Universidade do Minho
Escola de Ciências

João Pedro Dioniso Rego

Avaliação de Metodologias de Controlo de Confidencialidade: Uma aplicação com bases de microdados

Avaliação de Metodologias de Controlo de Confidencialidade:
Uma aplicação com bases de microdados
João Pedro Dioniso Rego
UMinho | 2021

dezembro, 2021

Universidade do Minho
Escola de Ciências

João Pedro Dioniso Rego

Avaliação de Metodologias de Controlo de Confidencialidade: Uma aplicação com bases de microdados

Projeto de Estágio
Mestrado em Estatística

Trabalho realizado sob orientação de
Professor Doutor Luís Filipe Meira Machado
Doutora Rita Cristina Sousa

dezembro, 2021

Despacho RT – 31 / 2019 – Anexo 3

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 udo 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

(Caso o autor pretenda usar uma das licenças Creative Commons, deve escolher e deixar apenas um dos seguintes ícones e respetivo Lettering e URL, eliminando o texto em itálico que se segue.

Contudo, é possível optar por outro tipo de licença, devendo, nesse caso ser incluída a informação necessária adaptando devidamente esta minuta)

Despacho RT - 31 /2019 - Anexo 4

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.

Agradecimentos

Gostaria de expressar os meus agradecimentos a diversas pessoas e instituições pela colaboração na realização deste projeto, sem os quais não seria possível concretizar.

Em primeiro lugar um agradecimento à orientadora deste projeto de estágio, Doutora Rita Cristina Sousa, por toda a orientação e ajuda que deu na elaboração deste projeto, uma vez que demonstrou sempre uma grande disponibilidade, profissionalismo e compreensão.

Quero agradecer também ao Professor Doutor Luís Filipe Meira Machado por toda ajuda, disponibilidade, palavras de força e positivismo, que sempre demonstrou na elaboração deste projeto.

Não podendo esquecer do Laboratório de Investigação em Microdados do Banco de Portugal (BPLIM), pelo grande profissionalismo e prontidão na resolução de problemas demonstrado, pelo acesso ao servidor, e pela disponibilização do ficheiro de microdados anonimizado.

Por último, e não menos importante, quero agradecer à minha família, nomeadamente ao meu irmão e minha cunhada, pelo apoio demonstrado, mas principalmente à pessoa que mais me incentivou e apoiou, Carla Almeida. Muito obrigado pela tua paciência, compreensão, dedicação, carinho, pela força e pelo teu amor.

Resumo

Avaliação de Metodologias de Controlo de Confidencialidade: Uma aplicação com bases de microdados

Devido à era da informação a que nos encontramos, cada vez mais, a procura por informação de qualidade por parte dos investigadores tem vindo a aumentar. Com a imposição das novas leis de confidencialidade, nos últimos anos, a utilização e a modernização de técnicas de controlo confidencial têm vindo a prosperar. Deste modo, é enaltecida a importância de proporcionar aos investigadores ficheiros de microdados, com um *trade-off* entre o risco de divulgação e utilidade, o mais equilibrado possível. Este trabalho aborda as técnicas de controlo de divulgação, sendo o seu propósito a avaliação e comparação dos métodos de divulgação para efeitos de investigação. A biblioteca sdcMicro (Templ *et al*., 2015) do *software* R forma a base deste estudo. Todos os métodos apresentados neste estudo, foram obtidos com recurso à biblioteca sdcMicro. Foram aplicados métodos de controlo de divulgação a uma base anonimizada, donde se concluiu que os melhores métodos são a adição de ruido não correlacionado, através do indicador erro quadrático médio, e a microagregação tendo como medida de grupo a mediana nos restantes indicadores da qualidade dos dados respetivamente.

Palavras-Chave: Microdados, Métodos Perturbativos, Anonimização, Confidencialidade, Risco de divulgação, Perda de Informação; Variáveis-chave

Abstract

Evaluation of Confidentiality Control Methodologies: An application with microdata bases

Due to the information age we are in, more and more, the demand for quality information by researchers has been increasing. With the imposition of the new confidentiality laws in recent years, the use and modernization of confidentiality control techniques have been flourishing. Thus, the importance of providing researchers with microdata files, with a trade-off between disclosure risk and utility, in order to be as balanced as possible, is emphasized. This paper covers disclosure control techniques, and its purpose is to evaluate and compare disclosure methods for research purposes . The sdcMicro library (Templ *et al.* 2015) of the R software forms the basis of this study. All methods presented in this study, were obtained using the sdcMicro library. Disclosure control methods were applied to an anonymized database, and it was concluded that the best methods are the addition of uncorrelated noise, through the mean square error indicator, and microaggregation with the median as a group measure in the remaining indicators of data quality respectively.

KEYWORDS: Microdata, Perturbative Methods, Anonymization, Confidentiality, Risk of Disclosure; Information Loss, Key Variables;

Índice

Lista de abreviaturas siglas e acrónimos

BdP	Banco de Portugal
BPLIM	Laboratório de Investigação em Microdados do Banco de Portugal
BSD	Book Statistical Disclosure
CASC	Statistical Disclosure Central Microdata
CBFS	Centraid – Based Fixed Size
DBA	Density – Based Algorithm
DEE	Departamento de Estudos Económicos
FI	Ficheiros de Investigação
IHSN	International House hold Survey NetWork
MDAV	Maximum Distance Average Vector
NIF	Número Identificação Fiscal
NISS	Número de Identificação de Segurança Social
PRAM	Post Randomization Method
PUF/FUP	Ficheiros de Uso Público
RGPD	Regulamento Geral de Proteção de Dados
SAS	Analytics Software & Solutions
SPSS	Statistic Powerful Statistical Software
SSE	Soma do quadrado dos erros
STATA	Software For Statistics And Data Science
TFRP	Two Fixed Reference Points

Índice de tabelas

Índice de figuras

Capítulo 1

1- Introdução

No âmbito do Mestrado em Estatística, ministrado na Universidade do Minho, foi realizado um estágio curricular no Banco de Portugal, mais precisamente no Laboratório de Investigação em Microdados inserido no Departamento de Estudos Económicos, situado na cidade do Porto.

O estágio foi desenvolvido à distância e repartiu-se em duas componentes:

- Componente teórica, que consistiu numa revisão de literatura acerca das técnicas de anonimização;
- Componente prática, dedicada à aplicação dos métodos e avaliação dos mesmos.

Com a imposição das novas leis de confidencialidade, nos últimos anos, a utilização e a modernização de técnicas de controlo confidencial têm vindo a prosperar. Deste modo, o objetivo deste estudo consiste na aplicação e avaliação dos métodos de controlo de divulgação numa base de microdados, sendo o foco os métodos perturbativos.

O trabalho está estruturado em 10 capítulos. O **Capítulo 1** apresenta um enquadramento do tema abordado, e uma descrição do Banco de Portugal, em particular do Laboratório de Investigação em Microdados, BPLIM. O **Capítulo 2** aborda a estrutura de dados que um ficheiro pode ter, seguindo-se com a definição de um ficheiro de microdados, e respetiva política de acesso aos dados no BPLIM. O **Capítulo 3** é dirigido à classificação de dados, onde se refere a classificação quanto à divulgação e a classificação das variáveis de um ficheiro de microdados. O **Capítulo 4** apresenta as técnicas de controlo de confidencialidade. Aqui, são apresentadas técnicas de anonimização e as técnicas de controlo de divulgação. O **Capítulo 5** é reservado para a utilidade dos dados e perda de informação, onde são apresentadas as métricas de permitem quantificar a perda de informação de um ficheiro de microdados. O **Capítulo 6** refere indicadores para o risco de divulgação. O **Capítulo 7** apresenta um levantamento de ferramentas para aplicação dos métodos de controlo de divulgação e as suas desvantagens e vantagens. No **Capítulo 8** é realizado um tutorial da libraria *sdcMicro* do *software R*. Ainda na presente secção é apresentada um exemplo prático

para a maioria dos métodos referidos no **Capítulo 4**. No **Capítulo 9** é reservado para o estudo de caso. São apresentadas as variáveis do estudo, assim como uma análise descritiva e uma análise univariada. De seguida, são aplicados métodos e é feita a avaliação dos mesmos, através de indicadores de utilidade. Por fim, o **Capítulo 10** engloba os resultados das análises e comparações de métodos que se propôs estudar no trabalho, apresenta os aspetos mais relevantes, e as linhas possíveis para futuras investigações.

De referir que as estatísticas e resultados apresentados neste trabalho foram obtidos do *software R*, através da libraria *sdcMicro*.

Capítulo 2

2- Banco de Portugal

O Banco de Portugal (BdP) é o banco central da República Portuguesa. Fundado em 1846 em Lisboa, local da sua sede. Define-se como uma pessoa coletiva de direito público, com autonomia administrativa, financeira e património próprio. São órgãos do BdP o Governador, o Conselho de Administração, o Conselho de Auditoria e o Conselho Consultivo.

O BdP é a entidade emissora da moeda nacional, no entanto o *core business* não se concentra apenas neste ramo. Tem como propósito a manutenção da estabilidade dos preços, bem como a promoção da estabilidade do sistema financeiro, entre outras.

Figura 1: Instalações do Banco de Portugal, Porto

Laboratório de Investigação em Microdados do Banco de Portugal (BPLIM)

O Departamento de Estudos Económicos (DEE) do BdP tem como funções principais a realização de previsões sobre a economia portuguesa. Este departamento, está dividido em diversas áreas, sendo uma delas o BPLIM (em inglês *The Banco de Portugal Microdata Researh Laboratory*).

O BPLIM é uma unidade autónoma, sendo a sua principal missão, o apoio à produção de projetos e estudos de investigação acerca da economia portuguesa. Iniciou atividade em meados de 2016 e, deste então, tem vindo a desenvolver um trabalho fulcral no que concerne ao melhoramento da produção de projetos de pesquisa e gestão de

microdados para uso de investigadores internos e externos sobre múltiplas vertentes da atividade económica portuguesa. O BPLIM distingue-se de outras instituições que disponibilizam microdados, na medida em que não só os disponibilizam como dão tambem suporte científico e computacional através da sua equipa especializada e respetivos colaboradores.

2.1 Objetivo

Encontrámo-nos numa era digital. O volume de dados é cada vez maior e a velocidade com que os necessitamos de aceder impôs o uso atual de sistemas informáticos ao invés dos arquivos físicos utilizados anteriormente. Devido à evolução tecnológica, a eficiência e eficácia das metodologias utilizadas no tratamento de dados das diferentes organizações bem como nos centros de investigação melhoraram, ao mesmo tempo que introduziram um novo nível de segurança.

Parte da informação armazenada nesses sistemas, corresponde, na prática a dados pessoais, ou dados referentes a informação privada e, mesmo sendo tratadas de modo a suscitar a sua segurança e a proteção dos seus respetivos utilizadores, quando acedidos e relacionados com outras bases de dados, isto é, realizar cruzamento de dados, podem tornar-se relevantes para outras finalidades. Contudo, este tipo de ações põe em causa a privacidade dos titulares, e pela sua importância, podem até atingir a integridade e a disponibilidade dos responsáveis pelo tratamento de dados, neste caso o BPLIM.

O novo Regulamento Geral de Proteção de dados (RGPD) vem obrigar a uma reestruturação organizacional, uma nova abordagem na gestão de tratamento de dados pessoais, de modo que sejam preservados os direitos dos titulares, dos dados e os riscos de quebra de segurança, sendo minimizados. Do mesmo modo, que o RGPD leva que os dados obsoletos sejam eliminados, o mesmo promove a anonimização ou a pseudominização (perturbação) através de um processo de modificação, substituição ou remoção das suas características individuais – potenciais identificadores, por outras codificadas ou modificadas.

No entanto, apesar dos dados anónimos não serem considerados pessoais, existem diversos casos em que pode ocorrer falhas de segurança. Um exemplo, é o cruzamento de informações que pode levar à re-identificação de uma informação confidencial.

Devido a este fator, o risco de re-identificação dos respetivos titulares não deve ser descurado. Deste modo, o propósito deste projeto vai do encontro com a análise, comparação e avaliação das metodologias presentes no mercado académico, do controlo da confidencialidade numa base de dados para efeitos de investigação, sendo o objetivo primordial encontrar um método/abordagem que se aproxime do ponto de equilíbrio entre a utilidade, o risco de re-identificação bem como das falhas de segurança presentes.

Capítulo 3

3 Dados e Divulgação

Já não é de agora, que entender a composição dos dados, as suas características, e o mais importante, a informação que estes carregam, é deveras muito importante. É primordial que o indivíduo que está encarregue do tratamento dos dados, tenha uma perceção clara do objetivo da base de dados que lhe é apresentado, como o seu propósito, de modo que consiga garantir um grau de confidencialidade aceitável, caso assim seja pretendido. Ao longo deste capítulo serão descritos os diversos tipos de dados, a forma que são apresentados e divulgados, consoante os seus destinatários, isto é, os vários tipos de públicos a que são destinados.

3.1 Tipos de dados

Segundo a literatura, a estrutura do ficheiro de dados pode ser apresentada de duas formas: na forma de microdados e na forma de macrodados. Esta classificação, constitui uma das etapas prévias da aplicação das metodologias de confidencialidade.

Diz-se que se está na presença de um ficheiro de microdados quando o conjunto de registos contém informação sobre os respondentes individuais, entidades económicas sendo apresentada relativamente a uma determinada lista de variáveis (Huang & Williamson, 2001). Já no caso do ficheiro de macrodados, os ficheiros são apresentados de forma agregada, num formato tabular (Castro, 2012).

Nas posteriores secções e capítulos, será dada apenas enfase aos ficheiros de microdados ao invés de macrodados, uma vez que é o propósito do estudo.

3.2 Microdados

Os ficheiros de microdados, cada vez mais, têm sido alvo de grande atenção uma vez que detêm informação elementar não-agregada, acerca de operações e das características dos agentes económicos (indivíduos, famílias, empresas ou outro tipo de unidades).

Existem dois tipos de ficheiros de microdados que podem ser divulgados pelos centros de investigação:

- **Ficheiros de Uso Público (*Public Use File* ou FUP):** Ficheiros de dados, preparados por investigadores ou por fornecedores de dados, com o objetivo de os tornar acessíveis ao público. Estes ficheiros, são de acesso livre e sem problemas de confidencialidade.

- **Ficheiros de Investigação (FI):** Ficheiros de dados, preparados por laboratórios de investigação ou entidades estatísticas que podem ser facultadas aos investigadores credenciados para que os mesmos possam utilizar a informação necessária para os seus projetos, com a salvaguarda da confidencialidade da informação.

Alguns centros de investigação permitem o acesso a estes ficheiros de microdados em laboratórios de dados, sendo em alguns casos com a possibilidade de acesso remoto, como é o caso do BPLIM.

