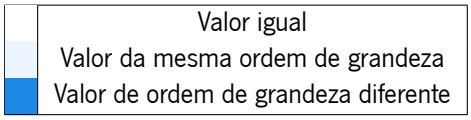
Tabela 32: Estatísticas principais da variável *investcand*

Medidas/ Métodos	M.I.	Valor Mínimo	1° Quartil	Mediana	Valor Médio	3° Quartil	Valor Máximo	Variância
Originais	–	0	20000	178400	723740	545771	480000000	1.731 x 10^{13}
Modelo Aditivo de Ruído Independente	NÃO	−827568	18965	242268	723949	604514	479737990	1.735 x 10^{13}
	SIM	0	20000	178400	723740	545771	480000000	1.731 x 10^{13}
Modelo Aditivo de Ruído Correlacionado	NÃO	−281007	−26008	32564	149876	112531	42124366	1.730 x 10^{13}
	SIM	0	20000	178400	723740	545771	480000000	1.731 x 10^{13}
Modelo EGADP	NÃO	−419460	−163	402070	723740	1121439	9694611	4.832 x 10^{12}
	SIM	0	20000	178400	723740	545771	480000000	1.731 x 10^{13}
Modelo de Ruído Multiplicativo	NÃO	0	19991	136972	722777	517473	526439203	1.888 x 10^{13}
	SIM	0	20000	178400	723740	545771	480000000	1.731 x 10^{13}
Data Shuffling	NÃO	0	20000	178400	723740	545771	480000000	1.731 $x10^{13}$
Rank Swapping	NÃO	0	20000	178400	723740	545771	480000000	1.731 x 10^{13}
Microagregação Mahalanobis	NÃO	4889	20286	201265	723740	609733	12758188	3.862 x10^{12}
	SIM	0	20000	178400	723740	545771	480000000	1.731 x 10^{13}
Microagregação MDVM	NÃO	4889	19948	173620	723740	570561	8855362	3.113 x 10^{12}
	SIM	0	20000	178400	723740	545771	480000000	1.731 x 10^{13}
Microagregação em Componentes Principais	NÃO	1	20000	252762	723740	619078	140223467	8.139 x 10^{12}
	SIM	0	20000	178400	723740	545771	480000000	1.731 x 10^{13}
Microagregação de *ranking* Individual	NÃO	5796	20000	169371	723740	483659	5957542	2.400 x 10^{12}
	SIM	0	20000	178400	723740	545771	480000000	1.731 x 10^{13}

Valor igual
Valor da mesma ordem de grandeza
Valor de ordem de grandeza diferente

Tabela 33: Estatísticas principais da variável *totpagam_realizado*

Medidas/ Métodos	M.I.	Valor Mínimo	1° Quartil	Mediana	Valor Médio	3° Quartil	Valor Máximo	Variância
Originais	–	0	0	0	70613	9945	40058305	2.137 x 10^{11}
Modelo Aditivo de Ruído Independente	NÃO	−83725	−10638	7731	70643	29800	40068326	2.142 x 10^{11}
	SIM	0	0	0	70613	9945	40058305	2.137 x 10^{11}
Modelo Aditivo de Ruído Correlacionado	NÃO	−199821	−23451	11667	70437	51677	39905545	2.141 x 10^{11}
	SIM	0	0	0	70613	9945	40058305	2.137 x 10^{11}
Modelo EGADP	NÃO	−102429	−42815	41853	70613	103657	3203419	3.662 x 10^{10}
	SIM	0	0	0	70613	9945	40058305	2.137 x 10^{11}
Modelo de Ruído Multiplicativo	SIM	0	0	0	69183	7287	25619093	2.069 x 10^{11}
	SIM	0	0	0	70613	9945	40058305	2.137 x 10^{11}
Data Shuffling	NÃO	0	0	0	70613	9945	40058305	2.137 x 10^{11}
Rank Swapping	NÃO	0	0	0	70613	9945	40058305	2.137 x 10^{11}
Microagregação Mahalanobis	NÃO	0	0	402.5	70613	14989.7	1081371.3	4.937 x 10^{10}
	SIM	0	0	0	70613	9945	40058305	2.137 x 10^{11}
Microagregação MDVM	NÃO	0	0	519.9	70613	16032.9	1819178.8	7.854 x 10^{10}
	SIM	0	0	0	70613	9945	40058305	2.137 x 10^{11}
Microagregação em Componentes Principais	NÃO	0	0	1929	70613	19847	13173526	1.462 x 10^{11}
	SIM	0	0	0	70613	9945	40058305	2.137 x 10^{11}
Microagregação de *ranking* Individual	NÃO	0	0	0	70613	4845	818219	4.699 x 10^{10}
	SIM	0	0	0	70613	9945	40058305	2.137 x 10^{11}



A Tabela 32 apresenta as estatísticas descritivas principais da variável *investcand*. Novamente, os métodos de Mapeamento Inverso, *Data Shuffling* e *Rank Swapping* preservam as estatísticas principais. O modelo Aditivo de Ruído Independente apresenta valores bastante próximos aos valores originais, apesar da existência de valores negativos. Já o modelo Aditivo de Ruído Correlacionado e o modelo EGADP apresentam valores com diferenças relativamente elevadas face aos originais, onde a maioria das estatísticas apresentam valores diferentes dos originais. Neste caso, o modelo de Ruído Multiplicativo apresenta resultados muito idênticos aos valores originais. Os métodos de microagregação apresentam diferenças elevadas na variância, no valor máximo e no valor mínimo, não preservando as observações com valor zero desta variável. Pela Tabela 32 é visível que grande parte dos métodos preservam o valor médio, a mediana e o 3º quartil da variável original, por outro lado não são preservados o valor mínimo e a variância por muitos dos métodos.

Para terminar, analisa-se a Tabela 33, que apresenta as estatísticas descritivas principais da variável *totpagam_realizado*. Conclui-se que os métodos de Mapeamento Inverso, *Data Shuffling* e *Rank Swapping* preservam todas as estatísticas. Quanto aos modelos de Ruído Correlacionado, Ruído Independente e EGADP, estes apresentam valores com diferenças elevadas comparativamente aos valores originais, preservando apenas o valor médio no caso do modelo EGADP. Já os modelos Aditivos de Ruído Independente e de Ruído Correlacionado apresentam resultados próximos para o valor médio, para o valor máximo e para a variância. Novamente o modelo de Ruído Multiplicativo não apresenta valores muito diferentes dos valores originais. Quanto aos métodos de microagregação, estes apresentam resultados com diferenças relativamente reduzidas, apresentando valores próximos para a maioria das estatísticas. A Tabela 33 mostra que maioria dos métodos preserva o valor mínimo, o 1º quartil e o valor médio, não sendo preservado o valor máximo por grande parte dos métodos.

Tendo em consideração a conclusão obtida da Tabela 28, os métodos com medidas de perda de informação mais reduzidas, eram os modelos de Ruído Independente, de Ruído Correlacionado e de Ruído Multiplicativo. Destes três modelos, apenas o ruído multiplicativo não apresenta valores negativos, mas apresenta estatísticas muito diferentes em algumas variáveis. Quanto aos outros dois modelos, estes possuem uma grande desvantagem que é a existência de valores negativos, por outro lado, grande parte das estatísticas são preservadas para a maioria das variáveis em estudo.

Da análise realizada às tabelas das estatísticas principais, destacam-se os métodos de Mapeamento Inverso, *Data Shuffling* e *Rank Swapping* que preservam todas as estatísticas na maioria das variáveis. Dos métodos apresentados, o que apresenta valores mais distintos nas 5 variáveis em estudo é o modelo EGADP, pois apresenta valores negativos e na maioria das variáveis apenas preserva o valor médio e a variância. Os métodos de Microagregação destacam-se em algumas variáveis com valores muito próximos dos valores originais, em particular a Microagregação em Componentes Principais e de *Ranking* Individual, o que era de esperar, visto que estes métodos apresentavam medidas de perda de informação bastante reduzidas.

Com base nas tabelas apresentadas, destacam-se os seguintes métodos: modelo de Ruído Multiplicativo, modelo Aditivo de Ruído Independente, modelo Aditivo de Ruído Correlacionado, Microagregação em Componentes Principais, Microagregação de *Ranking* Individual, Mapeamento Inverso, *Data Shuffling* e *Rank Swapping*. Por um lado, o método de *Rank Swapping* preserva as estatísticas principais de todas as variáveis, por outro lado, apresenta o maior risco de identificação e em certas medidas de perda de informação apresenta o pior

cenário, considerando por exemplo, as medidas EQM e EAM dos coeficientes de correlação (Tabela 28). Assim,na próxima análise, exclui-se o método de *Rank Swapping* uma vez que este método influencia muito a relação entre as variáveis, traduzindo-se numa elevada perda de informação (Figura 17 e Tabela 28).

A base de microdados em estudo contém diversas variáveis relacionadas entre si e caso não sejam preservadas as observações com valores nulos em determinadas variáveis originais, a perturbação das variáveis sensíveis pode originar resultados sem sentido, como por exemplo, valores negativos nos montantes de financiamento. Assim, um dos principais objetivos desta perturbação é fornecer a maior utilidade possível, ou seja, é necessário que a base de microdados perturbada cumpra determinadas condições. Os modelos Aditivos de Ruído Independente e de Ruído Correlacionado resultam em montantes de financiamento negativos, mas como estes modelos apresentam medidas de perda de informação bastante reduzidas e preservam maioria das estatísticas, decidiu-se realizar a perturbação de forma a que estes métodos preservem as observações com valores iguais a zero e que apenas existam valores positivos. O mesmo se aplica ao modelo EGADP. Para que as bases de microdados perturbadas por estes modelos cumpram as condições referidas, aplica-se o módulo às observações que possuem valores negativos e força-se que observações com valores nulos não sofram perturbação.

De forma a que a perda de informação seja o mais reduzida possível, escolhe-se o método de Mapeamento Inverso após o modelo EGADP, pois é um método que apresenta medidas de perda de informação relativamente reduzidas e permite que os valores nulos sejam preservados. Repetimos o mesmo procedimento no modelo *Data Shuffling*. Esta opção deve-se ao facto dos métodos EGADP e *Data Shuffling* serem teoricamente os métodos com melhor desempenho.

Na Tabela 34 são apresentadas as medidas de perda de informação e risco de identificação para estes métodos em particular. Já nas Figuras 8.13 e 8.14 apresentam-se os gráficos que comparam a perda de informação com o risco de identificação nos diversos métodos aplicados.

Como é percetível pela Tabela 34 e considerando a restrição de valores positivos, o modelo Aditivo de Ruído Independente contém valores superiores em algumas medidas de perda de informação (IL1, EQM e EAM das variâncias). No entanto, quando comparado com os valores iniciais, percebe-se que continuam bastantes reduzidos e que até se verifica uma descida em grande parte das medidas de perda de informação. De referir que neste caso o risco de identificação sofreu um aumento relativamente elevado, apresentando um valor ainda assim inferior a 50% (42.50%). O modelo Aditivo de Ruído Correlacionado apresenta uma diminuição na maioria das medidas de perda de informação e um aumento de risco de identificação para 16.76%, que ainda é um valor perfeitamente aceitável. É possível perceber que as restrições impostas nos dois modelos de ruído provocaram alterações nas variâncias e covariâncias das variáveis sensíveis. Ainda assim, e apesar do aumento, este modelos continuam a ser os que apresentam estas medidas com valores mais reduzidos. O modelo EGADP apresenta pequenas reduções na maioria das medidas, continuando a apresentar resultados bastante razoáveis e um ótimo compromisso entre a perda de informação e o risco de identificação.

Quando se aplica o Mapeamento Inverso no modelo EGADP com a restrição de valores nulos, existe uma diminuição acentuada na medida de perda de informação IL1 e um aumento acentuado no risco de identificação, que se encontra agora nos 23.15%. Quanto ao modelo *Data Shuffling*, este apenas sofreu uma alteração considerável em uma das medidas, a medida IL1 que aumentou. Como é visível pela Tabela 34, as medidas de perda de informação do método *Data Shuffling* e do método de Mapeamento Inverso após EGADP são muito próximas entre si, o que seria de esperar, já que o modelo *Data Shuffling* utiliza o processo de mapeamento inverso após a perturbação do modelo EGADP.

Assim, é possível afirmar que, em geral, a perda de informação diminui pelo facto de impormos a restrição dos

valores serem sempre positivos e dos valores nulos permanecerem nulos, o que seria de esperar. Relativamente ao risco de identificação, os modelos de Ruído apresentam aumentos consideráveis.

Tabela 34: Medidas de perda de informação e risco de identificação

Medidas/ Métodos	Com Restrição	IL1	IL1s	Valores Próprios	EQM	EQM Σ	EQM ρ	EAM	EAM Σ	EAM ρ	Risco (RL)
Modelo Aditivo de Ruído Independente	NÃO	2.62 x 10^{9}	6.12 x 10^{3}	0.048	1.29 x 10^{10}	8.06 x 10^{15}	2.28 x 10^{-10}	7.34 x 10^{4}	3.46 x 10^{5}	1.00 x 10^{-7}	14.60%
	SIM	↑4.52 x 10^{9}	↓3.4 x 10^{3}	↓0.033	↓9.97 x 10^{9}	↑5.46 x 10^{16}	4.29 x 10^{-10}	↓5.25 x 10^{4}	↑1.46 x 10^{6}	3.61 x 10^{-7}	↑42.50 %
Modelo Aditivo de Ruído Correlacionado	NÃO	1.03 x 10^{10}	1.22 x 10^{4}	0.209	5.15 x 10^{10}	4.20 x 10^{16}	3.82 x 10^{-9}	1.47 x 10^{5}	6.12 x 10^{6}	6.00 x 10^{-7}	0.73%
	SIM	↓9.06 x 10^{9}	↓6.70 x 10^{3}	↓0.105	↓3.88 x 10^{10}	↑7.95 x 10^{17}	3.22 x 10^{-9}	↓1.03 x 10^{5}	↓5.79 x 10^{6}	5.00 x 10^{-7}	↑16.76%
Modelo EGADP	NÃO	9.15 x 10^{9}	5.34 x 10^{4}	3.550	4.66 x 10^{12}	1.60 x 10^{21}	9.78 x 10^{-6}	4.68 x 10^{5}	2.28 x 10^{8}	2.63 x 10^{-5}	0.00 %
	SIM	↓6.91 x 10^{9}	↓4.05 x 10^{4}	↓1.235	↓4.12 x 10^{12}	↓1.58 x 10^{21}	↓1.24 x 10^{-6}	↓3.42 x 10^{5}	↑2.12 x 10^{8}	↓8.87 x 10^{-6}	↑2.36 %
Mapeamento Inverso após EGADP	NÃO	1.40 x 10^{10}	5.20 x 10^{4}	10.644	1.03 x 10^{13}	1.61 x 10^{20}	1.17 x 10^{-5}	6.01 x 10^{5}	8.72 x 10^{7}	3.04 x 10^{-5}	11.90 %
	SIM	↓1.55 x 10^{9}	↓2.57 x 10^{4}	↓8.055	↓8.74 x 10^{12}	↑2.36 x 10^{20}	1.27 x 10^{-5}	↓4.14 x 10^{5}	↑1.12 x 10^{8}	3.28 x 10^{-5}	↑23.15 %
Data Shuffling	NÃO	9.33 x 10^{8}	4.54 x 10^{4}	10.804	9.90 x 10^{12}	1.28 x 10^{20}	1.52 x 10^{-5}	5.42 x 10^{5}	8.25 x 10^{7}	3.50 x 10^{-5}	19.22 %
	SIM	↑8.95 x 10^{9}	↓2.6 x 10^{4}	10.804	↓9.02 x 10^{12}	↑3.70 x 10^{20}	↑2.10 x 10^{-5}	↓4.24 x 10^{5}	↑1.32 x 10^{8}	↑7.28 x 10^{-5}	19.93 %

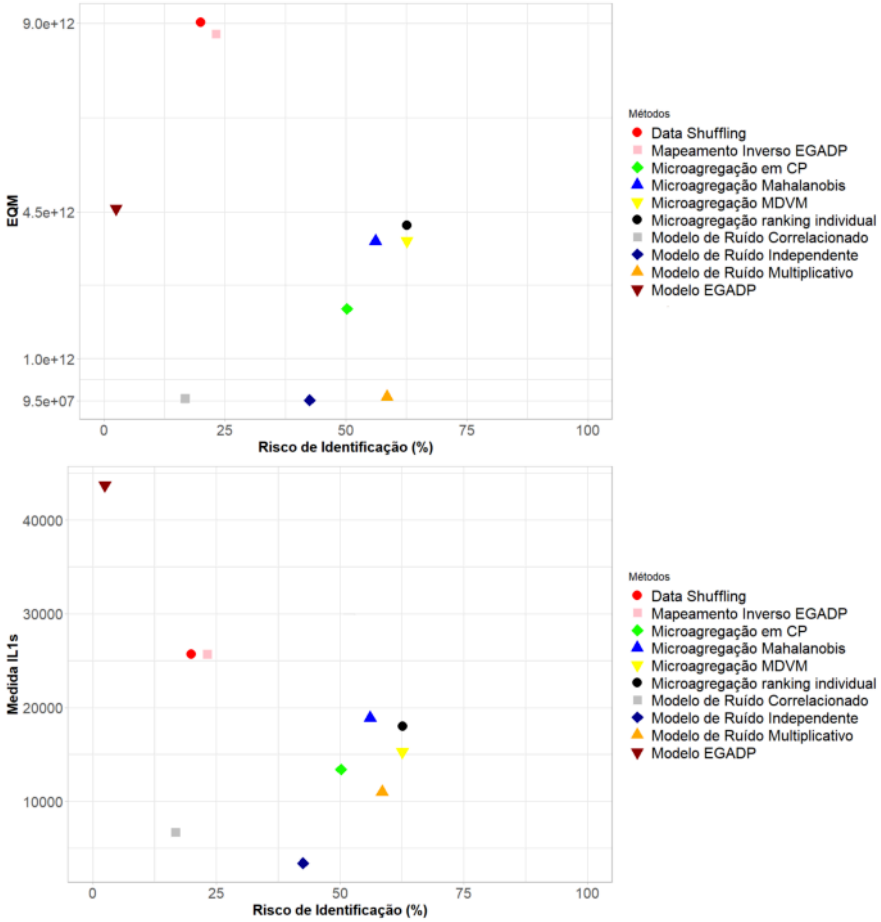


Figura 19: Gráficos de Perda de informação (IL1s e EQM) vs Risco de Identificação