3.2.1 Investigadores que usam dados do BPLIM

Como foi referido previamente, o objetivo primordial do BPLIM consiste em proporcionar uma maior facilidade a nível de acesso e de utilização dos microdados acerca da economia portuguesa. O BPLIM dispõe de dois servidores, o interno e o externo. O servidor interno é destinado maioritariamente para os investigadores internos do BdP, no entanto, o servidor externo destina-se a investigadores externos e apresenta uma mais-valia, o acesso remoto.

Características dos conjuntos de dados

Todas as bases de dados do BPLIM são tratadas de forma minuciosa de maneira que não apresentem elementos (como por exemplo, o NIF[1] e o NISS[2]) que permitam a identificação direta de empresas ou até dos indivíduos.

Os conjuntos de dados estão disponíveis no formato STATA, e são armazenados de forma eficiente de modo que o tamanho do arquivo seja minimizado e detêm a nomenclatura BPLIM por convenção. Os conjuntos de dados disponibilizados, são sempre acompanhados por um manual e outros ficheiros de suporte, de forma, que quando solicitados, os usuários tenham ao seu dispor uma descrição minuciosa do ficheiro de microdados, de modo que consigam efetuar o seu trabalho de modo mais acessível. Os conjuntos de dados são também regularmente atualizados. A lista completa dos conjuntos de dados pode ser consultada no site do BPLIM, *https://bplim.bportugal.pt/* .

Acesso aos microdados

No ponto de acesso aos microdados, a política do BdP é bem esclarecedora. Só é permitido o acesso, caso os investigadores pretendam utilizar os microdados para fins científicos. Os colaboradores/investigadores internos podem ter acesso a todos os conjuntos de dados anonimizados disponibilizados no BPLIM no servidor interno. Contudo, o mesmo já não acontece com os investigadores externos. Os investigadores externos têm acesso aos microdados que são colocados no servidor externo, e de forma a preservar a confidencialidade dos dados, o BPLIM fornecerá versões modificadas dos dados originais, dependendo do grau da confidencialidade existente. Estes conjuntos de dados são disponibilizados para facilitar a elaboração dos códigos que estruturam e analisam os dados. No entanto, os resultados são obtidos com dados modificados, estes não são válidos para fins científicos. Portanto, após a finalização do código, os investigadores devem solicitar que os códigos sejam executados nos dados originais.

[1] Número de Identificação Fiscal
[2] Número de identificação da Segurança Social

Capítulo 4

4 Classificação dos dados

Antes dos dados serem disponibilizados é necessário entender a estrutura da privacidade inerente ao ficheiro de dados, isto é, saber se é possível divulgar a informação sem que os respondentes sejam identificados. Facto este, que leva a uma medição do risco de divulgação para constituir uma parte importante do processo de controlo de divulgação.

A utilização de estratégias adequadas que permitam medir e quantificar o risco são necessários quando a tomada de decisão vai de encontro com o facto de o ficheiro ser seguro ou não. Contudo, para medir e quantificar o risco, é necessário que de antemão, um panorama dos cenários possíveis de divulgação, estejam presentes.

Segundo a literatura, usualmente são definidos três tipos de divulgação (Matthias, 2017):

- **Divulgação da identidade:** A divulgação ocorre se um utilizador associa um indivíduo conhecido ou organização com os registos de dados disponibilizados;
- **Divulgação por atributo:** A divulgação ocorre se um utilizador determinar novas características de um indivíduo ou organização baseado nos registos de dados disponibilizados;
- **Divulgação por inferência:** A divulgação ocorre se um utilizador determina novas características de um indivíduo ou organização através de análises estatísticas;
  Exemplo: O utilizador utiliza um modelo de regressão com uma alta percentagem de previsão e através dos dados disponibilizados, o intruso consegue inferir acerca de outras características acerca do indivíduo ou organização sem os mesmos estarem identificados no ficheiro de dados anonimizado.

Embora a divulgação por inferência tenha sido definida como um tipo de divulgação, os métodos de controlo de divulgação para microdados tem como objetivo principal prever a divulgação da identidade e dos atributos, uma vez que o propósito é que os investigadores consigam realizar as devidas análises exploratórias, como as inferências estatísticas de modo que consigam entender as relações e os comportamentos das variáveis.

Classificação de Variáveis

No âmbito do processo de divulgação, a classificação das variáveis é descrita na Tabela 1.

Tabela 1 Classificação de variáveis

Variáveis identificadoras: contêm informações que podem levar à identificação dos respondentes e podem ser categorizadas como:

- **Identificadoras Diretas:** atributos que sem ambiguidade identificam o indivíduo (exemplos: Nome, apelido, nº passaporte, NIF, *etc*). A remoção dos identificadores diretos é a primeira etapa para o lançamento *à posteriori* de uma base de microdados mais segura. No entanto, realizar apenas a extração dos identificadores diretos, geralmente não é suficiente para uma base de microdados segura.

- **Identificadoras indiretas:** estas variáveis, também denominadas como variáveis-chave, contêm informações que, quando combinadas com outras identificadoras indiretas no conjunto de dados, pode levar à identificação dos respondentes (exemplos: localidade, género, idade, nº telefone, setor de atividade, etc). Estas variáveis ao contrário das identificadoras diretas, são variáveis que apresentam um grau de sensibilidade associado, isto é, carregam informação que podem ser importantes para as análises dos investigadores, e por isso, não devem ser removidas da base de dados.

Variáveis não identificadoras: São variáveis que não podem ser usadas para identificação dos respondentes. Estas variáveis não contêm nenhuma informação que possa identificar o indivíduo.

Capítulo 5

5 Técnicas de controlo de confidencialidade

5.1 Técnicas de anonimização

A publicação de dados pode levar a vários riscos de violação de privacidade devido à existência dos identificadores diretos e indiretos. Isso pode acarretar em consequências graves, por causa do uso não autorizado de informações sensíveis pertencentes aos indivíduos. Como forma de solucionar esse problema, uma estratégia ingênua seria a não publicação dos dados (Matthias, 2017). No entanto, estas medidas levariam a que os governos, as organizações e outras entidades não pudessem tirar proveito de análises, tal como identificação dos padrões e das tendências para a sociedade, dificultando assim, o possível crescimento das mesmas. Uma abordagem mais promissora para solucionar o problema da preservação da privacidade numa publicação é anonimizar os dados antes de qualquer disponibilização (Fung, 2010).

Entenda-se por anonimização, o uso de técnicas que convertem os dados confidenciais em dados anónimos (DGEEC, 2021). No presente caso de estudo, o processo de anonimização vai de encontro com a disponibilização da informação para o investigador de forma adequada. Como o objetivo, é não divulgar a informação que é confidencial, poderão ser utilizados procedimentos que levem a uma diminuição do tamanho base de dados, bem como, ocultá-la de forma que não seja reconhecível. No entanto, estas abordagens podem variar consoante o objetivo da respetiva base de dados.

As técnicas de anonimização reduzem o nível de identificação das entidades ou indivíduos de um conjunto de dados original para um nível mais aceitável. Uma avaliação do risco de identificação deve ser executada antes e após o processo de anonimização. A avaliação inicial assegura a estrutura e a informação dentro de um atributo que são claramente identificadas e compreendidas. Já a avaliação final irá determinar o risco residual de re-identificação. Desta forma, é importante no último passo haver um profissional que apresente um conhecimento sobre o assunto em causa.

As diferentes características das várias técnicas de anonimização levam a que algumas possam ser mais adequadas para uma situação do que outras. As técnicas mais utilizadas estão apresentadas na Tabela 2.

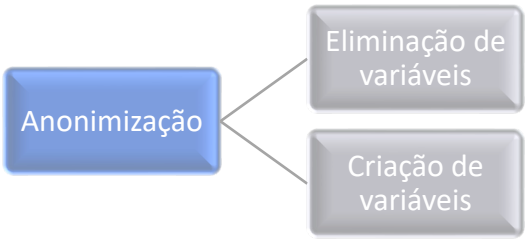


Tabela 2 Técnicas de anonimização

Eliminação de variáveis:

- Remoção de informação de identificação do conjunto de dados;

Criação de variáveis:

- Mascaramento de informação de identificação do conjunto de dados que pode passar pela recodificação dos identificadores diretos;

5.2 Classificação de Métodos de Controlo de Divulgação

Segundo a literatura (Matthias, 2017), existem três grandes grupos de técnicas de controlo de divulgação:

- Técnicas não perturbativas – Não alteram os valores das variáveis de identificação e /ou sensíveis que consistem principalmente numa redução de detalhe da base de microdados;
- Técnicas perturbativas – Modificam os valores das variáveis sensíveis antes da sua publicação;
- Técnicas que geram dados sintéticos – dados gerados de forma aleatória que preservam certas estatísticas ou relações de acordo com os arquivos originais.

5.3 Métodos Não Perturbativos

5.3.1 Amostragem

A amostragem consiste em selecionar parte de uma população e observá-la com vista a estimar uma ou mais características para a totalidade da população. Esta técnica - não perturbativa- para além da divulgação de apenas uma parte dos dados, têm como objetivo alcançar a confidencialidade dos respondentes, através da redução do risco de divulgação.

Com a divulgação da respetiva amostra, os intrusos[3] não conseguem identificar os respondentes, mesmo sabendo quem são os indivíduos que deram origem aos dados, uma vez que não existe conhecimento por parte dos usuários quais os elementos incluídos e excluídos na respetiva amostra.

No entanto Reiter & Dreshler (2010), menciona que na presença de valores referentes às variáveis sensíveis, os intrusos podem fazer cruzamento de informações entre outros ficheiros de dados e conseguirem identificar os respetivos respondentes. E, portanto, desta forma é que a confidencialidade dos indivíduos não pode ser posta em causa, devendo-se realizar o tratamento dos valores sensíveis - como por exemplo, o método da reamostragem que se resume na substituição dos valores sensíveis das respetivas variáveis contínuas pela média amostral.

5.3.2 Recodificação Global

A recodificação global é um método não perturbativo, que pode ser aplicado a variáveis-chave de privacidade, tanto a variáveis categóricas como contínuas.

Para uma variável categórica, a ideia de recodificação consiste em combinar várias categorias numa, com uma contagem de frequência mais alta e com menos informação, uma vez que está agregada (Loukides & Lomax, 2021).

Para uma variável contínua podem ser usados os meios de recodificação das variáveis, de forma que estas mesmas sejam alteradas.

[3] Utilizador que abusa de dados divulgados e tenta divulgar informações sobre um indivíduo ou organização em particular

5.3.3 Limite superior e inferior

A codificação superior baseia-se na definição de um limite superior a *priori*, relativamente a todos os valores de uma variável. De seguida, é efetuada a substituição dos valores da variável, e que sejam superiores ao limite estabelecido pelo próprio limite.

No caso da codificação inferior, o cenário é análogo, mas, no entanto, com um limite inferior definido a *priori* (Matthias, 2017).

5.3.4 Supressão Local

A supressão local consiste na omissão de um ou mais valores sobre uma combinação insegura, isto é, que possa re-identificar o indivíduo ou entidade em questão. Este método, quando utilizado para proteção dos dados, os valores suprimidos passam a ter um valor em falta. O método é geralmente utilizado para variáveis categóricas, que, no entanto, também pode ser utilizado para variáveis contínuas (Hundepool, 2009).

5.3.5 Método de aleatorização a *posteriori* (PRAM)

Alguns métodos referidos a *priori*, como o método da recodificação global, a supressão global e codificação '*top and bottom'* podem conduzir a um grau de perda de informação elevado. O método PRAM (*Post Randomization Method*) é outra alternativa, uma vez que mantém a quantidade de detalhe. Esta técnica, consiste em classificar de forma errada algumas das variáveis categóricas, usando probabilidades de má classificação, e divulgar parte dos dados incorretamente especificados juntamente com essas probabilidades (Hout, 2006).

Segundo Hundepool (2009), o PRAM é um método de controlo de divulgação que pode ser aplicado em dados categóricos. É um método perturbativo e probabilístico que é utilizado para aumentar o grau de confidencialidade dos dados, e por sua vez proteger os ficheiros de microdados.

PRAM é um método que surgiu em 1997 e é definido em termos de probabilidades de transição, resumidas numa matriz PRAM (Wolf, 1998). Este produz ficheiros de

microdados, em que os valores de algumas variáveis categóricas para determinados registos são alterados em relação aos valores dos ficheiros de microdados originais. É aplicado normalmente às variáveis de identificação (Mendes, 2010).

5.3.6 Arredondamento

O método de arrendondamento, como o próprio o nome indica, consiste na substituição do valor das variáveis originais por valores arredondados. Estes valores, para uma determinada variável, são escolhidos através de um conjunto de valores arredondados (Domingo, 2001). Já numa base de dados original multivariada, o método também é possível ser usado (Waal, 2001). O princípio é idêntico ao anterior, só é necessário realizar a abordagem individualmente para cada variável.

5.4 Métodos Perturbativos

5.4.1 Rank Swapping

A troca de dados foi inicialmente introduzida como sendo um método de controlo da divulgação para variáveis categóricas. A ideia base é transformar um ficheiro de dados através de uma permutação de valores do mesmo atributo (dois registos diferentes). Esta abordagem mantém algumas características estatísticas dos dados, como a contagem e a frequência relativa dos atributos (Domingo, 2001). Para Reiss, *et al*. (1982), a troca de dados foi introduzida para proteger os microdados contínuos, por outro lado, Reiss (1984) refere que é utilizada para a proteção de microdados categóricos.

A hierarquia da troca é uma variante de troca de dados, sendo utilizada originalmente por variáveis ordinais (Greenberg, 1987), que também pode ser utilizada por variáveis numéricas (Moore, 1996). Os valores das variáveis são, classificados por ordem crescente. Depois cada valor ordenado é trocado aleatoriamente, dentro de um intervalo restrito, por cada outro valor ordenado. As estatísticas calculadas a partir deste algoritmo são menos 'distorcidas' do que as calculadas após uma troca livre.

5.4.2 Microagregação

A microagregação é uma técnica perturbativa aplicável a variáveis numéricas (Defays & Nanopoulos, 1993), no entanto existem variações que admitem outros tipos de variáveis.