Na Figura 19 representam-se os resultados das medidas de perda de informação EQM e IL1 face ao risco de identificação. É possível observar que o modelo Aditivo de Ruído Independente é o que apresenta o menor valor nas duas medidas de perda de informação (EQM e IL1). Os modelos de Ruído Correlacionado, Ruído Multiplicativo e Microagregação em Componentes Principais são os que, para além do modelo Aditivo de Ruído Independente, apresentam valores também bastantes reduzidos nas medidas de perda de informação em ambos os gráficos. Quanto ao método EGADP, este possui o risco de identificação mais baixo de todos os métodos, no entanto, contém o maior valor para a medida IL1s. Pelos gráficos apresentados, verifica-se que o método de Mapeamento Inverso após EGADP e o modelo *Data Shuffling* estão muito próximos entre si, no entanto, apresentam os valores mais elevados para a medida EQM. Este resultado já seria de esperar uma vez que estes métodos efetuam trocas entre as observações, aumentando as distâncias entre os valores originais e perturbados. Após a análise da tabela e dos gráficos, conclui-se que os modelos de Ruído continuam a ser os que apresentam melhores resultados nas medidas de perda de informação. Em particular, o modelo Aditivo de Ruído Correlacionado apresenta um compromisso ideal entre a perda de informação e o risco de identificação, apresentando em ambos os casos dos valores mais reduzidos comparativamente aos restantes métodos.

Até ao momento percebe-se que não existem alterações consideráveis quando se impõem as restrições de valores nulos e de valores positivos. De forma a perceber melhor as vantagens de se aplicar estas restrições, analisam-se de seguida as estatísticas principais das variáveis sensíveis.

De seguida compara-se as estatísticas principais das variáveis perturbadas pelos novos métodos.

Valor igual
Valor da mesma ordem de grandeza
Valor de ordem de grandeza diferente

Tabela 35: Estatísticas principais da variável *dec_investaprov* para os métodos com restrições

Medidas/ Métodos	Com Restrição	Valor Mínimo	1° Quartil	Mediana	Valor Médio	3° Quartil	Valor Máximo	Variância
Originais	–	0	2878	69813	541879	419794	120476104	4.754×10^{12}
Modelo Aditivo de Ruído Independente	NÃO	−430744	1548	133902	542600	436896	120457922	1.153×10^{12}
	SIM	0	2626	129978	561057	431380	120487224	4.742×10^{12}
Modelo Aditivo de Ruído Correlacionado	NÃO	−892316	−26897	178184	540110	488595	120504449	4.754×10^{12}
	SIM	0	3865	184555	590741	485546	119897167	4.687×10^{12}
Modelo EGADP	NÃO	−528538	10535	373284	541879	963137	6850457	1.587×10^{12}
	SIM	0	7338	178557	468478	802154	6897824	5.138×10^{11}
Mapeamento Inverso após EGADP	NÃO	0	2878	69813	541879	419794	120476104	1.142×10^{11}
	SIM	0	2833	39500	537788	391688	120476104	5.346×10^{12}
Data Shuffling	NÃO	0	2878	69813	541879	419794	120476104	4.754×10^{12}
	SIM	0	2975	64704	535793	413825	120476104	4.602×10^{12}

Tabela 36: Estatísticas principais da variável *dec_investeleg* para os métodos com restrições

Medidas/ Métodos	Com Restrição	Valor Mínimo	1° Quartil	Mediana	Valor Médio	3° Quartil	Valor Máximo	Variância
Originais	–	0	0	10000	307151	160720	120476104	2.829×10^{12}
Modelo Aditivo de Ruído Independente	NÃO	−328086	−25357	50543	307418	182771	120536574	2.839×10^{12}
	SIM	0	0	20049	320796	177543	120447025	2.826×10^{12}
Modelo Aditivo de Ruído Correlacionado	NÃO	−715614	−62770	78149	305908	256604	120015326	2.831×10^{12}
	SIM	0	0	32071	339926	242069	119815592	2.798×10^{12}
Modelo EGADP	NÃO	−471106	−102848	259081	307151	509420	5623249	8.669×10^{11}
	SIM	0	0	76579	293321	477991	5156791	2.673×10^{11}
Mapeamento Inverso após EGADP	NÃO	0	0	10000	307151	160720	120476104	8.807×10^{10}
	SIM	0	0	10000	287827	157602	63605000	1.900×10^{12}
Data Shuffling	NÃO	0	0	10000	307151	160720	120476104	2.829×10^{12}
	SIM	0	0	10000	320429	162324	120476104	2.845×10^{12}

Tabela 37: Estatísticas principais da variável *dec_incentivoapov* para os métodos com restrições

Medidas/ Métodos	Com Restrição	Valor Mínimo	1° Quartil	Mediana	Valor Médio	3° Quartil	Valor Máximo	Variância
Originais	–	0	0	0	149843	68217	42166636	4.950×10^{11}
Modelo Aditivo de Ruído Independente	NÃO	−152460	−11176	20463	149945	77781	42174724	4.962×10^{11}
	SIM	0	0	0	152191	72350	42180552	4.950×10^{11}
Modelo Aditivo de Ruído Correlacionado	NÃO	−281007	−26008	32564	149876	112049	41983856	4.952×10^{11}
	SIM	0	0	0	156289	90513	41983856	4.905×10^{11}
Modelo EGADP	NÃO	−242919	−61886	128420	149843	261184	1976334	1.298×10^{11}
	SIM	0	0	0	132174	214196	1998293	7.44×10^{10}
Mapeamento Inverso após EGADP	NÃO	0	0	0	149843	68217	42166636	4.950×10^{11}
	SIM	0	0	0	164963	63377	42166636	1.840×10^{11}
Data Shuffling	NÃO	0	0	0	149843	68217	42166636	4.950×10^{11}
	SIM	0	0	0	165855	69536	42166636	5.791×10^{11}

Tabela 38: Estatísticas principais da variável *investcand* para os métodos com restrições

Medidas/ Métodos	Com Restrição	Valor Mínimo	1° Quartil	Mediana	Valor Médio	3° Quartil	Valor Máximo	Variância
Originais	–	0	20000	178400	723740	545771	480000000	1.731×10^{13}
Modelo Aditivo de Ruído Independente	NÃO	−827568	18965	242268	723949	604514	479737990	1.735×10^{13}
	SIM	0	120441	276788	794559	610514	480177526	1.724×10^{13}
Modelo Aditivo de Ruído Correlacionado	NÃO	−281007	−26008	32564	149876	112531	42124366	1.730×10^{13}
	SIM	0	193436	420382	900343	788921	477952649	1.702×10^{13}
Modelo EGADP	NÃO	−529460	−293	429070	723740	1241339	7693621	5.812×10^{12}
	SIM	0	154054	379782	792497	1382968	7415986	9.97×10^{11}
Mapeamento Inverso após EGADP	NÃO	0	20000	178400	723740	545771	480000000	1.731×10^{13}
	SIM	0	20000	178400	723740	545771	480000000	1.731×10^{13}
Data Shuffling	NÃO	0	20000	178400	723740	545771	480000000	1.731×10^{13}
	SIM	0	20000	178400	723740	545771	480000000	1.731×10^{13}

Tabela 39: Estatísticas principais da variável *totpagam_realizado* para os métodos com restrições

Medidas/ Métodos	Com Restrição	Valor Mínimo	1° Quartil	Mediana	Valor Médio	3° Quartil	Valor Máximo	Variância
Originais	–	0	0	0	70613	9945	40058305	2.137×10^{11}
Modelo Aditivo de Ruído Independente	NÃO	−83725	−10638	7731	70643	29800	40068326	2.142×10^{11}
	SIM	0	0	0	71919	11834	40063421	2.135×10^{11}
Modelo Aditivo de Ruído Correlacionado	NÃO	−199821	−23451	11667	70437	51677	39905545	2.141×10^{11}
	SIM	0	0	0	74351	18079	39868096	2.113×10^{11}
Modelo EGADP	NÃO	−132539	−62905	50873	70613	153359	1205429	4.568×10^{10}
	SIM	0	0	0	79417	157138	1258789	2.77×10^{10}
Mapeamento Inverso após EGADP	NÃO	0	0	0	70613	9945	40058305	2.137×10^{11}
	SIM	0	0	0	78641	10053	40058305	2.529×10^{12}
Data Shuffling	NÃO	0	0	0	70613	9945	40058305	2.137×10^{11}
	SIM	0	0	0	79062	10500	40058305	2.850×10^{11}

Pelas tabelas apresentadas é visível que as estatísticas principais dos modelos Aditivos de Ruído Correlacionado e de Ruído Independente com restrições possuem valores próximos dos originais. Relativamente ao modelo EGADP com restrição, este apresenta valores mais próximos dos originais comparativamente ao

modelo EGADP sem restrição, no entanto, continuam a existir diferenças elevadas para algumas variáveis. O método de Mapeamento Inverso após EGADP com restrição não possui grandes melhorias nas suas estatísticas face ao método aplicado sem restrições, obtendo-se valores aproximados em vez de valores iguais. O modelo de *Data Shuffling* com restrição não sofre grandes alterações, apenas deixa de possuir tantos valores iguais, passando a ter valores aproximados. Desta forma, é possível afirmar que a modificação dos métodos provocou vantagens nos modelos de Ruído Correlacionado, Ruído Independente e EGADP, pois com algumas descidas das medidas de perda de informação obtiveram-se estatísticas mais próximas das originais e apenas com valores positivos. Quanto aos outros dois modelos (Mapeamento Inverso com restrição e *Data Shuffling* com restrição), as vantagens/desvantagens não são claras, pois a perda de informação e o risco de identificação não sofreram alterações consideráveis, no entanto, as estatísticas principais apresentam algumas diferenças face às variáveis originais. Assim, apesar das diferenças maiores existentes em alguns métodos, as restrições impostas conduzem a resultados consistentes ao longo das variáveis, provocando aumentos consideráveis apenas no risco de identificação, que em todos os casos permanece inferior a 50%.