Esta técnica, consiste na criação de *clusters* sendo estes os mais homogéneos possível de tamanho sendo igual ou superior a k (parâmetro de segurança), de forma que seja evitada uma perda de informação (Oganian & Domingo, 2001). A microagregação pode ser vista como um problema de clusterização, com limitações ao nível do tamanho dos *clusters* (Solanas, 2008). Ao conjunto de subconjuntos criados denomina-se como k-partição (Solanas & Martinez, 2021)

Segundo Solana (2021), a microagregação ótima, é definida como um processo que produz uma partição de ordem k, de forma que maximize a homogeneidade dentro dos *clusters*. A microagregação ótima requer abordagens heurísticas, uma vez que se trata de um problema *NP-hard* (Chettri, *et al*., 2012) para dados multivariados. As heurísticas de microagregação podem ser classificadas da seguinte forma:

- **Microagregação de <u>tamanho fixo</u>:** a partir do ficheiro original, os dados são agrupados por *clusters* de ordem k, exceto de um grupo que apresente o tamanho entre k e 2k-1, quando o número de registos não é divisível por k;
- **Microagregação de <u>tamanho variável</u>:** a partir do ficheiro original são agrupados *clusters* cuja ordem pertence ao intervalo (k, 2k-1).

Portanto, um processo de microagregação funciona com base em dois passos, assim, numa primeira fase o conjunto de dados originais é dividido em subconjuntos os mais homogéneos possíveis, com pelo menos k elementos (obtém-se assim a partição-k), sendo chamado como passo da partição. Numa segunda fase, o passo da agregação, começa por calcular os centroides (ou seja, o vetor médio) dos vários subconjuntos criados no passo anterior, e depois cada elemento é substituído pelo centroide do grupo a que pertence (Domingo, *et al*., 2006), criando assim um conjunto de dados k-anónimo, o que permite reduzir a perda de informação causada pela agregação.

A soma do quadrado dos erros (SSE) é utilizada normalmente para medir a homogeneidade em cada grupo. Como referido, o objetivo é maximizar a

homogeneidade, por isso, maximizar a homogeneidade dentro dos grupos é equivalente a encontrar uma k-partição que minimize o SSE (Solanas & Martinez, 2021).

O problema da microagregação consiste em encontrar uma partição-k com SSE mínimo. Segundo a literatura, há uma grande variedade de heurísticas para resolver o problema da microagregação multivariada, e muitas destas, são baseadas na teoria dos grafos. Um dos métodos mais conhecidos é a Distância Máxima do Vetor Médio (*Maximum Distance Average Vector* – MDAV), proposto por Domingo, *et al*., (2006). Este método cria de forma interativa *clusters* de k, elementos considerando sempre os registos mais distantes do centroide do conjunto de dados. Uma variante do MDAV foi proposta por Laszlo (2005), nomeadamente o método *Centroid-Based Fixed Size* (CBFS)(Laszlo & Mukherjee, 2005), que também otimizou versões baseadas na *kd-tree neighborhood search*, como KD-CBFS e KD-CBFSapp (Solé, 2021). Cheng *et al*., (2007), propôs o método dos dois pontos de referência fixos (*Two Fixed Reference Points – TFRP)*, que consiste em usar dois pontos extremos de um conjunto de dados em cada iteração como referência para criar os *clusters*. Yang *et al*., (2020) criou uma variante do algoritmo MDAV que utiliza as correlações entre os atributos para selecionar o ruído mínimo de forma a atingir o nível da privacidade desejado. Solanas *et al*., (2021) baseado no método MDAV introduziu *Variable Group-Size Heuristic based on* MDAV*.* No seu trabalho, os autores referem que este método tem o objetivo de diminuir as restrições, que normalmente estão presentes nos métodos como a cardinalidade da microagregação de tamanho fixo, onde permite que os *clusters* se adaptem melhor aos dados e levando a uma redução do SSE.

Laszlo & Mukherjee (2005) abordou o problema da microagregação através das árvores abrangentes mínimas. Lin *et al.,* (2010), propôs *Density- Based Algorithm (DBA)* que consiste na construção dos grupos de registos através da ordem decrescente da respetiva densidade, e de seguida, é realizado um ajuste desses grupos na ordem inversa. Mais abordagens que vão de encontro com a minimização do SSE bem como a eficiência do procedimento de microagregação podem ser encontrados em vários estudos segundo Panagiotakis (2011) & Ferrer (2006) *et al*.

Recentemente, Solanas *et al*. (2021) apresentou uma abordagem melhorada da microagregação multivariada. O trabalho dos autores é semelhante ao de Heaton & Mukherjee (2011) que utilizam heurísticas de otimização através do algoritmo do

caixeiro-viajante (Shmoys, et al.1985) de forma a conseguirem definir um caminho criado com as informações de um método de microagregação multivariado (por exemplo, MDAV, MD, CBFS). Solanas *et al*. (2021), também propõe a utilização de heurísticas de construção ao invés das de otimização. Dessa forma conseguem eliminar a necessidade de utilizar um método de microagregação multivariado como uma etapa de pré-processamento, e consequentemente, diminuem o tempo computacional sem prejudicar a utilidade dos dados.

5.4.3 Adição de Ruído

Adição de ruído, é um método perturbativo geralmente aplicado a variáveis contínuas. A ideia deste método consiste em adicionar um número estocástico ou aleatório a um ficheiro de dados.

Equação 1 Adição de ruído

Adição de Ruído Não Correlacionado

A adição de ruído não correlacionado (Hundepool, *et al*., 2007), a um ficheiro de dados pode ser expresso da seguinte forma:

$$\boldsymbol{Z}_i = \boldsymbol{X}_i + \varepsilon_i \quad , \qquad i = 1, \dots, n$$

Onde:

- $\boldsymbol{Z}_i$: representa o vector dos valores perturbados da variável i
- $\boldsymbol{X}_i$: representa o vector dos valores originais da variável i

Habitualmente, requer-se que o ruído, $\boldsymbol{\varepsilon_i}$, siga uma distribuição normal[4], cujo valor esperado é nulo e a variância é proporcional ao dos dados originais.

$$E[\varepsilon_i] = 0 \; e \; V[\varepsilon_i] = kV[X_i], \qquad onde \; k \in \mathbb{R}^+, \qquad i = 1, \dots, n$$

[4] É usual considerar que o ruído segue uma distribuição Gaussiana

Uma vez que é adicionado ruído não correlacionado, então:

Covariância entre os ruídos dos dados

$$Cov[\varepsilon_i, \varepsilon_j] = 0\,, \qquad \forall\, i \neq j$$

Covariância entre os valores originais do ficheiro e respetivos ruídos

$$Cov[X_i, \varepsilon_i] = 0\,, \qquad \forall\, i$$

Valor esperado dos Dados Perturbados

$$E[Z_i] = E[X_i + \varepsilon_i] = E[X_i] + E[\varepsilon_i] = E[X_i]$$

Variância dos Dados Perturbados

$$V[Z_i] = V[X_i + \varepsilon_i] = V[X_i] + 2Cov[X_i, \varepsilon_i] + V[\varepsilon_i] = V[X_i] + kV[X_i] =$$

$$= (1 + k)V[X_i] \text{ (1)}$$

Covariância dos Dados Perturbados:

$$Cov[Z_i, Z_j] = Cov[X_i + \varepsilon_i, X_j + \varepsilon_j] = Cov[X_i, X_j] + Cov[X_i, \varepsilon_j] + Cov[\varepsilon_i, X_j] +$$

$$+ Cov[\varepsilon_i, \varepsilon_j] \text{ (2)} \quad , \forall\, i \neq j$$

Como os ruídos são independentes entre si e os dados originais, tem-se de (Castro, J., 2012):

$$Cov[Z_i, Z_j] = Cov[X_i, X_j]$$

Correlação dos Dados Perturbados:

$$\rho_{Z_i Z_j} = \frac{Cov[Z_i, Z_j]}{\sqrt{V[Z_i]V[Z_j]}}$$

De (Huang & Williamson, 2001) e (Castro, 2012), temos:

$$\rho_{Z_i Z_j} = \frac{Cov[X_i, X_j]}{(1 + k)\sqrt{V[X_i]V[X_j]}} = \frac{1}{1 + k}\rho_{X_i X_j}$$

onde $\rho_{Z_i Z_j}$ é a correlação amostral referente ao ficheiro de dados que sofreu a perturbação, e $\rho_{X_i X_j}$ a correlação amostral ao ficheiro de dados original.

A partir destes resultados, constatamos que a adição de ruído não correlacionado preserva a média e as covariâncias.

Adição de Ruído Correlacionado

A metodologia de ruído correlacionado (Brand, 2002) é semelhante ao supramencionado, no entanto agora o ruído introduzido no ficheiro de dados originais apresenta outras características, é correlacionado. E, portanto:

$$Cov[\varepsilon_i, \varepsilon_j] = kCov[X_i, X_j], \qquad \forall\, i \neq j$$

Consideremos agora ε_i segue uma distribuição normal $N(0, k\Sigma)$, onde Σ é a matriz de covariância dos ruídos.

Valor esperado dos Dados Perturbados:

$$E[Z_i] = E[X_i + \varepsilon_i] = E[X_i] + E[\varepsilon_i] = E[X_i]$$

Covariância dos Dados Perturbados:

$$Cov[Z_i, Z_j] = Cov[X_i + \varepsilon_i, X_j + \varepsilon_j] = Cov[X_i, X_j] + Cov[X_i, \varepsilon_j] + Cov[\varepsilon_i, X_j] +$$

$$+ Cov[\varepsilon_i, \varepsilon_j] = Cov[X_i, X_j] + kCov[X_i, X_j] = (1 + k)Cov[X_i, X_j] \quad (3)$$

Variância dos Dados Perturbados:

$$V[Z_i] = V[X_i + \varepsilon_i] = V[X_i] + 2Cov[X_i, \varepsilon_i] + V[\varepsilon_i] = V[X_i] + kV[X_i] =$$

$$= (1 + k)V[X_i] \quad (4)$$

Correlação dos Dados Perturbados:

$$\rho_{Z_i, Z_j} = \frac{Cov[Z_i, Z_j]}{\sqrt{V[Z_i]V[Z_j]}}$$

De (Matthias, 2017) e (Wong & Fu, 2010) temos:

$$\rho_{Z_i, Z_j} = \frac{(1 + k)\, Cov[X_i, X_j]}{(1 + k)\sqrt{V[X_i]V[X_j]}} = \rho_{X_i, X_j}$$

Como se pode observar, em contraste com a adição de ruído não correlacionado que preserva a média e as covariâncias, este método perde a propriedade relativamente à

covariância, no entanto os dados perturbados apresentam a mesma estrutura de correlação dos dados originais (Domingo, & Torra, 2001), e, portanto, preserva a correlação.

É importante realçar que embora a metodologia supramencionada pareça simples, a adição de ruído deve ser realizada com algum cuidado, uma vez que os resultados dependem da quantidade de ruído atribuído e esta atribuição em demasia, embora leve a um risco de divulgação menor, apresenta uma perda de informação muito maior.

Estes dois tipos de adição de ruído ilustram claramente um dos grandes problemas da privacidade dos dados. É difícil transformar os dados de forma que estas propriedades estatísticas da qual estamos interessados sejam transitadas para o ficheiro transformado. Como vimos, no caso da adição de ruído não correlacionado, a média e a covariância são preservadas, mas já na adição de ruído correlacionado apenas a média é preservada. Desta forma, a pessoa que está a perturbar o ficheiro deverá ter em conta quais as propriedades que farão mais sentido preservar.

5.4.4 Ruído Multiplicativo

O ruído multiplicativo é um método perturbativo aplicado a variáveis contínuas. A ideia deste método consiste em multiplicar um número estocástico ou aleatório a um ficheiro de dados.

O ruído multiplicativo (Jay & William *et al*., 2003) a um ficheiro de dados pode ser expresso da seguinte forma:

$$\boldsymbol{Z}_i = \boldsymbol{X}_i \varepsilon_i$$

onde:

- $\boldsymbol{Z}_i$: representa o vector dos valores perturbados da variável i
- $\boldsymbol{X}_i$: representa o vector dos valores originais da variável i
- ε_i: representa o vector dos ruídos da variável i, que segue uma $N(1, \sigma)$

Jay & William *et al*. (2003) defendem que os valores gerados devem ter comportamento gaussiano, com média 1, mas, no entanto, a variância deve apresentar uma ordem de grandeza pequena.

Uma vez que os dados originais e os ruídos aleatórios são independentes, temos as seguintes propriedades:

Valor esperado dos dados perturbados:

$$E[Z_i] = E[X_i \varepsilon_i] = E[X_i]E[\varepsilon_i]$$

Variância dos dados perturbados:

$$V[Z_i] = E\left[{Z_i}^2\right] - E[Z_i]^2 = E[(X_i \varepsilon_i)^2] - E[X_i \varepsilon_i]^2 = E\left[{X_i}^2\right]E[{\varepsilon_i}^2] - (E[X_i]E[\varepsilon_i])^2$$

Hwang (1996) conjeturou que o uso do método ruído multiplicativo é mais adequado comparativamente com adição de ruído, quando os dados estão relacionados com variáveis económicas, como rendimentos , salários, *etc*.

5.4.5 Embaralhamento (*Shuffling*)

Esta técnica introduzida por Muralidhar *et al*., (2006) é semelhante à troca de dados, no entanto usa um modelo de regressão subjacente para determinar quais variáveis devem ser trocadas. O embaralhamento pode ser usado para variáveis contínuas. Este método tem a vantagem de manter as distribuições marginais nos dados "embaralhados". No entanto, este método requer uma classificação completa dos dados, pois em termos computacionais pode ser muito intenso dependendo da quantidade de dados que o usuário está a lidar.

No trabalho Muralidhar *et al*., (2006) explica de forma detalhada o processo. A ideia consiste na classificação dos indivíduos com base nas respetivas variáveis originais. De seguida, é ajustado um modelo de regressão com as variáveis a serem protegidas como variáveis dependentes e um conjunto de variáveis (regressores) que consigam explicar muito bem a variável, ou seja, com um grau de correlação elevado. Este modelo de regressão irá gerar valores para cada variável que necessita de ser protegida. Os valores

gerados são também classificados e cada valor original será correspondido ao valor gerado com a mesma classificação. Isto, de certo modo, indica que todos os valores originais estão no ficheiro embaralhado, simplesmente não se encontram no mesmo local.

5.5 Geração de Dados Sintéticos

A publicação de dados sintéticos é uma outra forma de proteger os dados contra a divulgação inapropriada. Estes tipos de técnicas baseiam-se, geralmente num modelo estatístico que é gerado a partir do conjunto de dados originais, de modo que, quando os dados reproduzidos, estes sejam o mais próximo possível da estrutura dos dados originais. São esses dados sintéticos que são disponibilizados ao usuário final, no nosso caso, ao investigador. A vantagem que esta técnica apresenta, tal como Hundepool *et al*. (2009) referem é que algumas das propriedades estatísticas ou relações presentes no ficheiro de dados original são preservadas. Contudo pode-se gerar também alguns valores que não fazem sentido no mundo real.

Muitas abordagens e ferramentas têm sido desenvolvidas para gerar dados sintéticos. Estas abordagens podem ser categorizadas em três grupos principais (Matthias Templ, 2017):

- Reconstrução sintética
- Otimização combinatória
- Geração baseada em modelos

Capítulo 6

6 Utilidade dos dados e Perda da informação

Após a apresentação de diversos métodos que podem ser aplicados consoante a sua natureza, deparamo-nos com questões que são cruciais para o indivíduo que está a realizar a anonimização do ficheiro.

- O risco de divulgação do ficheiro é alto? Baixo?
- Quão útil é o ficheiro anonimizado?
- Houve perda de informação?
- Qual é o panorama entre o risco de divulgação e a utilidade do ficheiro anonimizado?

Como referido anteriormente, o propósito da anonimização do ficheiro é que o mesmo apresente características que permita ao investigador realizar as suas análises, sem que exista um alto risco de re-identificação nem que ponha em causa a confidencialidade dos indivíduos ou entidades em questão. É, portanto, aqui que entra a necessidade da análise da utilidade dos microdados e da perda de informação (Figura 2).

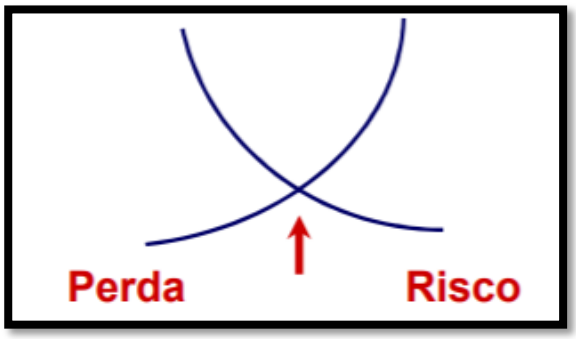


Figura 2 Ponto de equilíbrio entre Utilidade e Risco de divulgação

Como o leitor, já deve ter notado o cenário ideal após a aplicação dos métodos de controlo de divulgação é que o ficheiro de microdados apresente um ponto de equilíbrio, Figura 2, entre a sua utilidade e, com isto entenda-se as estatísticas dos dados originais serem preservadas no novo ficheiro e a perda de informação, de forma que ambas as partes saiam favorecidas.

Objetivo:

- Risco de identificação aceitável
- Perda de Informação mínima
- Informação do respondente segura

Karr *et al*., (2005) afirma que a qualidade da informação é avaliada pela capacidade dos mesmos serem utilizados de forma eficiente, económica e rápida para informar e avaliar no processo de tomada de decisão. Os mesmos autores definem a utilidade dos dados como a capacidade das propriedades estatísticas serem transitadas para os microdados divulgados.

De seguida, iremos ver indicadores que nos quantifiquem, de certo modo, a variação da nova base de dados.

6.1 Medidas de Utilidade

De modo a avaliar a perda de informação causada por um método de divulgação num conjunto de microdados, é necessário efetuar um *trade-off* entre o conjunto de dados original e o conjunto de dados modificado. Podemos dizer que houve pouca perda de informação e que a estrutura de ambas bases de dados for semelhante e vice-versa. De facto, a motivação de preservar a estrutura inerente no ficheiro de dados original é garantir que o ficheiro modificado seja válido analiticamente, e interessante no ponto que os investigadores, e não só, consigam realizar as análises pretendidas. Vejamos de seguida as medidas de utilidade para variáveis categóricas e continuas.

6.1.1 Variáveis Categóricas

Comparação de Valores omissos

Os valores omissos podem ser contados usando uma medida de utilidade simples que realiza a contagem dos dois valores omissos nos dados originais X e nos dados anonimizados Y. Sejam $R^{(X)}$ e $R^{(Y)}$ (indicadores) matrizes com a mesma dimensão que X e Y, respetivamente. Uma célula/elemento de $R^{(X)}$ é 1 quando X apresenta um valor omisso na exata posição, e 0 se o valor correspondente não está em falta. O raciocínio

é análogo para $R^{(Y)}$ e Y. Assim, $R^{(X)}$ e $R^{(Y)}$ são matrizes binárias, apresentando $0's$ e $1's$ consoante a posição dos valores omissos nas matrizes X e Y.

Seja R a matriz:

$$r_{ij} = \begin{cases} 0 \;\; se \;\; r_{ik}^{(X)} = r_{ik}^{(Y)} = 0 \\ 1 \; se \; r_{ik}^{(X)} = 1 \; \wedge \; r_{ik}^{(Y)} = 1 \\ 0 \; se \; r_{ik}^{(X)} = 0 \; \wedge \; r_{ik}^{(Y)} = 1 \\ 0 \; se \; r_{ik}^{(X)} = 1 \; \wedge \; r_{ik}^{(Y)} = 0 \end{cases}$$

Repare-se que a matriz original X, apresenta sempre uma quantidade de informação igual ou superior à matriz anonimizada Y.

Contagem de valores omissos adicionais

A variação dos valores omissos entre a matriz original e a matriz anonimizada é dada por:

$$m_j = \sum_{i}^{n} r_{ij} \;, \qquad j \in \{1, \dots, p\}$$

Assim, quanto maior for o valor de m_j, maior será a perda de informação no ficheiro de microdados anonimizado.

Comparar Informação agregada

Ao invés de compararmos diretamente os valores das variáveis categóricas, uma alternativa é comparar a tabela de contingência $T^{(X)}$, calculada a partir das variáveis categóricas da matriz original X, e a tabela de contingência $T^{(Y)}$ da matriz anonimizada Y. Uma maneira de avaliar a qualidade dos dados é através da soma das distâncias absolutas entre as respetivas células das tabelas de contingência.

Considere-se que $T^{(X)}$ e $T^{(Y)}$ apresentam n_1 linhas e n_2 colunas, e a mesma dimensão.

Temos então,

$$UT = \frac{1}{n_1 n_2} \sum_{i=1}^{n_1} \sum_{j=1}^{n_2} \left| T_{ij}^{(X)} - T_{ij}^{(Y)} \right|$$

O valor de UT é inversamente proporcional à qualidade presente nos dados anonimizados, ou seja, um valor alto de UT é indicador de uma fraca qualidade e vice-versa.

6.1.2 Variáveis contínuas

Winkler (1998) define que um ficheiro de microdados é analiticamente válido se:

1. As médias e as covariâncias num pequeno conjunto de subdomínios são aproximadamente preservadas;
2. Os valores marginais de pequenas tabulações de dados forem aproximadamente iguais;
3. Pelo menos uma característica da distribuição for preservada;

Winkler (1998) também refere, que um ficheiro de microdados é analiticamente interessante se seis variáveis de importantes subdomínios podem ser analisados devidamente.

Segundo Domingo & Mateo (2009), existem várias formas de averiguar que a estrutura dos dados originais é preservada. Uma das abordagens que autor refere é perante bases de microdados com variáveis contínuas, comparar estatísticas do ficheiro original e no ficheiro modificado é uma boa metodologia para ter um bom indicador da utilidade do ficheiro modificado.

Seja X o conjunto de dados original e X' o conjunto de dados modificado. Seja V e V' as matrizes variância-covariância das matrizes X e X' respetivamente. Analogamente, R e R' as matrizes de correlação. A Tabela 3, sumariza as medidas propostas por Domingo & Torra, (2001). Nessa tabela, p é o número de variáveis, n o número de registos e x_{ij} corresponde ao i-ésimo elemento da coluna j da respetiva matriz.

	Erro Quadrático Médio	Erro absoluto médio	Variação média
$\boldsymbol{X - X'}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}(x_{ij} - x'_{ij})^2}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\lvert x_{ij} - x'_{ij}\rvert}{np}$	$\frac{\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{\lvert x_{ij} - x'_{ij}\rvert}{\lvert x_{ij}\rvert}}{np}$
$\boldsymbol{\bar{X} - \bar{X}'}$	$\frac{\sum_{j=1}^{p}(\bar{x}_{ij} - \bar{x}'_{ij})^2}{p}$	$\frac{\sum_{j=1}^{p}\lvert \bar{x}_{ij} - \bar{x}'_{ij}\rvert}{p}$	$\frac{\sum_{j=1}^{p}\frac{\lvert \bar{x}_{ij} - \bar{x}'_{ij}\rvert}{\lvert \bar{x}_{ij}\rvert}}{p}$
$\boldsymbol{V - V'}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(v_{ij} - v'_{ij})^2}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert v_{ij} - v'_{ij}\rvert}{\frac{p(p+1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert v_{ij} - v'_{ij}\rvert}{\lvert v_{ij}\rvert}}{\frac{p(p+1)}{2}}$
$\boldsymbol{S - S'}$	$\frac{\sum_{j=1}^{p}(v_{jj} - v'_{jj})^2}{p}$	$\frac{\sum_{j=1}^{p}\lvert v_{jj} - v'_{jj}\rvert}{p}$	$\frac{\sum_{j=1}^{p}\frac{\lvert v_{jj} - v'_{jj}\rvert}{\lvert v_{jj}\rvert}}{p}$
$\boldsymbol{R - R'}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}(r_{ij} - r'_{ij})^2}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\lvert r_{ij} - r'_{ij}\rvert}{\frac{p(p-1)}{2}}$	$\frac{\sum_{j=1}^{p}\sum_{1\le i\le j}\frac{\lvert r_{ij} - r'_{ij}\rvert}{\lvert r_{ij}\rvert}}{\frac{p(p-1)}{2}}$

Tabela 3 Indicadores de qualidade

Quando estamos perante variáveis contínuas, a comparação dos dados originais e anonimizados requer a definição de distância. As medidas de perda de informação e de qualidade dos dados pode ser baseado na definição clássica (Euclidiana) ou robustas entre dados originais e perturbados. Uma métrica multivariada considerada robusta é a distância de Mahalanobis.

Distância de Mahalanobis

A distância entre os dados originais e os dados disponibilizados pode ser vista da seguinte forma:

$$MD(x_i) = [(x_i - U)^T C^{-1}(x_i - U)]^{\frac{1}{2}} \qquad i = 1, \dots, n$$

onde x_i é a observação, U e C os estimadores de localização e de covariância respetivamente. *(Matthias Templ., 2017)*

Nota: Para uma consideração fiável (estatísticas robustas), tanto C e T, estimadores de covariância e de localização respetivamente, devem ser estimados de forma robusta, e não da maneira tradicional, como por exemplo a média aritmética e covariância amostral. Estimativas robustas de localização e covariância podem ser obtidas através da Covariância Mínima Determinante – CMD (Rousseeuw & Dressen., 1998), (Matthias Templ., 2017).

A par da distância de *Mahalanobis* existem mais duas métricas que são geralmente usadas para mediar a perda de informação e a utilidade dos dados.

Considere-se, $X = \{x_{ij}\}$ o conjunto dos dados originais e $Y = \{y_{ij}\}$ o conjunto dos dados perturbados. Considere-se que ambos os conjuntos de dados apresentam n observações e p variáveis.

As medidas da perda de informação são dadas por:

$$IL1s = \frac{1}{pn}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{|x_{ij} - y_{ij}|}{0.5(|x_{ij}| + |y_{ij}|)}$$

Esta medida, foi proposta por Yancey *et al*., (2002), e pode ser interpretada como uma *scaled distance*, entre os valores originais e perturbados.

$$IL1 = \frac{1}{pn}\sum_{j=1}^{p}\sum_{i=1}^{n}\frac{|x_{ij} - y_{ij}|}{\sqrt{2}S_j}$$

Onde S_j é o desvio padrão da j-ésima variável do conjunto de dados original.

Qualidade de Previsão

Esta medida é obtida através da diferença entre os estimadores de um modelo de regressão definido *à priori* dos dados originais e dos dados perturbados.

$$\left|\frac{\hat{y}_w^o - \hat{y}_w^m}{\hat{y}_w^o}\right|$$

$\hat{y}_w^o$: Valores ajustados do modelo, relativamente ao conjunto de dados originais

$\hat{y}_w^m$: Valores ajustados do modelo, relativamente ao conjunto de dados perturbados;

w: pesos considerados quando se ajusta o modelo;

Estas duas últimas medidas estão presentes na biblioteca sdcMicro (será abordado no próximo capítulo).

A análise que é efetuada nas três metodologias é semelhante. O comportamento dos indicadores é inversamente proporcional à utilidade dos dados. Portanto, um valor alto do indicador implica uma baixa utilidade e vice-versa.

Capítulo 7

7 Medidas do Risco de Divulgação

As características que o ficheiro a ser divulgado possui são importantes. Como referido anteriormente, o método que otimiza a relação entre a perda de informação e o risco de divulgação deverá ser o escolhido. Desta forma, avaliar esta relação é um dos requerimentos do analista encarregue do ficheiro de microdados.

Segundo a literatura (J. Domingo-Ferrer & V. Torra., 2001) o risco de divulgação baseia-se basicamente nos métodos não perturbados, onde este risco é medido como uma probabilidade, isto é, uma taxa relativa entre a amostra e a respetiva população. Relativamente aos métodos perturbativos Domingo-Ferrer (2001), realça que a literatura não é vasta no sentido de haver metodologias de avaliação do risco para a classe dos métodos perturbativos. Adam & Wortmann (1989) dizem que as metodologias tendem a ser utilizadas de forma individual para cada método. No entanto, métodos empíricos tal como a ligação de registos (*record linkage*) proporcionam uma abordagem mais unificada, no sentido de poder ser aplicado na classe dos métodos perturbativos. Geralmente, são utilizadas três metodologias para a ligação de registos (Matthias Templ, 2017).

Ligação de registos baseado em distâncias:

Pagliuca & Seri (1999) descrevem esta abordagem no seu trabalho. Contudo, no seu artigo este algoritmo é aplicado num caso específico. O método utilizado foi o de microagregação e o algoritmo para o cálculo das distâncias foi o euclidiano. No entanto, o algoritmo de ligação de registos baseado em distâncias (*Distance – Based Record Linkage*) pode ser generalizado, e utilizado em qualquer método perturbativo. Geralmente, de forma a evitar problemas de escala, as distâncias são estandardizadas (Domingo-Ferrer, 2001).

Considere-se X o conjunto de dados original e Y o conjunto de dados perturbado.

Para cada registo em Y, é calculado as distâncias a todos registos em X e considera-se os valores do mais próximo e do segundo mais próximo. Suponhamos que identificamos x1

e x2 do conjunto de dados original como o primeiro e o segundo valor mais próximo de um elemento xi do conjunto de dados Y. Se x1 e xi correspondem ao mesmo respondente, então classificamos xi como "ligado". De forma semelhante, x2 e xi correspondem ao mesmo respondente então classificamos como "ligado ao 2º mais próximo". E procede-se de forma análoga para elemento do conjunto de dados Y. Finalmente, o risco de divulgação será definido como a percentagem de elementos de Y classificados como "ligado" ou "ligado ao 2º mais próximo". Este método, uma vez que é baseado em distâncias é computacionalmente intenso, e por isso, não é aconselhado para grandes bases de dados (Matthias Templ, 2017).