Visto que o modelo de EGADP continua a apresentar estatísticas distintas dos valores originais, este modelo não apresenta os resultados pretendidos quando comparado com outros métodos apresentados. Como foi referido no Capítulo 4, não é aconselhável aplicação deste modelo a bases de microdados de grande dimensão. Como no caso em estudo a base de microdados apresenta uma dimensão relativamente elevada, é justificável a elevada perda de informação nos resultados apresentados para este modelo.

Neste momento a escolha do método está entre os métodos de Mapeamento Inverso após EGADP com restrição, o modelo Aditivo de Ruído Correlacionado com restrições, o modelo Aditivo de Ruído Independente com restrições, *Data Shuffling* com restrição, o modelo de Ruído Multiplicativo e a Microagregação em Componentes Principais. Todos estes métodos apresentam estatísticas relativamente próximas das variáveis originais, bem como medidas de perda de informação reduzidas e um risco máximo de identificação próximo ou inferior do limite estabelecido (50%).

Para uma melhor escolha do método, apresentam-se nas Tabelas 40 e 41 os diagramas de dispersão das variáveis perturbadas e das variáveis originais dos métodos referidos anteriormente.

Por análise das Tabelas 40 e 41, é percetível que o modelo Aditivo de Ruído Correlacionado com restrições e o modelo Aditivo de Ruído Independente com restrições são os que apresentam melhores resultados, uma vez que a representação ggráfica destas duas variáveis aproxima-se a uma reta. Já o modelo de Ruído Multiplicativo e a Microagregação em Componentes Principais apresentam um conjunto de pontos com a forma de uma nuvem, o que representa valores relativamente próximos dos originais, mas não suficientemente próximos para que seja possível visualizar uma reta nos gráficos.

Quanto aos métodos de *Data Shuffling* com restrição e Mapeamento Inverso após EGADP com restrição, estes apresentam um pequeno conjunto de pontos que formam uma reta, no entanto, existem diversos pontos dispersos que correspondem a grandes discrepâncias entre os valores originais e os valores perturbados. Como seria de esperar, estes dois métodos apresentam gráficos muito parecidos.

Após a análise de tabelas e gráficos apresentados, já se possui informação suficiente para uma escolha mais adequada do método de perturbação a aplicar nesta base de dados. Das medidas de perda de informação apresentadas nas Tabelas 40 e 41 e nas Figuras 18 e 19, o modelo Aditivo de Ruído Independente com restrições é o que apresenta menores valores, no entanto, o risco de identificação máximo apresenta um valor relativamente elevado, cerca 42.50%. Quanto às estatísticas principais dos métodos, o modelo de *Data Shuffling* com restrição e de Mapeamento Inverso após EGADP com restrição, em geral, são os que apresentam estatísticas mais próximas dos valores originais.

Tabela 40: Gráficos entre as Variáveis perturbadas e as variáveis originais

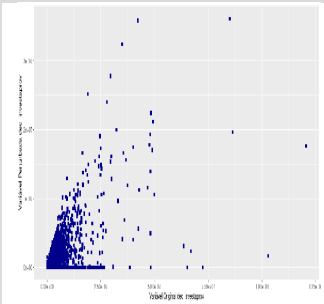
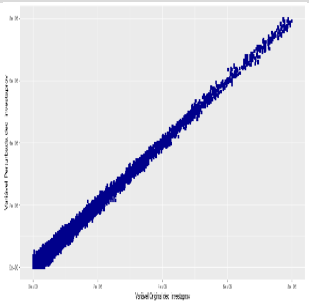
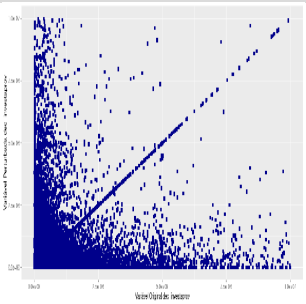
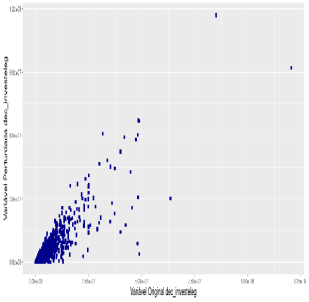
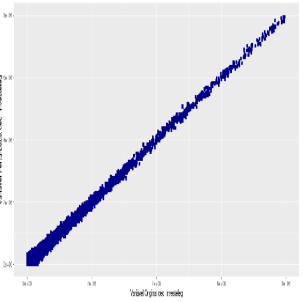
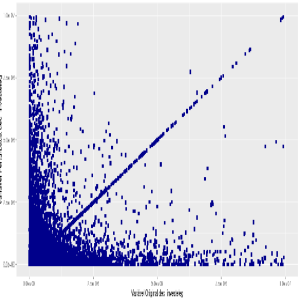
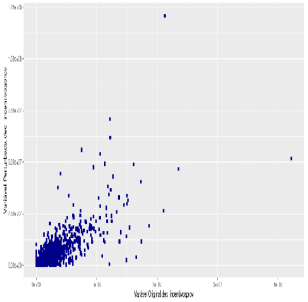
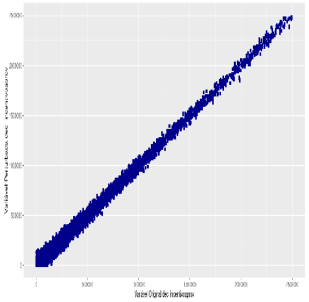
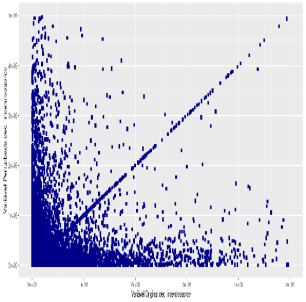
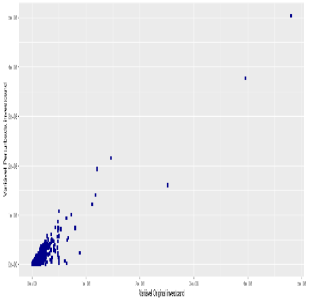
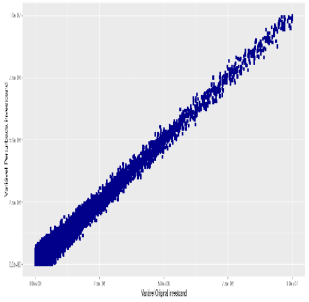
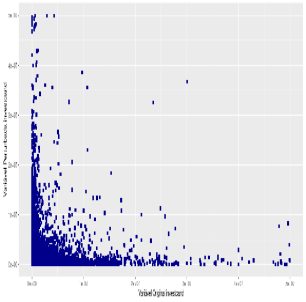
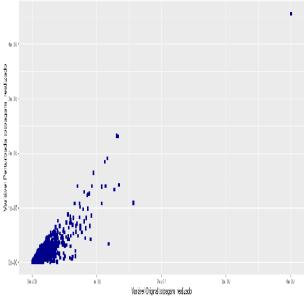
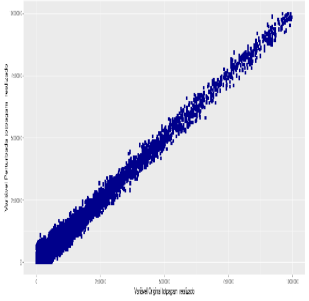
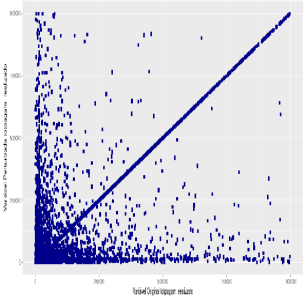
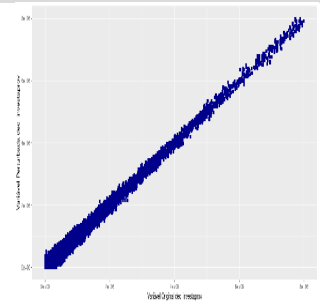
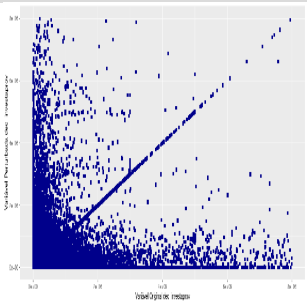
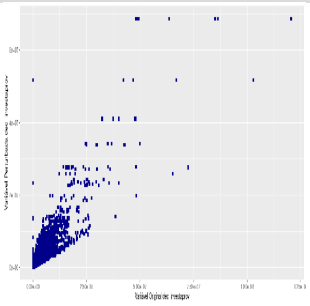
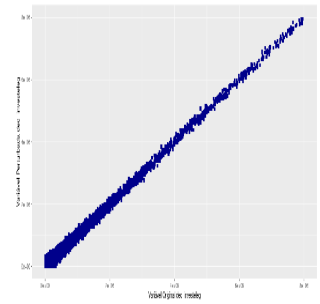
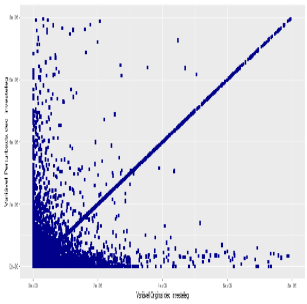
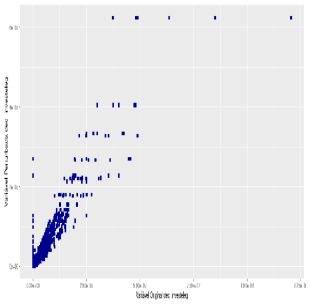
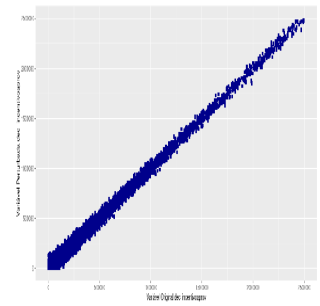
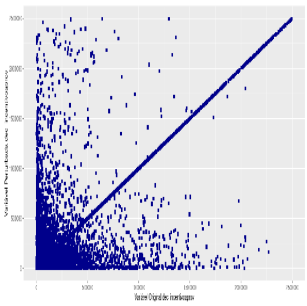
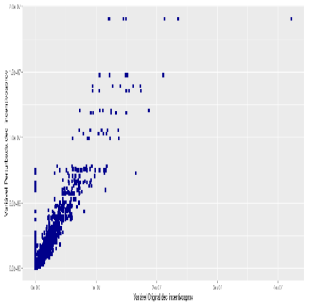
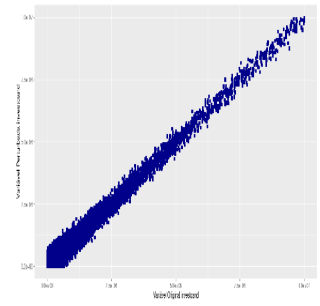
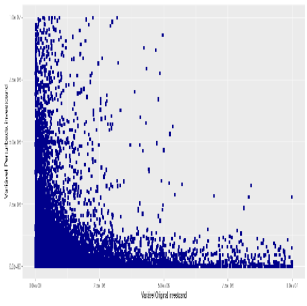
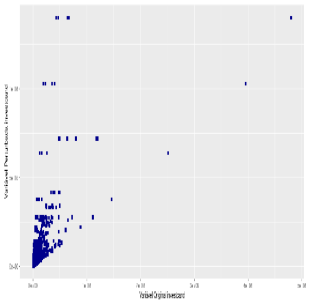
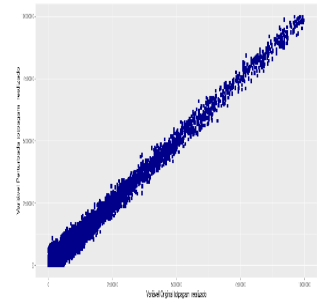
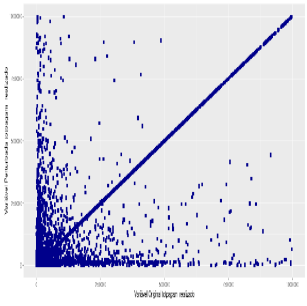
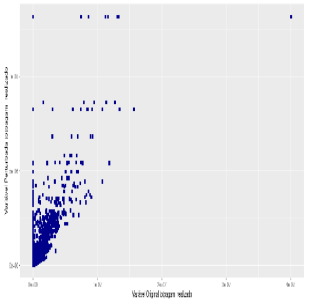
Métodos /Variáveis	Modelo de Ruído Multiplicativo	Modelo de Ruído Independente	Data Shuffling
dec_investaprov			
dec_investeleg			
dec_incentivoaprov			
investcand			
totpag_realizado			

Tabela 41: Gráficos entre as variáveis perturbadas e as variáveis originais

Métodos	Mapeamento Inverso Ruído Correlacionado	Mapeamento Inverso EGADP	Microagregação em CP
dec_investaprov			
dec_investeleg			
dec_incentivoaprov			
investcand			
totpag_realizado			

No entanto, os modelos de ruído com restrições apresentam também valores muito idênticos, em certos casos mais próximos aos valores originais que os outros dois modelos. Já nas relações entre as variáveis perturbadas e originais, é claro que os melhores resultados são dos modelos Aditivos de Ruído Independente e Correlacionado com restrições, pois a representação gráfica é aproximadamente uma reta para todas as variáveis perturbadas. Assim se o pretendido é obter a menor perda de informação possível, o modelo que se escolhe, é o modelo Aditivo de Ruído Independente com restrições, pois apresenta medidas de perda de informação bastante reduzidas face aos outros modelos, as principais estatísticas das variáveis perturbadas são próximas dos valores originais e possui uma correlação próxima de 1 entre as variáveis originais e as variáveis perturbadas. Estes resultados podem conduzir a um risco de identificação mais elevado, mas neste caso o risco de identificação não ultrapassa os 42.50%, ou seja, um valor inferior ao limite estabelecido no início da perturbação.

Se o objetivo da perturbação é obter o melhor compromisso entre o risco de identificação e a perda de informação, então a escolha deve ser o modelo Aditivo de Ruído Correlacionado com restrições, que apresenta um risco máximo de identificação de 16.76%, apresentando valores mais elevados, mas ainda próximos, para as medidas de perda de informação face ao modelo Aditivo de Ruído Independente com restrições.

Caso o objetivo da perturbação seja obter o menor risco de identificação possível, então a escolha seria o modelo EGADP com restrição, obtendo-se um risco máximo de identificação de 2.46%. Por outro lado, neste caso a perda de informação é consideravelmente maior que a perda de informação apresentada nos modelos de ruído com restrições.

Com o método já escolhido (modelo Aditivo de Ruído Independente com restrição dos valores positivos e dos valores nulos), apresenta-se na Tabela 42 as variâncias e as covariâncias das variáveis perturbadas e das variáveis originais. Como é visível as variâncias sofreram alterações mínimas, são valores muito próximos dos originais.

Tabela 42: Variâncias das variáveis perturbadas e originais do modelo Aditivo de Ruído Independente

Variáveis	dec_investaprov	dec_investeleg	dec_incentivoaprov	investcand	totpagam_realizado
dec_investaprov	4.742e+12 4.754e+12	3.014e+12 3.026e+12	1.149e+12 1.153e+12	4.590e+12 4.639e+12	5.906e+11 5.929e+11
dec_investeleg	3.014e+12 3.026e+12	2.826e+12 2.829e+12	1.084e+12 1.086e+12	2.960e+12 2.988e+12	5.697e+11 5.709e+11
dec_incentivoaprov	1.149e+12 1.153e+12	1.084e+12 1.086e+12	4.949e+11 4.950e+11	1.133e+12 1.144e+12	2.576e+11 2.579e+11
investcand	4.590e+12 4.639e+12	2.960e+12 2.989e+12	1.133e+12 1.144e+12	1.724e+13 1.731e+13	5.825e+11 5.886e+11
totpagam_realizado	5.906e+11 5.929e+11	5.697e+11 5.709e+11	2.576e+11 2.579e+11	5.825e+11 5.886e+11	2.135e+11 2.137e+11

Na Tabela 43 apresenta-se os coeficientes de correlação entre as variáveis originais, entre variáveis perturbadas pelo modelo Aditivo de Ruído Independente e as variáveis originais e entre as variáveis perturbadas pelo modelo Aditivo de Ruído Correlacionado e as variáveis originais.

Os coeficientes de correlação são idênticos nos três cenários apresentados. No entanto, é percetível que a correlação entre as variáveis perturbadas pelo modelo de Ruído Independente com restrições são superiores às do modelo de Ruído Correlacionado com restrições, o que seria de esperar, pois o risco de identificação do modelo de ruído independente com restrições é bastante superior à do ruído correlacionado com restrições.