Ligação probabilística de registos :

Alternativamente, a ligação de registos probabilísticos (Jaro, 1989) consiste no emparelhamento dos registos do conjunto de dados original e o protegido, e utiliza um algoritmo que atribui um peso para cada par de dados, e que indica a probabilidade de os dois registos se referirem ao mesmo respondente. Pares que apresentem pesos a um limite especificado são identificados como "vinculados". À taxa relativa dos registos considerados como vinculados é o indicador do risco de divulgação.

Intervalo de divulgação:

A terceira medida de risco é chamada intervalo de divulgação (Pagliuca & Seri, 1999), uma vez que simplifica a ligação de registos baseada em distâncias. Este método é mais utilizado em grandes bases de dados. Nesta abordagem, depois dum método de divulgação ser aplicado ao ficheiro original, é construído um intervalo em torno de cada registo obtido no ficheiro disponibilizado. A amplitude do intervalo é baseada no valor que a variável toma ou no seu desvio padrão. Depois, é verificado se o valor original está dentro ou não do intervalo. A medida da divulgação do risco, é portanto, a proporção dos valores que estão dentro do intervalo. Como será de esperar, o pior cenário, é quando o indicador apresenta um valor de 100%.

Capítulo 8

8 Exemplos de ferramentas

$\mu - Argus$

Através de um programa da União Europeia, e devido à necessidade de anonimização e perturbação com o objetivo de proteger os ficheiros de microdados contra a divulgação, o *software* $Argus$ foi desensolvido. Em 1995, foi lançado a primeira versão do *software* pelo Departamento de Métodos Estatísticos na Holanda. (Hundepool & Willenborg, 1998) e ainda se encontra a ser aprimorado. A ferramenta $\mu - Argus$ foi originalmente desenvolvida em visual *basic* até à versão 4.2 e agora é escrito em *Java* e pode ser obtido de forma gratuita no site do CASC http://neon.vb.cbs.nl/casc/mu.htm (*Statistical Disclosure control Microdata*, 2017).

Código C++ da *International Household Survey Network*

O código c++ foi desenvolvido pela *International Household Survey Network* (IHSN) para a anonimização de microdados, com o fim de apoiar a disseminação segura dos dados. Além disso, o *software* foi desenvolvido de forma que fosse possível correr código estatístico, isto é, proveniente de diferentes *softwares* tais como STATA, SPSS e SAS. Enquanto o *software* desenvolvido a partir do IHSN é gratuito e de código aberto, o uso dos *softwares* referidos previamente requerem a posse de uma licença, uma vez que é restrito para uso comercial. O código IHSN é totalmente integrado (e melhorado) no sdcMicro. (*Statistical Disclosure Control Microdata,* 2017)

sdcMicro e *sdcMicroGUI*

A biblioteca *sdcMicroGUI* (Kowarik et al. 2013) fornece uma interface gráfica do utilizador para *sdcMicro* e serve como uma ferramenta fácil de manusear e altamente interativa para os utilizadores que desejam usar a biblioteca *sdcMicro* para controlo de divulgação, que não são familiarizados com a *interface* da linha de comando do R. A GUI executa de forma automatizada os cálculos e exibe a informação de forma sucinta as

contagens de frequência, medidas de risco individuais e globais, perda de informações e utilidade dos dados após cada etapa de anonimato. Mudanças efetuadas, como no risco e medidas de utilidade nos dados originais também são exibidas de forma convenientemente na interface gráfica do utilizador (GUI) e o código é salvo num *script*, que pode ser facilmente exportado, modificado e reutilizado, possibilitando a reprodução de quaisquer resultados. Recentemente, uma nova reimplementação do *sdcMicroGUI* foi feita utilizando o *shiny*[5] (Chang *et al*., 2006).

8.1 Diferenças entre as Ferramentas

A Tabela 4 dá uma visão geral do *software* e dos métodos disponíveis de três produtos de *software*: o *software* $\mu - Argus$ (Hundepool *et al*, 2008) do Departamento de Estatística da Holanda, os pacotes do R *sdcMicro* (Templ *et al*., 2015) e *sdcMicroGui* (Kowarik *et al*., 2013) e p código C++ que foi escrito pelo IHSN.

Relativamente ao $\mu - Argus$ e *sdcMicro*, as diferenças não se prende apenas com quantidade de métodos como pode ser observado na Tabela 4. As principais vantagens do *sdcMicro* é o seu aplicativo que é orientado a objetos e sua forma amigável de ser lidado, a facilidade de importar dados (em $\mu - Argus$ um *script* de importação que determina a estrutura hierárquica dos dados deve ser escrita) e sua flexibilidade de uso. Note-se também, que $\mu - Argus$ não apresenta a geração automática de código, nem apresenta os recálculos automáticos. Além disso, muitos métodos não estão presentes em $\mu - Argus$ como por exemplo SUDA, *shuffling* ou estimativas de risco baseadas em modelos.

[5] *Shiny*: É uma biblioteca do *software* R que facilita a construção de aplicações *web* interativas diretamente a partir deste *software*.

Métodos|Medidas		$\mu - Argus$	Ferramentas BioMed	sdcMicro	sdcMicroGui	IHSN
Estimativas de Risco	Contagens de Frequências	✓	✓	✓	✓	✓
	Risco individual (RI)	✓		✓	✓	✓
	RI nas famílias	✓		✓	✓	✓
	L-diversidade		✓	✓	✓	✓
	SUDA2			✓		✓
	Risco global (RG)	✓		✓	✓	✓
	RG com mod log-lin			✓		
Métodos	Recodificação	✓	✓	✓	✓	✓
	Supressão Local	✓		✓	✓	✓
	Rank Swapping	✓		✓		✓

Tabela 4: Métodos e medidas presentes nas ferramentas

Relativamente à velocidade de computação, não podemos comparar *sdcMicro* e $\mu - Argus$, no entanto *Templ* constata que $\mu - Argus$ não é adequado para grandes quantidades de dados, pois se torna muito lento e fica sem memória muito rápido mesmo presente com uma base de dados média, no entanto se recorrermos ao *sdcMicro* tal não acontece uma vez que o processamento da informação é realizado em segundos independentemente da base de dados utilizado (*Statistical Disclosure Control*, 2017).

Métodos|Medidas		$\mu - Argus$	Ferramentas BioMed	sdcMicro	sdcMicroGui	IHSN
Métodos	PRAM	✓		✓	✓	✓
	Adição de ruído correlacionado			✓	✓	✓
	Microagregação	✓		✓	✓	✓
	Shuffling			✓	✓	
Utilidade	Medidas de utilidade	✓	✓	✓	✓	
Características	GUI	✓	✓		✓	
	CLI			✓		✓
	Relatório	✓		✓	✓	
	Plataforma independente		✓	✓	✓	✓
	Grátis e código aberto		✓	✓	✓	✓

Tabela 5: Métodos e medidas presentes nas ferramentas

8.2 Vantagens *sdcMicro*

A biblioteca *sdcMicro* (Templ *et al.*, 2005) inclui todos os métodos SDC apresentados. É gratuito, de código aberto e encontra-se disponível na abrangente rede de arquivos R (CRAN). Esta biblioteca implementa métodos de divulgação estatística populares para estimativa de risco, como o algoritmo suda2, a abordagem de risco individual ou medição de risco usando o modelo *log-linear*. Além disso, métodos perturbativos como recodificação global, supressão local, pós-randominização, microagregação, adição de ruído correlacionado, embaralhamento e outros métodos estão integrados (Tabela 4 e 5). Com a biblioteca *sdcMicro*, o controlo da divulgação pode ser realizado de forma exploratória, interativa e de uma forma muito amigável. Todos os resultados são

guardados de uma forma estruturada e os mesmos são atualizados assim que um método é aplicado. Métodos de impressão e de sumarização permitem resumir o nível do risco de divulgação e da utilidade dos dados, e relatórios podem ser gerados de forma automatizada. Além disso, a maioria dos métodos aplicados podem ser realizados com a *interface* gráfica do usuário (GUI) *sdcMicroGUI* (Kowarik *et al.*, 2013) sem recorrer ao *software* R ou caso seja pretendido através do *software* R utilizando a função *sdcApp* da biblioteca *sdcMicro*. A mais recente versão corre num *browser* e é baseado no *shiny*. (Chang *et al*., 2016). Uma biblioteca de *software* com um conceito semelhante ao do *sdcMicro* é a biblioteca *simPop* (Templ *et al*., 2017) – é utilizado para gerar dados sintéticos (*Statistical Disclosure Control Microdata*, 2017).

Desta forma, no presente estudo será utilizado a ferramenta *sdcMicro*.

Capítulo 9

9 Controlo de Divulgação - *sdcMicro*

9.1 Informações gerais sobre *sdcMicro*

A primeira versão, 1.00, da biblioteca *sdcMicro* foi lançada em 2007 na rede abrangente de arquivos R e foi introduzido por Templ (2008). No entanto, esta versão incluía apenas alguns métodos e a biblioteca consistia numa pequena coleção de funções que eram aplicáveis a pequenos conjuntos de dados. A versão atual 5.6.0, é um grande passo à frente comparativamente com a inicial. Assim, os métodos são implementados de forma orientada a objetos (usando classes S4) e têm sido escritos internamente através de implementações em C++ ou pela biblioteca *data.table* (Dowle *et al*., 2013). Estes últimos factos, permitem que sejam efetuados cálculos com alto desempenho.

O IHSN forneceu o código C++ para muitos métodos que foram reescritos do zero (exceto *suda2* e *rankswap*) e integrado ao *sdcMicro* (Matthias Templ., 2017).

9.2 Estrutura da classe S4 da biblioteca *sdcMicro*

Esta secção é baseada principalmente em Templ *et al*., (2005) que apresenta a implementação de *sdcMicro* de forma detalhada. Os seguintes pontos fornecem uma visão geral sobre o principal objetivo de *sdcMicro*:

- *SdcMicro* inclui conjunto de método de proteção de microdados bastante compreensiva;
- Apresenta uma implementação de uma classe S4 bem definida o que proporciona uma implementação amigável e facilita o uso das funcionalidades do *sdcMicroGUI*;
- As funções de utilidade extraem informações de objetos de classe S4 bem definidos;
- Depois de aplicado o método, as células das bases de dados são atualizadas automaticamente;
- Por motivos de desempenho, os métodos são implementados internamente em C++ ou através da biblioteca *data.table* (Dowle *et al*., 2013);

- Relatórios dinâmicos dos resultados do processo de anonimato podem ser gerados.

A ideia é gerar um objeto em R que contém informações das variáveis que pretendemos proteger. Para tal é necessário informar o R, quais as variáveis que pretendemos avaliar de forma que informações adicionais que são estimadas a partir dos dados estejam armazenadas de forma adequada. Para realizar o processo supramencionado, devemos criar o objeto através da função *createSdcObj.*

Os parâmetros da função *createSdcObj* são, por exemplo, as variáveis chave categóricas ou contínuas, o vetor de pesos amostrais e opcionalmente estratificação e os ID's de *cluster*.

O seguinte código, apresentado nas Figuras 3 e 4, mostra como podemos gerar um objeto a partir de dados de teste de uma pesquisa das Filipinas (base de dados contida na biblioteca *sdcMicro*).

```
require("sdcMicro")
data("testdata", package="sdcMicro")
sdc <- createSdcObj(testdata,
          keyVars=c('urbrur','water','sex','age'),
          numVars=c('expend','income','savings'),
          pramVars=c("walls"),
          w='sampling_weight',
          hhId='ori_hid')
```

Figura 3: Criação de um objeto sdcMicro

```
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"
```

Figura 4: Atributos do método sdc

9.3 Métodos *sdcMicro*

Na Tabela 6, podemos constatar os métodos que se encontram disponíveis e as respetivas funções da biblioteca *sdcMicro*.

Método	Classificação	Natureza dos dados	Função no sdcMicro
Recodificação Global	Não Perturbativo	Contínua e categórica	globalRecode , groupVars
Limite superior e inferior	Não Perturbativo	Contínua e categórica	topBotCoding
Supressão Local	Não Perturbativo	Categórica	localSuppression, localSupp
PRAM	Perturbativo	Categórica	Pram
Micro-agregação	Perturbativo	Contínua	microaggregation
Adição de Ruído	Perturbativo	Contínua	addNoise
Shuffling	Perturbativo	Contínua	Shuffle
Rank swapping	Perturbativo	Contínua	rankSwap

Tabela 6 Métodos SDC micro

9.4 Aplicação de métodos

Na presente secção, iremos aplicar os métodos de controlo de divulgação referidos na Tabela 6. Embora, o presente estudo se baseie essencialmente com os métodos perturbativos, serão apresentados exemplos ilustrativos das famílias dos métodos perturbativos, bem como, não perturbativos para ficheiros de microdados.

9.4.1 Métodos não perturbativos

Recodificação Global

O método de recodificação global é um método que pode ser aplicado a variáveis continuas ou categóricas. Sendo aplicado a uma variável categórica, a ideia deste método consiste no agrupamento das categorias, de forma que as mesmas apresentem uma frequência absoluta mais elevada e, consequentemente, um grau de re-identificação mais baixo. Caso a variável seja contínua, uma abordagem será classificá-la por níveis, categorizando-a, e procedendo do mesmo modo supramencionado.

Considere-se o seguinte exemplo apresentado na Figura 5, onde a base de dados considerada é a *testdata* incorporada na libraria *sdcMicro*. Selecionou-se a variável *age* (idade) para recodificar.

```
require("sdcMicro")
data(testdata, package="sdcMicro")
sdc <- createSdcObj(testdata,
                    keyVars=c("urbrur","water","sex","age","relat"),
                    numVars=c("expend","income","savings"),
                    pramVars=c("walls"),
                    w="sampling_weight",
                    hhId="ori_hid")

table(testdata$age)

## 0   1   2   3   4   5   6   7   8   9  10  11  12  13  14
## 98  90 134 112 128 133 136 125 144 126 151 127 108 143 103
## 15  16  17  18  19  20  21  22  23  24  25  26  27  28  29
##111 106 101  96  64  88  61  64  55  55  69  56  68  69  50
## 30  31  32  33  34  35  36  37  38  39  40  41  42  43  44
## 90  51  72  49  59  86  61  68  51  42  67  43  44  49  40
## 45  46  47  48  49  50  51  52  53  54  55  56  57  58  59
## 65  37  31  43  28  44  28  28   8  19  31  28  17  27  20
## 60  61  62  63  64  65  66  67  68  69  70  71  72  73  74
## 36  24  28  14  12  40   8  16  14   8  20   6  14   4   6
## 75  76  77  78  79  80  82  83  84  85  88  90  95
##  6   4   5   3   2   5   1   1   1   1   1   2   1
```

Figura 5: Criação de objeto sdcMicro e tabela de frequências absolutas

A variável *age*, apresenta muitos valores distintos observados. Note-se que os valores mais elevados apresentam poucas ocorrências, portanto, de forma a diminuir o risco de re-identificação, será aumentado o número de observações por categoria, por outras palavras, iremos considerar a variável *age* por classes (Figura 6).

```
labs <- c("1-9","10-19","20-29","30-39",
          "40-49","50-59","60-69","70-79","80-100")

sdc <- globalRecode(sdc, column=c("age"),
                    breaks=c(0,9,19,29,39,49,59,69,79,100),
                    labels= labs)

print(sdc)

##Infos on 2/3-Anonymity:

## Number of observations violating
##     - 2-anonymity: 111 (2.424%) | in original data: 653 (14.258%)
##     - 3-anonymity: 184 (4.017%) | in original data: 1087 (23.734%)
##     - 5-anonymity: 345 (7.533%) | in original data: 1781 (38.886%)
```

Figura 6: Método Recodificação global

Repare-se que ainda existem algumas observações que estão a violar o k-anonimato (número de combinações necessárias entre as variáveis chave para identificar um atributo de um indivíduo/entidade). Uma possível causa, poderá dever-se à dimensão das observações de algumas classes. Veja-se então, na Figura 7, o tamanho das respetivas classes:

```
table(sdc@manipKeyVars$age)

##  1-9  10-19  20-29  30-39  40-49  50-59  60-69  70-79 80-100
## 1128   1110    635    629    447    250    200     70     13
```

Figura 7: Tabela de frequências (classes)

```
table(sdc1@manipKeyVars$age)

##  1-9  10-19  20-29  30-39  40-49  50-59 60-100
## 1128   1110    635    629    447    250    283


print(sdc1)
##Number of observations violating
##     - 2-anonymity: 85 (1.856%) | in original data: 653 (14.258%)
##     - 3-anonymity: 162 (3.537%) | in original data: 1087 (23.734%)
##     - 5-anonymity: 294 (6.419%) | in original data: 1781 (38.886%)
```

Figura 8: Tabela de frequências (redução de classes) e output do método de recodificação global

As últimas duas classes, apresentam um número de observações muito reduzido em comparação com as restantes. Assim sendo, agrupemos as últimas três classes que estão representadas na Figura 8.

Note-se que houve um decréscimo percentual relativamente à transgressão do k-anonimato, e posto isto, o risco de re-identificação baixou. Observe-se também que o procedimento de agrupamento poderia ser efetuado novamente, no entanto teríamos que ter um cuidado acrescido com o grau de utilidade obtido.

Limite Superior e Inferior

Um caso especial da recodificação global, é o método limite superior e inferior. O *sdcMicro* não possibilita a aplicação dos dois métodos em simultâneo. E, por isso, deve-se executar o limite superior separadamente do limite inferior. Este método, também pode ser executado através da função anterior *globalRecode().*

Considere-se que dispomos de um conjunto de idades de indivíduos e pretendemos aplicar o método do limite superior e inferior. Como podemos ver no exemplo apresentado na Figura 9, a primeira função diz respeito à aplicação do limite superior. O que esta função irá fazer, é considerar a variável idade, e todos os valores que apresentem uma idade superior a 65, os seus respetivos valores irão ser substituídos pela idade 65.

```
#Limite Superior: Idade 65
sdctop<-topBotCoding(obj = sdc, value = 65, replacement = 65,
                     kind = 'top', column = 'age')

#Limite Inferior: Idade 5
sdcbottom <- topBotCoding(obj = sdcInitial, value = 5, replacement = 5,
                          kind = 'bottom', column = 'age')
```

Figura 9: Método limite superior e limite inferior

No entanto, a segunda função, já diz respeito ao limite inferior. A função, irá considerar, novamente a variável idade, e todos os valores que apresentem uma idade inferior a 5, os seus respetivos valores serão substituídos pela idade 5.

Embora o exemplo apresentado, seja ilustrativo para entendermos a aplicação do método no *sdcMicro* bem como as suas funcionalidades, é importante destacar, que numa situação onde a presença de *outliers* na variável de interesse seja relevante, este método não perturbativo reduz a sensibilidade dos valores da variável.

Supressão Local

A biblioteca *sdcMicro*, dispõe de duas funções para aplicar a supressão local: *localSupression()* e *localSup().* A primeira é geralmente utilizada, uma vez que permite a supressão de valores de específicas variáveis indiretas, de modo que seja possível alcançar o menor número de cruzamentos possíveis (k -anonimato). A função (Figura 10) por defeito irá suprimir em maior escala valores de variáveis que apresentem um maior número de categorias, e em menor os que contenham poucas categorias.

Vejamos então a aplicação do método:

```
sdcSup <- localSuppression(sdc, k = 5)
print(sdcSup, 'ls')

##Local suppression (applied per strata given by variable(s) hhcivil)

##  KeyVar | Suppressions (#) | Suppressions (%)
##  urbrur |                2 |            0.044
##   water |               44 |            0.961
##     sex |                4 |            0.087
##     age |             1756 |           38.341
##   relat |              144 |            3.144
```

Figura 10: Método Supressão local

Primeiro, definiu-se o número de cruzamentos possíveis. De forma ilustrativa escolheu-se um valor para k=5. Observando o *output*, podemos constatar que existiu um número de supressões mais elevado na variável *age.* Tal como tínhamos referido, o método de supressão local no *sdcMicro* analisa as categorias das respetivas variáveis e suprime, caso necessário, em maior escala, as variáveis que apresentam mais categorias. Neste exemplo ilustrativo, a variável *age* é a que apresenta um maior número de categorias.

9.4.2 Métodos Perturbativos

PRAM

PRAM (*Post Randomization Method*) é um método perturbativo para dados categóricos. Este método é baseado na teoria de *Markov*, e baseia-se numa matriz de transição P, que contém as respetivas probabilidades de transição, isto é, a probabilidade de uma certa variável permanecer inalterada ou ser alterada com qualquer um dos K-1 valores. K é o número de categorias ou níveis dentro da variável que será aplicada o PRAM.

Considere-se a variável *region* que apresenta três categorias: *capital, rural1* e *rural2.* A matriz de transição que será aplicada no método PRAM tem a dimensão 3x3:

$$P = \begin{bmatrix} 1 & 0 & 0 \\ 0.05 & 0.8 & 0.15 \\ 0.15 & 0.15 & 0.8 \end{bmatrix}$$

Os valores da diagonal da matriz quadrada são as probabilidades de a respetiva categoria não ser alterada. Por exemplo a entrada (1,1) da matriz P, significa que todos os indivíduos da categoria *Capital* irão permanecer na categoria *Capital*. Na entrada (2,2) o valor 0.8 é a probabilidade dos indivíduos permanecerem na categoria *rural1*, mas no entanto as restantes entradas da mesma linha correspondem a uma mudança para categoria *capital* com probabilidade 0.05 e *rural2* com probabilidade 0.15. O raciocínio é análogo para as restantes entradas.

No *sdcMicro*, a função *pram()* permite que o método PRAM seja executado. Esta função, permite que o método seja aplicado a um subconjunto de microdados de forma independente. Para tal, é necessário especificar a variável estratificada definida nos subconjuntos. No caso de na função, não estar especificado a variável estratificada, a função aplica o PRAM a todas as variáveis.

Considere-se agora, a aplicação prática do método PRAM:

Comecemos por criar o objeto da classe *sdcMicro*, tendo em conta que as variáveis categóricas devem ser indicadas como *fator*. Indicou-se uma variável estratificada, e de seguida procedeu-se à aplicação do PRAM a uma variável (Figura 11).

```
require("sdcMicro")
data(testdata, package="sdcMicro")

# Variáveis categóricas devem ser consideradas como factor
vars <- c("urbrur", "water", "sex", "age", "relat", "walls", "roof")
testdata[, vars] <- lapply(testdata[, vars], as.factor)

sdc <- createSdcObj(testdata,
                    keyVars = c("urbrur", "water", "sex", "age", "relat"),
                    numVars = c("expend", "income", "savings"),
                    pramVars = c("walls"),
                    w = "sampling_weight",
                    hhId = "ori_hid",
                    strataVar = "hhcivil" )
```

Figura 11: Criação de objeto sdcMicro com variável pram

Como se pode observar na Figura 12, o output do PRAM , apresenta-me o estado final da matriz de transição, com as probabilidades de transição especificadas. Por fim, é apresentado o número e a taxa percentual de alteração das observações da variável que foi alvo do método PRAM.

```
sdc <- pram(sdc)
print(sdc, "pram")

##Post-Randomization (PRAM):

##  Variable:walls

##--> final Transition-Matrix:

##   2          3           9
## 2 0.94110562 0.05660232 0.002292054
## 3 0.02046666 0.97475127 0.004782064
## 9 0.05514683 0.31819857 0.626654599

#Changed observations:

#  variable nrChanges percChanges
#1    walls       180         3.93
```

Figura 12: Método PRAM

Microagregação

A função da biblioteca *sdcMicro, microaggregation(),* pode ser utilizada para a microagregação univariada. O argumento *aggr* especifica o tamanho do grupo. Esta função requer que todos os grupos tenham o mesmo tamanho. Por defeito, o tamanho do grupo *aggr* assume uma dimensão de 3. A escolha do tamanho do grupo depende da homogeneidade entre grupos (SSE), e do nível de proteção pretendida. Geralmente, verifica-se que um valor do tamanho do grupo é diretamente proporcional ao nível de proteção. Com isto, um maior valor definido previamente irá levar a um nível de proteção mais elevado. No entanto, existe uma grande desvantagem. Quanto maior for o tamanho do grupo, maior será a perda de informação.

A função de microagregação (Figura 13 e 14), por defeito, substituiu os valores do grupo utilizando como medida de localização, a média. No entanto, uma medida alternativa pode ser utilizada, a mediana. Geralmente, este último é mais útil quando na nossa

variável apresenta uma quantidade considerável de *outliers*. Vejamos as seguintes aplicações dos métodos de microagregação univariada, utilizando como medida a média e a mediana respetivamente (Tabela 7).

```
sdc1<-microaggregation(obj = sdc, variables = c("income"),
                       aggr = 3, measure = "mean")
```

Figura 13: Método de microagregação univariada, tendo como medida de grupo a média

```
sdc2<-microaggregation(obj = sdc, variables = c("income"),
                       aggr = 3, measure = "median")
```

Figura 14: Método de microagregação uunivariada, tendo como medida de grupo a mediana

Grupo	Income	Microagregação (Média)	Microagregação (Mediana)
1	2300	2245	2300
2	2434	3608	2434
1	2123	2245	2300
1	2312	2245	2300
2	6045	3608	2434
2	2345	3608	2434

Tabela 7: Primeiras 6 observações dos valores originais, e dos métodos de microagregação aplicados

No caso de mais de que uma variável ser candidata à microagregação, uma forma de atuar é aplicar a microagregação univariada a cada uma, ou aplicar a microagregação multivariada.

É importante referir, que caso seja pretendido aplicar uma microagregação multivariada, é recomendado observar as matrizes de covariância e de correlação, pois geralmente ocorre menos perda de informação quando as variáveis são fortemente correlacionadas.

Veja-se na Tabela 8, os métodos disponíveis de microagregação multivariada existentes, e depois a aplicação de um método multivariado.

Método	Descrição
Mdav	Agrupamento é baseado na distância euclidiana
rmd	Agrupamento é baseada na distância multivariada de Mahalanobis
pca	Agrupamento é baseado na análise de componentes principais
clustpppca	Agrupamento é baseado nos clusters e na análise de componentes principais de cada cluster
influence	Agrupamento é baseado nos clusters e a agregação é realizada dentro os clusters

Tabela 8: Descrição das medidas dos métodos de microagregação multivariado

```
sdc1<-microaggregation(obj = sdc, variables = c("expend", "income", "savings"),
                       method = "mdav", measure = "mean")
```

Figura 15: Microagregação multivariada, usando como medida Mdav

Adição de Ruído

No *sdcMicro* a adição de ruído é implementada através da função *addNoise() (Figura 16)*. Tal como referido anteriormente, este método baseia-se na adição de um ruído gaussiano. No entanto, este ruído pode ser ou não correlacionado.

Vejamos primeiro, a aplicação do método adição de ruído não correlacionado, aplicando um ruído na ordem de 0.5:

```
sdcNoise <- addNoise(obj = sdc,
                     variables = c("expend", "income", "savings"),
                     noise = 0.5, method = "additive")
```

Figura 16:Adição de ruído não correlacionado com ordem de ruído 0.5

Observemos na Tabela 9, as primeiras 6 observações das respetivas variáveis numéricas antes e depois da aplicação do método:

Obs.	Valores Originais			Valores Perturbados		
	Expend	Income	Savings	Expend	Income	Savings
1	90929693	57800000	116258.5	90880976	57850316	106889.9
2	27338058	25300000	279345.0	27658859	25284777	324503.8
3	26524717	69200000	5495381.0	26349608	68970171	5505219.0
4	18073948	79600000	8695862.0	17900879	79755149	8692465.5
5	6713247	90300000	203620.2	6879964	90286864	189558.3
6	49057636	32900000	1021268.0	49189644	32960837	1049186.2

Tabela 9: Primeiras 6 observações dos valores originais e após a aplicação da adição de ruído

Observando os valores das variáveis após a perturbação, verifica-se que sofreram alterações, uma vez que foi aplicado um ruído não correlacionado na ordem de 0.5.

Dependendo do objetivo, o ruído alocado às variáveis também pode ser correlacionado. A função *addNoise()* dispõe de dois métodos de aplicação de ruído correlacionado. O método *'correlated'* e o método *'correlated2'*. O primeiro assume que as variáveis são aproximadamente gaussianas e a segundo assume que as variáveis não são gaussianas. Portanto, antes de aplicar um dos métodos, o usuário deve investigar se os dados apresentam um comportamento gaussiano ou não. Para tal, pode-se utilizar o teste de *Shapiro - Wilk*, que está presente na libraria *stats* ou o teste de *Jarque-Bera* que está presente na libraria *tseries* no R (Shapiro., (1965); Jarque (1987).