Tabela 43: Coeficientes de correlação

Variáveis	dec_investaprov	dec_investeleg	dec_incentivoaprov	investcand	totpagam_realizado
dec_investaprov	1.00 **0.9992** **0.9969**	0.8250 0.8246 0.8212	0.7514 0.7505 0.7492	0.5113 0.5105 0.5037	0.5882 0.5875 0.5863
dec_investeleg	0.8250 0.8237 0.8230	1.00 **0.9994** **0.9978**	0.9175 0.9166 0.9159	0.4270 0.4260 0.4215	0.7343 0.7338 0.7332
dec_incentivoaprov	0.7514 0.7507 0.7485	0.9175 0.9171 0.9146	1.00 **0.9995** **0.9982**	0.3908 0.3904 0.3842	0.7931 0.7929 0.7916
investcand	0.5113 0.5082 0.5086	0.4270 0.4248 0.4241	0.3908 0.3882 0.3893	1.00 **0.9991** **0.9967**	0.3060 0.3040 0.3045
totpagam_realizado	0.5882 0.5876 0.5859	0.7343 0.7340 0.7324	0.7931 0.7926 0.7920	0.3060 0.3055 0.3010	1.00 **0.9997** **0.9987**

Como o modelo de ruído correlacionado com restrições apresenta valores muito distintos nas medidas de perda de informação IL1, na diferença de valores próprios e EQM das variâncias e ainda as correlações entres os valores originais e os valores perturbados são inferiores comparativamente ao modelo de Ruído Independente com restrições, escolhe-se o modelo de ruído independente com restrições, que apresenta um risco de identificação inferior ao limite estabelecido e maioria das medidas de perda de informação são as menores de todos os métodos apresentados. No entanto, o modelo de Ruído Correlacionado com restrições é uma ótima escolha para um cenário de compromisso ótimo entre perda de informação e o risco de identificação.

É ainda possível realizar a comparação dos Coeficientes de Gini das variáveis originais e das variáveis perturbadas pelo modelo de Ruído Independente com restrições. Na Tabela 44 estão apresentados os coeficientes para cada variável original, variável perturbada pelo modelo de Ruído Independente com restrições, variável perturbada pelo modelo de Ruído Correlacionado com restrições e as respetivas diferenças entre os mesmos.

Tabela 44: Coeficiente de Gini

Variáveis	Coeficiente de Gini (%)	Diferença (%)
dec_investaprov	82.2049 79.6055 76.7565	 2.5994 5.4484
dec_investeleg	89.1551 86.9204 84.5504	 2.2347 4.6047
dec_incentivoaprov	89.7022 89.0430 88.0190	 0.6591 1.6831
investcand	79.1964 71.6055 65.0962	 7.5993 14.1002
totpagam_realizado	94.2918 93.7971 92.8930	 0.4947 1.3989

Como é visível a desigualdade observada nos valores das variáveis perturbadas pelo modelo de Ruído Independente com restrições é semelhante à verificada nos valores das variáveis originais, sendo 7.5993% a maior diferença entre os coeficientes de Gini (variável *investcand*). O modelo de Ruído Correlacionado com restrições apresenta coeficientes mais distintos dos originais, sendo as diferenças destes coeficientes o dobro das diferenças do modelo de Ruído Independente com restrições. Portanto a desigualdade observada nos valores das variáveis perturbadas pelo modelo de Ruído Correlacionado apresenta diferenças consideráveis face à desigualdade presente nos valores das variáveis originais. Por outro lado, os valores das variáveis perturbadas pelo modelo de Ruído Independente com restrições contêm aproximadamente o mesmo nível de desigualdade observada nos valores das variáveis originais, uma característica importante a ser preservada para as análises dos utilizadores.

Para finalizar, apresenta-se nos Anexos I e II os gráficos entre as variáveis originais e os gráficos entre as variáveis perturbadas pelo modelo de Ruído Independente com restrições e como é percetível os gráficos são semelhantes nos dois casos, isto é, existe grande utilidade na base de microdados perturbada. Concluindo-se que o modelo Aditivo de Ruído Independente com restrição de valores positivos e valores nulos provoca alterações relativamente reduzidas na base de microdados original e fornece um risco de identificação satisfatório.

8. Conclusão e Trabalho Futuro

A proteção da confidencialidade dos dados é uma tema relativamente recente. Com o aumento de procura de informação, sentiu-se a necessidade de melhorar as técnicas disponíveis para a proteção de informação confidencial. Atualmente, a literatura oferece um conjunto diversificado de métodos que modificam a base de microdados por forma a assegurar que informação confidencial não é divulgada e ao mesmo tempo possibilitam aos investigadores retirar conclusões plausíveis face aos dados originais.

Nos últimos anos o BPLIM tem permitido que os investigadores externos nacionais e estrangeiros possam aceder a bases de microdados que contêm informação sobre entidades (indivíduos, empresas, etc) e sobre a economia portuguesa. Assim, é importante que estas bases de microdados divulgadas não forneçam informação confidencial.

O estudo realizado nesta Dissertação de Mestrado pretende aplicar e comparar, através da aplicação e comparação de métodos de Controlo de Divulgação Estatística (CDE) numa base de microdados disponibilizada pelo BPLIM. Sendo o foco a comparação dos métodos perturbativos, isto é, perceber qual dos métodos perturbativos oferece uma base de microdados perturbada mais próxima dos dados originais e com o menor risco de identificação possível.

Após o levantamento das metodologias e respetivos indicadores de avaliação, conclui-se que na literatura o modelo *Exact General Additive Data Pertubation* (EGADP) e o modelo *Data Shuffling* são os métodos que oferecem o melhor compromisso entre o risco de identificação e a perda de informação. O modelo EGADP é aconselhado na perturbação de base de microdados com uma dimensão relativamente reduzida e o modelo *Data Shuffling* é aconselhado na perturbação de base de microdados onde se pretende que relações não lineares sejam preservadas. Para além destes, outros métodos são apresentados como métodos que preservam diversas propriedades da base de microdados original, por exemplo, o modelo Aditivo de Ruído Correlacionado e o modelo aditivo de Ruído Independente preservam diversas estatísticas da base de microdados original.

Um conjunto de ferramentas disponíveis para a aplicação destes métodos foi analisado e conclui-se que a linguagem de programação R é a que oferece uma maior variedade de métodos e medidas, uma vez que dispõe do *package* **sdcMicro** que permite aplicar a maioria dos métodos descritos na literatura e ainda métodos disponíveis em outros *softwares*. Este *package* possibilita a aplicação dos métodos de uma forma simples. Na perturbação da base de microdados em estudo percebeu-se que esta ferramenta possui a desvantagem de um elevado tempo computacional em certos casos, mas que muitas vezes se deve à elevada complexidade de algumas metodologias.

Uma breve descrição sobre o *package* **sdcMicro** foi apresentada, onde se explica como aplicar as diversas funções correspondentes aos diferentes métodos e medidas usando uma base de microdados disponibilizada pela linguagem de programação R, com o objetivo de fornecer uma melhor compreensão da aplicação dos métodos com o *package* **sdcMicro**.

Com uma base de microdados disponibilizada pelo BPLIM com 43333 observações e 60 variáveis aplicaram-se e compararam-se os métodos perturbativos. Para dar inicio à perturbação, analisaram-se as variáveis de forma a perceber quais seriam as variáveis chave e quais seriam as variáveis sensíveis. No caso das variáveis chave, realizou-se a análise de várias combinações de variáveis, concluindo-se que as variáveis escolhidas seriam: *medida* (Medida específica dentro dos respetivos sistemas de incentivo da operação), *compagamentos* (Variável binária que assume valor 1 caso o projeto tenha recebido pagamentos e 0 caso contrário), *pofinan* (Programa Operacional (PO) do projeto submetido), *dom_interv* (Domínio de intervenção de acordo com a Comissão Europeia) e *apoiado* (Variável binária que assume valor 0 quando o projeto não é financiado e valor 1 caso contrário). Quanto às variáveis sensíveis, com ajuda de colaboradores do BPLIM percebeu-se que as variáveis *dec_investaprov* (Montante total decidido no momento de aprovação do projeto), *dec_investeleg* (Montante total de

incentivos aprovados no momento de aprovação do projeto), *dec_incentivoaprov* (Montante total de incentivos aprovados no momento de aprovação do projeto), *investcand* (Despesas totais co-financiadas, não co-financiadas e as não elegíveis necessárias de forma a cumprir os objetivos estabelecidos) e *totpagam_realizado* (Montante de pagamentos realizados) possuíam informação confidencial que não deveria ser relevada aos utilizadores.

Após a escolha das variáveis, iniciou-se a perturbação das variáveis chave. Esta perturbação teve como objetivo a eliminação das combinações únicas de variáveis chave. Assim, aplicou-se o método de Supressão Local em observações que possuiram um risco de identificação individual elevado. Para além deste método, ainda se aplicou o método PRAM às variáveis *proj_i40, estadofse e fundo*. A perturbação realizada através destes dois métodos provocou alterações mínimas nas variáveis chave e nas relações entre as variáveis, existindo apenas um aumento de 0.5% de valores em falta na base de microdados. As tabelas de contingência, entre as variáveis que sofreram PRAM, são idênticas antes e após a aplicação do método. Concluiu-se assim que as perturbações realizadas, resultaram numa boa utilidade dos dados e num baixo risco de identificação, não existindo portanto combinações únicas.