Uma vez que o presente capítulo, é meramente ilustrativo, iremos apenas mostrar a aplicação dos métodos *'correlated' e correlated2'*. Vejamos então, a aplicação dos métodos na Figura 17:

```
sdcNoise1 <- addNoise(obj = sdc,
                      variables = c("expend", "income", "savings"),
                      noise = 0.5, method = "correlated")

sdcNoise2 <- addNoise(obj = sdc,
                      variables = c("expend", "income", "savings"),
                      noise = 0.5, method = "correlated2")
```

Figura 17: Aplicação da adição de ruído correlacionado assumindo que os dados são e não são gaussianos

Rank Swapping

O método *Rank Swapping* é implementado no *sdcMicro* através da função *rankSwap()*. As variáveis que serão trocadas devem ser especificadas no argumento '*variables*'. Por defeito, o método aplica o método do limite superior e inferior. Este, considera como limite superior e inferior o 95º percentil e o 5º percentil respetivamente, e são substituídos pelo valor esperado da variável. Através dos argumentos '*TopPercent*' e '*BottomPercent*' podemos escolher os percentis. Uma vez que o algoritmo é executado de forma aleatória, é recomendado a utilização da semente (*seed()*) antes da aplicação do método, de modo a garantir os resultados produzidos. As variáveis que apresentem valores omissos, o método também procede à recodificação dos mesmos tornando-os *NA.*

Vejamos então como aplicar o método na Figura 18:

```
set.seed(12345)
rankSwap(sdc, variables = c("expend", "income"), missing = NA)
```

Figura 18: Método Rank swapping

É importante referir que este método não é útil, quando as variáveis apresentam poucos valores diferentes, ou bastantes valores omissos, uma vez que a troca nesse caso não vai resultar num valor alterado.

Embaralhamento

O método do embaralhamento é semelhante ao *rank swapping*, no entanto utiliza um modelo de regressão para determinar as variáveis que serão trocadas. Este método é implementado no *sdcMicro* através da função *shuffle().* Este método, através do modelo de regressão, gera um conjunto de valores, que por sua vez, ocuparão o lugar dos valores da variável do ficheiro original que apresentem a mesma classificação. Por defeito, a função aplica o método *'ds' ,* que consiste na aplicação do método *EGADP* (em inglês, *Enhanced Data Perturbation*) nos percentis. Segundo Matthias Templ. (2017), é aconselhado a utilização do método referido, uma vez que apresenta um melhor *trade-off* entre a divulgação e a utilidade dos dados. No argumento '*form*' deve ser especificado as variáveis que serão consideradas as variáveis regressoras do modelo. Contudo, este método, exige que sejam apresentadas pelo menos duas variáveis

dependentes e que haja um poder explicativo significativo entre as variáveis explicativas e independentes. Por outras palavras, o coeficiente de indeterminação R^2, deve ser alto.

Vejamos então, como aplicar o método na Figura 19:

```
sdcShuf <- shuffle(sdc, method=c('ds'),regmethod= c('lm'), covmethod=c('spearman'),
              form=savings+expend ~ urbrur+walls)
```

Figura 19: Método Embarallhamento

Capítulo 10

10 Estudo de caso: Base de dados CRC

A base de dados presente é referente à Central de Responsabilidades de Crédito(CRC). É anonimizada e disponibilizada pelo Laboratório de Investigação em Microdados (BPLIM). Esta, reporta informação acerca do crédito do fornecedor de todas as instituições de crédito em Portugal a partir de 1999. Os dados são obtidos mensalmente, com o objetivo de apoiar os participantes na avaliação de risco de concessão de crédito. Este conjunto de dados é atualizado anualmente. O ficheiro de dados fornecido apresenta dados desde 2010 até 2018. No entanto, uma vez que o propósito é comparação das técnicas de controlo de divulgação, apenas será utilizado o ficheiro referente ao ano de 2017. É importante referir, que a base de dados não corresponde à base de dados real, pois foi randomizada, isto é, foram aplicados métodos de divulgação para que não se tenha acesso aos dados confidenciais. Deste modo, informa-se que as estatísticas apresentadas em diante, são apenas para fins científicos, uma vez que o propósito do estudo é a comparação entre as bases de dados modificadas e a original (randomizada).

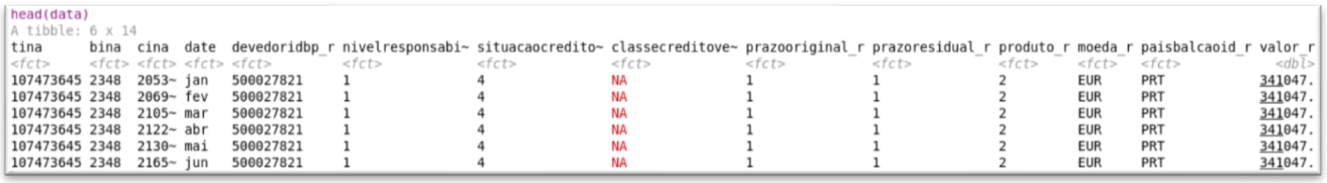
10.1 Análise Descritiva dos dados

Começou-se por realizar uma análise preliminar dos dados. Verificou-se que o conjunto de dados tem 13166441 observações referentes a 330233 indivíduos/entidades e apresenta um total de 14 variáveis (Figura 20).

```
library(haven)
CRC_R_MFRMEXP_2017_APR19_BAL_V01 <- read_stata("/bplimext/projects/p083_JoaoRego/initial_dataset/CRC_R_MFRMEXP_2017_APR19_BAL_V01.dta")
data<-CRC_R_MFRMEXP_2017_APR19_BAL_V01

#Dimensão dos dados
dim(data)
# 13166441       14

#Nomes das variaveis
colnames(data)
# "tina"                    "bina"                    "cina"                    "date"                    "devedoridbp_r"
#  "nivelresponsabilidade_r" "situacaocredito_r"       "classecreditovencido_r"  "prazooriginal_r"         "prazoresidual_r"
#  "produto_r"               "moeda_r"                 "paisbalcaoid_r"          "valor_r"

# tina: Número de identificação fiscal
length(unique(data$tina))
# 330 233
```

Figura 20: Base de dados CRC e respetivas variáveis

No entanto, deparou-se que todas as variáveis apresentavam uma natureza de variável numérica, enquanto a sua descrição tal não indicava. Por isso, foram modificadas para variáveis categóricas. Vejamos o formato da base de dados na Figura 21:



```
head(data)
A tibble: 6 x 14
tina      bina  cina  date  devedoridbp_r nivelresponsabi~ situacaocredito~ classecreditove~ prazooriginal_r prazoresidual_r produto_r moeda_r paisbalcaoid_r valor_r
<fct>     <fct> <fct> <fct> <fct>         <fct>            <fct>            <fct>            <fct>           <fct>           <fct>     <fct>    <fct>          <dbl>
107473645 2348  2053~ jan   500027821     1                4                NA               1               1               2         EUR      PRT            341047.
107473645 2348  2069~ fev   500027821     1                4                NA               1               1               2         EUR      PRT            341047.
107473645 2348  2105~ mar   500027821     1                4                NA               1               1               2         EUR      PRT            341047.
107473645 2348  2122~ abr   500027821     1                4                NA               1               1               2         EUR      PRT            341047.
107473645 2348  2130~ mai   500027821     1                4                NA               1               1               2         EUR      PRT            341047.
107473645 2348  2165~ jun   500027821     1                4                NA               1               1               2         EUR      PRT            341047.
```

Figura 21: Primeiras 6 observações da base de dados CRC

E agora, as variáveis que se encontram enumeradas e descritas na Tabela 10[6]:

Variável	**Natureza**	**Descrição**
Tina	Categórica	Número de Identificação fiscal anonimizado
Bina	Categórica	Número de identificação anonimizado do Banco
Cina	Categórica	Número de identificação anonimizado do crédito
Date	Categórica	Data de referência (Mês/ Ano)
Devedoridbp_r	Categórica	Código de identificação do devedor
Nivelresponsabilidade_r	Categórica	Nível de responsabilidade
Situacaocredito_r	Categórica	Situação do crédito
Classecreditovencido_r	Categórica	Classe de crédito vencido
Prazooriginal_r	Categórica	Maturidade do crédito definido no momento do acordo contratual
Prazoresidual_r	Categórica	Intervalo de tempo entre a data de referência e o momento do acordo contratual
Produto_r	Categórica	Produto Financeiro
Paisbalcao_r	Categórica	Código do país de residência do devedor
Moeda_r	Categórica	Código da moeda do país de residência do devedor
Valor_r	Contínua	Valor da dívida

Tabela 10: Descrição das variáveis

[6] As variáveis estão descritas em mais detalhe no anexo.

10.2 Análise univariada das variáveis

Variáveis contínuas:

A Tabela 11, apresenta as medidas de localização relativamente à variável contínua *valor_r*, onde se pode constatar o valor mínimo, valor máximo, média, bem como os 3 quartis.

Medidas de localização	
Valor mínimo	25.007
Valor máximo	3212487650
Média	140190.4
Mediana	10306.92
1º Quartil	2278.144
2ºQuartil	10306.02
3ºQuartil	36982.4
Valores omissos	0

Tabela 11 Medidas de localização

Note-se que 75% dos valores dos créditos se encontram abaixo de 36982,4 euros, enquanto o valor máximo é 3212487650. Claramente que este valor representa um *outlier* severo.

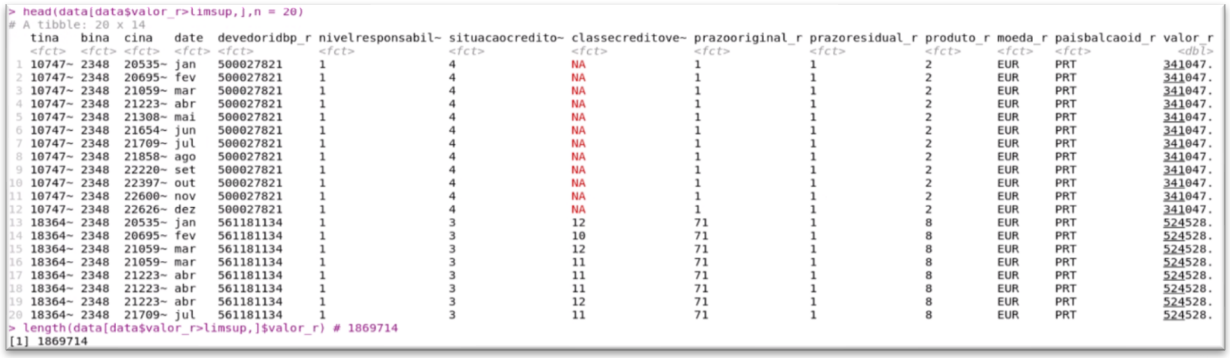
Considere-se a um *outlier*:

- a é moderado: Se $a \notin [Q_1 - 1.5AIQ; Q_3 + 1.5AIQ]$
- a é severo: Se $a \notin [Q_1 - 3AIQ; Q_3 + 3AIQ]$

É importante avaliar que mais créditos são *outliers*, isto, porque existindo, significa que são créditos com elevado risco de identificação.

Para um crédito apresentar um valor de dívida que seja considerado um *outlier* moderado tem que ser superior a 89038.78 e inferior a -49778.24. Já seria de esperar, pelas medidas de localização observadas que a presença dos *outliers*, seria do lado do limite superior, até porque a distribuição é assimétrica à direita e principalmente pela distância do valor máximo e do 3ºQuartil ser consideravelmente grande.

No *output* apresentado na Figura 22, é visível 20 observações que são considerados pelo menos *outliers* moderados. Repare-se também, que no *output* é apresentada a contagem dos *outliers* moderados, 1869714. E, portanto, estes créditos apresentam um elevado risco de divulgação.



```
> head(data[data$valor_r>limsup,],n = 20)
# A tibble: 20 x 14
   tina   bina  cina   date  devedoridbp_r nivelresponsabil~ situacaocredito~ classecreditove~ prazooriginal_r prazoresidual_r produto_r moeda_r paisbalcaoid_r valor_r
   <fct>  <fct> <fct>  <fct> <fct>         <fct>             <fct>            <fct>            <fct>           <fct>           <fct>    <fct>   <fct>          <dbl>
 1 10747~ 2348  20535~ jan   500027821     1                 4                NA               1               1               2        EUR     PRT            341047.
 2 10747~ 2348  20695~ fev   500027821     1                 4                NA               1               1               2        EUR     PRT            341047.
 3 10747~ 2348  21059~ mar   500027821     1                 4                NA               1               1               2        EUR     PRT            341047.
 4 10747~ 2348  21223~ abr   500027821     1                 4                NA               1               1               2        EUR     PRT            341047.
 5 10747~ 2348  21308~ mai   500027821     1                 4                NA               1               1               2        EUR     PRT            341047.
 6 10747~ 2348  21654~ jun   500027821     1                 4                NA               1               1               2        EUR     PRT            341047.
 7 10747~ 2348  21709~ jul   500027821     1                 4                NA               1               1               2        EUR     PRT            341047.
 8 10747~ 2348  21858~ ago   500027821     1                 4                NA               1               1               2        EUR     PRT            341047.
 9 10747~ 2348  22220~ set   500027821     1                 4                NA               1               1               2        EUR     PRT            341047.
10 10747~ 2348  22397~ out   500027821     1                 4                NA               1               1               2        EUR     PRT            341047.
11 10747~ 2348  22600~ nov   500027821     1                 4                NA               1               1               2        EUR     PRT            341047.
12 10747~ 2348  22626~ dez   500027821     1                 4                NA               1               1               2        EUR     PRT            341047.
13 18364~ 2348  20535~ jan   561181134     1                 3                12               71              1               8        EUR     PRT            524528.
14 18364~ 2348  20695~ fev   561181134     1                 3                10               71              1               8        EUR     PRT            524528.
15 18364~ 2348  21059~ mar   561181134     1                 3                12               71              1               8        EUR     PRT            524528.
16 18364~ 2348  21059~ mar   561181134     1                 3                11               71              1               8        EUR     PRT            524528.
17 18364~ 2348  21223~ abr   561181134     1                 3                11               71              1               8        EUR     PRT            524528.
18 18364~ 2348  21223~ abr   561181134     1                 3                11               71              1               8        EUR     PRT            524528.
19 18364~ 2348  21223~ abr   561181134     1                 3                12               71              1               8        EUR     PRT            524528.
20 18364~ 2348  21709~ jul   561181134     1                 3                11               71              1               8        EUR     PRT            524528.
> length(data[data$valor_r>limsup,]$valor_r) # 1869714
[1] 1869714
```

Figura 22: Observação de 20 outliers moderados e a quantidade de outliers moderados existentes

A Tabela 12, apresenta as medidas de dispersão relativamente à variável *valor_r*:

Medidas de dispersão	
Variância	$7{,}5\times 10^{12}$
Desvio padrão	2740545

Tabela 12: Medidas de dispersão

Estes resultados, vão de encontro ao verificado anteriormente, uma vez que os valores se encontram bem dispersos relativamente à média.