Os métodos perturbativos de CDE foram aplicados às variáveis numéricas. Nesta perturbação foi visível que os métodos produziram bases de microdados muito diferentes entre si, isto é, existem métodos que resultam em variáveis idênticas às variáveis originais e outros métodos que resultam em variáveis distintas das originais. Para além da aplicação dos métodos, sentiu-se a necessidade de impor algumas restrições em certos métodos, desta forma obteve-se variáveis perturbadas apenas com valores positivos, já que se tratavam de variáveis relativas a montntes de financiamento. Na aplicação dos métodos de CDE à base de microdados real, concluiu-se que existem métodos com valores reduzidos para as medidas de perda de informação, mas ainda assim, que conduziam a valores sem sentido para o estudo, como por exemplo, valores negativos. Alguns dos métodos apresentavam valores perturbados idênticos ou iguais no que diz respeito às estatísticas principais das variáveis originais, no entanto, apresentavam medidas de perda informação bastante elevadas. Como as variáveis sensíveis se referem a montantes monetários e a base de microdados possui diversas variáveis relacionadas com as variáveis sensíveis, realizou-se a alteração de quatro métodos (*Data Shuffling*, modelo Aditivo de Ruído Correlacionado, modelo Aditivo de Ruído Independente e modelo EGADP) impondo restrições de forma que estes métodos produzissem apenas valores positivos e as observações com valor nulo nos dados originais não sofressem alterações. Desta forma, foi possível obter variáveis perturbadas mais próximas das variáveis originais e com menor perda de informação, mas por consequência com maior risco de identificação.

As conclusões retiradas dos resultados da aplicação a uma base de microdados real diferem das conclusões obtidas na literatura, o modelo *Data Shuffling* não apresenta medidas de perda de informação muito reduzidas comparativamente a alguns métodos, no entanto, preserva a maioria das estatísticas principais de todas as variáveis sensíveis. Por outro lado, o modelo EGADP é dos modelos que mais alterações provoca nas principais estatísticas das variáveis sensíveis, possuindo medidas de perda de informação bastante elevadas. No entanto, na literatura aconselhou-se a aplicação deste modelo em bases de microdados de dimensão reduzida, e a base de microdados utilizada apresentava uma dimensão relativamente elevada, sendo assim, uma possível justificação para o mau desempenho do método. Com a base de microdados fornecida pelo BPLIM, concluiu-se que o modelo Aditivo de Ruído Independente com restrições é o modelo que oferece os melhores resultados nas medidas de perda de informação, contendo um risco máximo de identificação de 42.50%. Assim, caso se pretenda um risco de identificação mais reduzido o modelo Aditivo de Ruído Correlacionado com restrições é o modelo a utilizar, fornecendo maior perda de informação, mas contém um compromisso ótimo entre o risco de identificação (16.76%) e a perda de informação. Para estes dois casos, analisou-se a correlação entre as variáveis perturbadas

e as variáveis originais, obtendo-se valores muito semelhantes aos valores originais. Assim, as restrições impostas nestes modelos, resultaram em variáveis perturbadas bem mais próximas das variáveis originais.

Como é percetível as principais conclusões retiradas na literatura e na prática são diferentes, o que mostra que a escolha do método para a perturbação dependerá da base de microdados em estudo. Nesta dissertação, a base de microdados a perturbar possui características que permitem que os modelos Aditivos de Ruído Independente e de Ruído Correlacionado obtenham resultados mais próximos que os outros métodos, mesmo sem impor qualquer tipo de restrição. Após as restrições, os métodos preservam as estatísticas importantes da base de microdados original, sem que o risco de identificação ultrapasse o limiar estabelecido.

Trabalho Futuro

Como referido a procura de bases de microdados tem aumentado significativamente nos últimos anos, com uma tendência cada vez mais acentuada para os próximos anos. Com este aumento de procura, naturalmente surgirão requisitos de dados com estruturas diferentes, por exemplo, dados longitudinais e dados geográficos. Assim, um trabalho a realizar futuramente seria o desenvolvimento e a aplicação de diversos métodos nos diferentes tipos de dados.

Atualmente, os métodos existentes para estruturas diferentes de microdados são em pequena quantidade e são métodos pouco eficientes. O aconselhável é a aplicação dos métodos descritos nesta dissertação a esse tipo de dados, por exemplo, aplicá-los aos diferentes períodos de tempo existentes numa base de microdados longitudinal. Como foi visto nesta dissertação, é possível realizar alterações nos métodos de forma a obter resultados mais seguros ou mais próximos dos valores originais.

Uma outra possível área de interesse é a geração de bases de Dados Sintéticas, pois estes tipos de métodos permitem a divulgação de bases de microdados com grande utilidade nos dados e baixo risco de identificação. Esta área de CDE contém diversos métodos capazes de realizar a geração de bases de microdados de forma eficiente e recentemente têm surgido novos métodos .

Assim, é percetível que a perturbação de uma base de microdados ainda contém muitas questões e muitas áreas de interesse, no entanto é provável que seja necessário desenvolver novos métodos capazes de perturbar uma base de microdados, com qualquer tipo de estrutura, de forma eficiente do ponto de vista do risco de identificação e do ponto de vista do utilizador.

Como trabalho futuro, CDE é uma área que pode conduzir a grandes desenvolvimentos sobre novas metodologiass e *softwares*. Assim, entende-se que esta é uma área promissora que permitirá gerar conhecimento em técnicas com elevado interesse de aplicação.

Bibliografia

[1] Aggarwal Charu e Yu Philip. *Privacy Preserving Data Mining: Models and Algorithms*. Springer New York, NY, (2008). DOI: https://doi.org/10.1007/978-0-387-70992-5.

[2] Banco de Portugal Microdata Research Laboratory (BPLIM): Incentives Systems Data. "Dataset". Em: *BANCO DE PORTUGAL* (2021). DOI: https://doi.org/10.17900/SI.APR2021.V1.

[3] Benschop Thijs, Machingauta Cathrine e Welch Matthew. "Statistical Disclosure Control: A Practice Guide". Em: *The World Bank* (2021).

[4] Burridge Jim. "Information preserving statistical obfuscation". Em: *Statistics and Computing* 13 (2003), pp. 321–327. ISSN: 1573-1375. DOI: https://doi.org/10.1023/A:1025658621216.

[5] Dalenius Tore. "Towards a methodology for statistical disclosure control". Em: *Statistics Sweden*. Vol. 15. (1977), pp. 429–444.

[6] Daniel Ting, Stephen E. Fienberg e Mario Trottini. "ROMM Methodology for Microdata Release". Em: *UNECE* ((2005)).

[7] Domingo-Ferrer Josep. *Confidentiality, Disclosure,and Data Access*. 1ª ed. Elsevier Science, (2001).

[8] Domingo-Ferrer Josep e Muralidhar Krish. "New directions in anonymization: Permutation paradigm, verifiability by subjects and intruders, transparency to users". Em: *Information Sciences* (2016), pp. 11–16.

[9] Domingo-Ferrer Josep e Torra Vicenç. *Privacy in Statistical Databases*. Springer, (2004).

[10] Eurostat. *Description of target variables: Cross-sectional and longitudinal*. Statistics, Social e Society, Information, (2004).

[11] Hundepool Anco et al. *Handbook on Statistical Disclosure Control*. Vol. 1. (2006).

[12] Kim Jay. "A method for limiting disclosure in microdata based on random noise and transformation". Em: *Proceedings of the American Statistical Association, Survey Research Methods Section* (1986), pp. 370–374.

[13] Mendes Elsa. "Confidencialidade de Dados: Aplicação e Comparação de Técnicas de Controlo da Divulgação Estatística". Português. Tese de mestrado. FEP, (2010).

[14] Menéndez Uría. Em: *Actualidad Jurídica*. 53ª sér. (2019), pp. 142–148. ISSN: 1578-956X.

[15] Mivule Kato. "Utilizing Noise Addition for Data Privacy, an Overview". Em: IKE, (2012). DOI: 10.13140/2.1.4629.2482.

[16] Muralidhar Krishnamurty, Rathindra Sarathy e Domingo-Ferrer Josep. "Reverse Mapping to Preserve the Marginal Distributions of Attributes in Masked Microdata". Em: *Privacy in Statistical Databases*. Ed. por Josep Domingo-Ferrer. Cham: Springer International Publishing, (2014), pp. 105–116.

[17] Muralidharand Krishnamurty, Parsa Rahul e Sarathy Rathindra. "A General Additive Data Perturbation Method for Database Security". Em: *Management Science 45(10):* (1999), pp. 1399–1415.

[18] Navarro-Arribas Guillermo e Torra Vicenç. *Advanced Research in Data Privacy*. Vol. 1. Springer, (2015).

[19] "Portugal 2020". Em: *Acordo de Parceria 2014-2020* (2014).

[20] Rao C. Radhakrishna. *Handbook of Statistics:Bioinformatics in Human Health and Heredity*. Vol. 28. (2012).

[21] Rathindra Sarathy e Krish Muralidhar. "Perturbation Methods for Protecting Numerical Data: Evolution and Evaluation". Em: ed. por Ranajit Chakraborty, C. Radhakrishna Rao e Pranab Sen. Vol. 28. Handbook of Statistics. Elsevier, (2012), pp. 513–531. DOI: https://doi.org/10.1016/B978-0-44-451875-0.00019-1.

[22] Ruiz Nicolas. "A General Framework and Metrics for Longitudinal Data Anonymization". Em: *Privacy in Statistical Databases*. Ed. por Domingo-Ferrer Josep e Montes Francisco. Springer International Publishing, (2018), pp. 215–230.

[23] Sarathy Rathindra e Muralidhar Krish. *Protecting Numerical Confidential Data using Data Shuffling: A Demonstration of Effectiveness of Approach and Flexibility of Delivery*. (2006).

[24] Templ Matthias. *Statistical Disclosure Control for Microdata: Methods and Applications in R*. English. 1ª ed. Vol. 1. Springer International Publishing, (2017).

[25] Templ Matthias, Meindl Bernhard e Kowarik Alexander. *Introduction to Statistical Disclosure Control (SDC)*. IHSN, (2021).

[26] Templ Matthias, Meindl Bernhard e Kowarik Alexander. *Package 'sdcMicro'*. Rel. téc. (2021).

[27] Tendick Patrick e Matloff Norman. "A modified random perturbation method for database security". Em: *ACM Transactions Database System* 19 (1994), pp. 47–63. DOI: https://doi.org/10.1145/174638.174641.

Anexo I - Gráficos das variáveis sensíveis originais

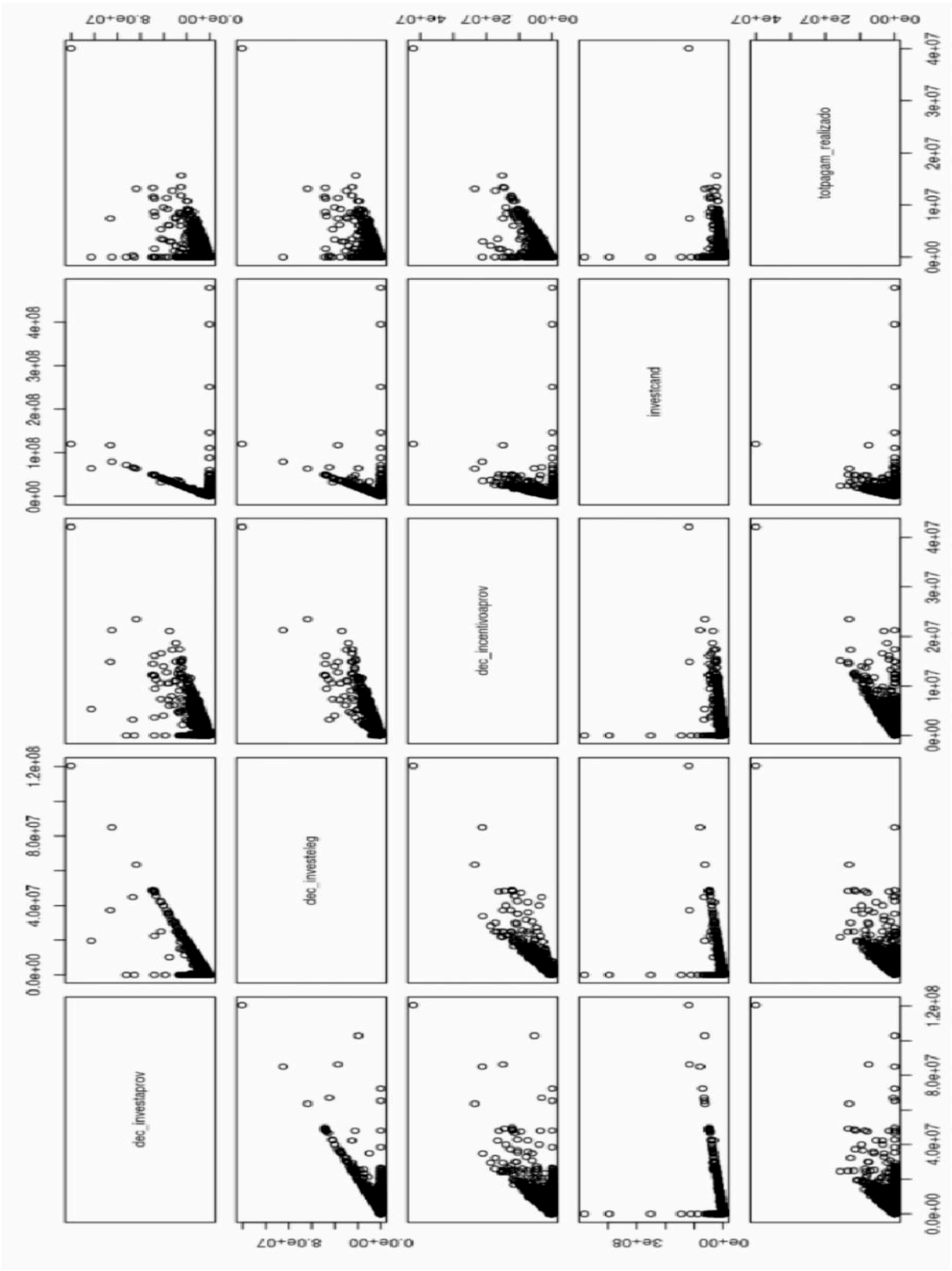


Figura 20: Gráficos entre as variáveis originais sensíveis

Anexo II - Gráficos das variáveis sensíveis perturbadas pelo modelo escolhido

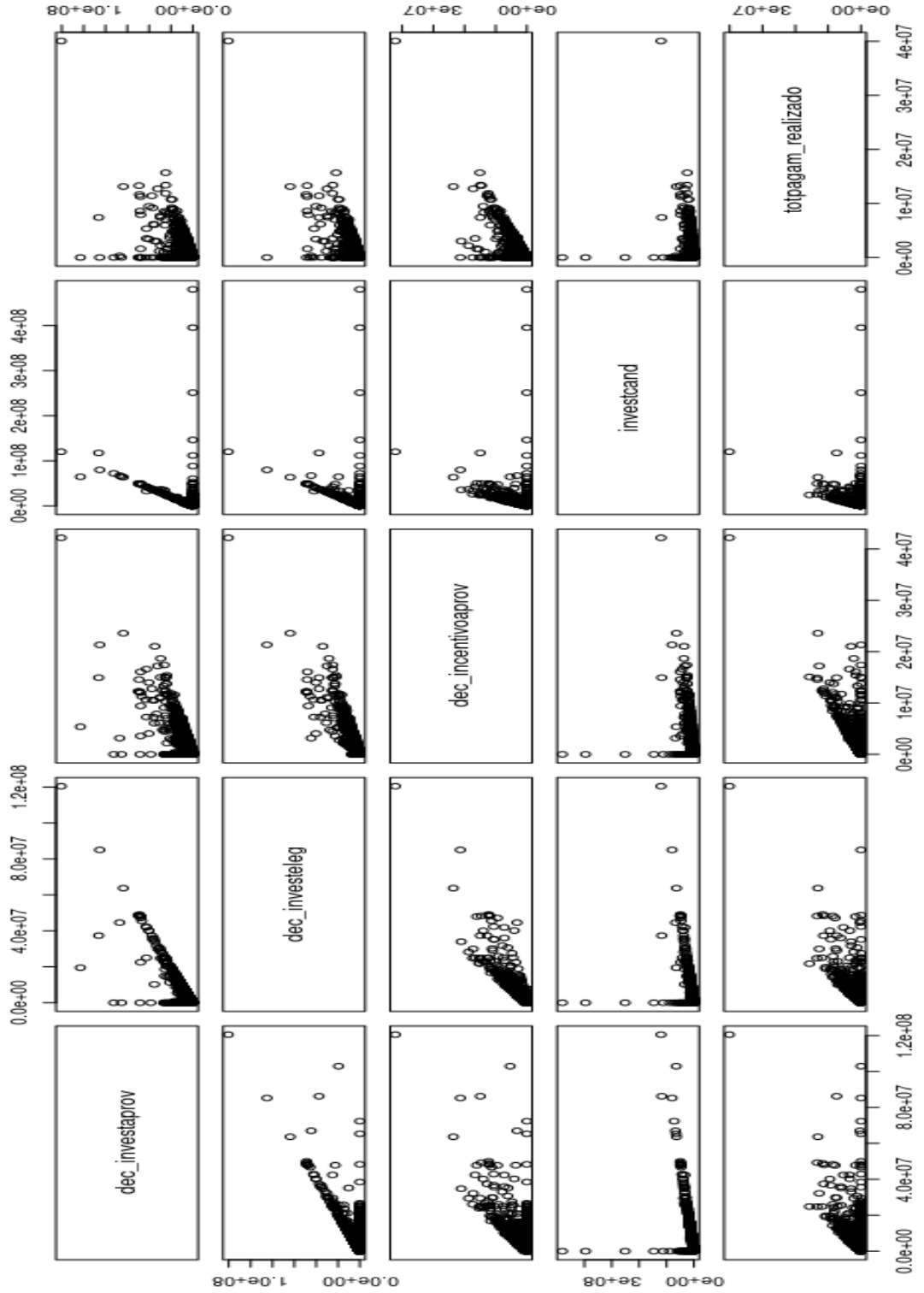


Figura 21: Gráficos entre as variáveis perturbadas sensíveis