10.3 Aplicação dos métodos de controlo de divulgação

Como foi visto previamente, o ficheiro de dados apresenta indicações que não está seguro, portanto a divulgação do ficheiro de microdados não deve ocorrer no seu formato original. Deste modo, a necessidade da proteção da confidencialidade, leva a que sejam aplicados os métodos de controlo de divulgação.

Uma vez que este estudo, têm como foco a aplicação de métodos perturbativos, o mesmo será feito sendo direcionado para variáveis contínuas. Assim sendo, serão aplicados os métodos de adição de ruído, microagregação, *rank swapping e shuffling*.

Comecemos por criar um objeto da classe *sdcMicro* (Figura 23):

```
#criar o objeto sdcMicro
library(sdcMicro)
kvars<-c('date','situacaocredito_r','prazooriginal_r',
         'prazoresidual_r')
nvars<-c('valor_r')

sdc<- createSdcObj(data,
                   keyVars = kvars,
                   numVars = nvars)
```

Figura 23: Criação de objeto sdcMicro

10.3.1 Adição de ruído não correlacionado

Como foi referido anteriormente, a adição de ruído pode ser aplicada maneiras diferentes. Começou-se por aplicar a adição de ruido não correlacionado normalmente distribuído. No entanto, como as estatísticas têm que ser preservadas aplicou-se ruído aditivo na ordem de 0.1,0.5,1,2 e 5 (veja-se na Figura 24 e Tabela 13).

```
sdc1<-addNoise(sdc, variables = nvars, noise= 0.1,method = 'additive')
sdc2<-addNoise(sdc, variables = nvars, noise= 0.5,method = 'additive')
sdc3<-addNoise(sdc, variables = nvars, noise= 1,method = 'additive')
sdc4<-addNoise(sdc, variables = nvars, noise= 2,method = 'additive')
sdc5<-addNoise(sdc, variables = nvars, noise= 5,method = 'additive')
```

Figura 24: Adição de ruído não correlacionado com as ordens de ruído 0.1, 0.5,1,2 e 5

	Valores originais	Ruído 0.1	Ruído 0.5	Ruído 1	Ruído 2	Ruído 5
Média	140190.4	140192.4	140192.4	140191.8	140182.2	140251.2
Variância	7.510586×10^{12}	7.510597×10^{12}	7.510778×10^{12}	7.11396×10^{12}	7.513708×10^{12}	7.528914×10^{12}
IL1		265 840 711 5	132 902 162 55	265 693 382 88	531 528 303 17	132 972 116 434

Tabela 13: Média, variância e IL1 para valores originais e adição de ruído não correlacionado

Uma vez que o ruído apresenta uma média nula, seria de esperar que a média permanecesse inalterada. Por sua vez o método de adição de ruído na ordem de 0.1 detêm a variância mais próxima dos dados originais, contudo o método de adição de ruído da ordem de 0.5 apresenta um indicador de IL1 mais baixo, e, portanto, apresenta uma menor perda de informação comparativamente aos outros.

10.3.2 Adição de ruído correlacionado

Após a aplicação de ruído não correlacionado, aplicou-se o método de adição de ruído correlacionado, no entanto, antes de partir para a aplicação do método verificou-se se a variável contínua apresentava um comportamento gaussiano ou não. Para isso, começou-se por aplicar o teste de *Shapiro - Wilk*, mas devido à dimensão da variável não foi possível obter resultados. Deste modo, decidiu-se aplicar o teste de *Jarque-Bera* (Shapiro (1965); Jarque (1987) (Figura 25).

Relembremos que o teste de *Jarque-Bera* tem como hipótese nula que os dados seguem uma distribuição normal, e a hipótese alternativa o contrário Jarque (1987).

```
library(tseries)
jarque.bera.test(data$valor r)
#Jarque Bera Test

#data:  data$valor_r
#X-squared = 7.3791e+16, df = 2, p-value < 2.2e-16
```

Figura 25: Teste Jarque-Bera

Como p-valor é aproximadamente zero, então rejeitamos a hipótese nula e podemos afirmar com um grau de confiança de 99% que os dados não apresentam um comportamento gaussiano.

Portanto, apliquemos o método de adição de ruído correlacionado (veja-se a Figura 26 e Tabela 14), que é robusto contra a assunção da normalidade dos dados, para as ordens de ruído correlacionado 0.1,0.5,1,2 e 5:

```
sdc6<-addNoise(sdc, variables = nvars, noise= 0.1,method = 'correlated2')
sdc7<-addNoise(sdc, variables = nvars, noise= 0.5,method = 'correlated2')
sdc8<-addNoise(sdc, variables = nvars, noise= 1,method = 'correlated2')
sdc9<-addNoise(sdc, variables = nvars, noise= 2,method = 'correlated2')
sdc10<-addNoise(sdc, variables = nvars, noise= 5,method = 'correlated2')
```

Figura 26 Adição de ruído correlacionado para as ordens de ruído 0.1, 0.5, 1, 2 e 5

	Valores originais	Ruído 0.1	Ruído 0.5	Ruído 1	Ruído 2	Ruído 5
Média	140190.4	140221.1	140167.3	140189.2	140182.5	140215.7
Variância	7.510586× 10^{12}	7.510288× 10^{12}	7.511415× 10^{12}	7.511252× 10^{12}	7.510621× 10^{12}	7.510288× 10^{12}
IL1		265910770022	265830675640	2.65947× 10^{11}	2.65731× 10^{11}	266208759234

Tabela 14: Média, variância e IL1 para dados originais e adição de ruído correlacionado

10.3.3 Microagregação

Uma vez que estamos na presença de apenas uma variável numérica, iremos aplicar a microagregação univariada. No entanto, uma vez que previamente verificamos que os montantes da dívida dos créditos apresentam um número considerável de *outliers*, iremos realizar a microagregação univariada não apenas como medida do grupo a média, mas também a mediana.

Aplique-se os métodos de microagregação univariada, tendo como medida de grupo a média e a mediana (Figura 27).

```
sdc11<-microaggregation(obj= sdc, variables = 'valor_r', aggr = 3, measure = 'mean')
sdc12<-microaggregation(obj= sdc, variables = 'valor_r', aggr = 3, measure = 'median')
```

Figura 27: Microagregação univariada, tendo como medida a média e a mediana

	Valores Originais	Micro (média)	Micro (mediana)
Média	140190.4	1350815	1348244
Variância	7.510586× 10^{12}	7.296201× 10^{12}	6.280631× 10^{12}
IL1		1662.911	1330.001

Tabela 15: Média, Variância e IL1 para dados originais e microagregações univariadas aplicadas

Observando a Tabela 15, podemos constatar que o método de microagregação tendo como medida de grupo a média, apresenta um valor esperado mais próximo da dos valores originais. No entanto, repare-se que o outro método apresenta, e, como seria de esperar, um menor valor da variabilidade dos dados relativamente à média, bem como um indicador de perda de informação menor. Note-se que, a melhoria presente

nestes últimos resultados mencionados, deve-se ao facto da estrutura da variável numérica apresentar um número acentuado de *outliers*, o que ainda reforça mais o uso da mediana, pois esta medida de localização é muito menos sensível a *outliers* do que a média.

10.3.4 Comparação dos métodos

A aplicação dos diferentes métodos de controlo de divulgação, leva à criação de novos ficheiros de microdados. No entanto, é necessário entender a partir desta amostra de métodos utilizados, qual destes se aplica melhor à nossa base de dados.

Para isso, serão utilizados os indicadores de qualidade: erro quadrático médio, erro absoluto médio, IL1 e IL1s (Tabela 16).

Métodos			**Erro Quadrático Médio**	**Erro Absoluto Médio**	**IL1**	**IL1s**
Adição Ruído Não correlacionado	Ruído 0.1	Sdc1	7512083	2.019110	2658460771	7429.483
	Ruído 0.5	Sdc2	187839003	10.1003	13298497972	37145.53
	Ruído 1	Sdc3	750476091	20.18321	26574106473	74255.68
	Ruído 2	Sdc4	3004016413	40.36957	53152351896	148575.6
	Ruído 5	Sdc5	18782980035	100.9757	132949083707	371489.1
Adição Ruído correlacionado	Ruído 0.1	Sdc6	75292828417	201.961	265910770022	743318.3
	Ruído 0.5	Sdc7	75325880070	201.9002	265830675649	743355.1
	Ruído 1	Sdc8	75288291138	201.9885	2.65947×10^{11}	743193.4
	Ruído 2	Sdc9	75292008790	201.8245	2.65731×10^{11}	743171.3
	Ruído 5	Sdc10	75306020095	202.1873	266208759234	743324.3
Microagregação	Média	Sdc11	214384625077	1.262992×10^{-6}	1662.911	1364.738
	Mediana	sdc12	340778693512	1.010145×10^{-6}	1330.001	1134.208

Tabela 16: Indicadores de qualidade

Como podemos observar na Tabela 16 apresentada, se considerarmos o erro quadrático médio como medida de qualidade, podemos afirmar que o método sdc1 é o método que origina menor perda de informação, enquanto nos restantes o método sdc12 é o eleito. Por outro lado, o método sdc7 é o pior utilizando as medidas do erro quadrático

médio e IL1s. No erro absoluto médio o método com o pior indicador é o sdc8 e no IL1 é o sdc9.

Capítulo 11

11 Conclusão

Cada vez mais o sigilo da informação é priorizado. A imposição de novas leis no que toca à confidencialidade dos dados, enaltece ainda mais a importância de assegurar o fluxo da informação. Como é sabido, o BPLIM, nos últimos anos tem agido de modo que os investigadores externos, nacionais e estrangeiros possam aceder a informação sobre bases de dados granulares, contruídas com informação ao nível das entidades, sobre a economia portuguesa, com total salvaguarda da confidencialidade da informação.

O estudo realizado incide nesse sentido, com a aplicação e comparação de métodos de controlo de divulgação para fins de investigação numa base de microdados disponibilizada pelo BPLIM randomizada, sendo o foco a comparação dos métodos perturbativos nomeadamente nas variáveis contínuas, entender quais os ficheiros de dados produzidos se encontram mais próximo do ficheiro original.

Após o levantamento das metodologias e respetivos indicadores para avaliar as mesmas para técnicas perturbativas e não perturbativas, analisou-se um conjunto de ferramentas disponíveis para tratar os microdados. Concluiu-se que o *software* R apresentava mais vantagens, uma vez que dispõe de uma biblioteca denominada *sdcMicro* que contém a maioria dos métodos desenvolvidos, perturbativos e não perturbativos, e variantes dos mesmos. Apresenta também uma estrutura com que é fácil de lidar, uma vez que é possível aplicar os métodos de modo simples e rápido.

Seguiu-se com um tutorial resumido da biblioteca *sdcMicro*. Nessa secção abordou-se informações gerais acerca da biblioteca, a estrutura que apresenta (classe S4), a criação de um objeto *sdcMicro* bem como os métodos existentes na biblioteca. De seguida, realizou-se uma ilustração de aplicação para cada um dos métodos, referindo as suas propriedades e variantes.

Terminado a aplicação dos métodos, prosseguiu-se para a parte prática dos dados. Aqui, começou-se por analisar as variáveis existentes na base de microdados randomizada, bem como a sua descrição. Realizou-se uma análise descritiva à base de dados onde se constatou que a base de dados contém 13166441 observações referentes a 330233 créditos. Seguiu-se uma análise univariada à variável de interesse *valor_r*, onde se

constatou que não havia a presença de valores omissos. Verificou-se também que esta variável apresenta uma distribuição assimétrica à direita, onde apresenta que um conjunto de 1868714 montantes da dívida dos créditos são *outliers* moderados.

É importante referir, que os resultados das análises estatísticas demonstrados ao longo deste projeto são meramente para fins investigacionais, uma vez que a base de dados disponibilizada é randominizada, e, portanto, já sofreu alterações relativamente aos dados reais.

Depois, aplicou-se os métodos de controlo de divulgação disponíveis na biblioteca *sdcMicro*. Uma vez que o propósito do estudo, incide na aplicação de métodos de controlo de divulgação nas variáveis contínuas, e como a base de dados contém apenas uma variável contínua, não se aplicou métodos multivariados.

Começou-se por aplicar o método da adição de ruído não correlacionado, donde se concluiu com um nível de 99% de confiança que os dados que dizem respeito às dívidas dos créditos não apresentam um comportamento gaussiano - para 5 diferentes ordens de ruído, nomeadamente 0.1, 0.5,1,2 e 5, e aplicou-se também a microagregação univariada, tendo como medida de grupo a média e a mediana.

Em cada método fez-se o balanço entre os valores originais e os perturbados em todos os novos ficheiros, apresentado a esperança matemática, a variância e um indicador de utilidade IL1.

Por fim realizou-se a comparação dos métodos, através de indicadores de qualidade, nomeadamente o erro quadrático médio, o erro absoluto médio, o IL1 e IL1s. Concluiu-se que nesta amostra de métodos o que apresenta uma menor perda de informação e risco, usando o erro quadrático médio como indicador, é o método sdc1 – adição de ruído não correlacionado com ordem de ruído 0.1 - e os restantes indicadores elegem o método sdc12 – microagregação tendo como medida de grupo a mediana - como o melhor dos considerados.

11.1 Trabalhos Futuros

Um assunto que poderá ser alvo de um trabalho futuro é aplicação de métodos multidimensionais, uma vez que neste trabalho tal não foi possível. Também seria muito interessante verificar como seria o *trade-off* entre o risco de divulgação do ficheiro e a utilidade do mesmo, com a introdução de um fator temporal, dados em painel.

12 Glossário

Anonimização: Uso de técnicas que convertem os dados confidenciais em dados anónimos/ remoção ou mascaramento de informações de informações de identificação de conjuntos de dados.

Ativo primário: Processo ou dados em que a disponibilidade, integridade e confidencialidade tem de ser protegida.

Ativo digital: Ficheiro, base de dados ou outra forma de organização que contenha informação em formato digital, incluindo a informação pessoal de um ou mais titulares de dados.

Ativo documental: Documento em papel ou outro meio que contenha informação, incluindo a informação pessoal de um ou mais titulares de dados.

Arquivo uso científico: Tipo de arquivo de microdados, que está disponível apenas para investigadores selecionados sob contrato.

Adição ruído: Método de anonimização baseado na adição ou multiplicação de um valor estocástico ou número aleatório com os valores originais para proteger os dados de correspondência exata com arquivos externos. É normalmente aplicada para variáveis contínuas.

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

Consentimento do titular dos dados: qualquer manifestação de vontade livre, específica e informada, nos termos da qual o titular aceita que os seus dados sejam objeto de tratamento.

Conservação a longo prazo: Ato de manter a informação numa forma independente e acessível a longo prazo.

Dados confidenciais: Os dados que permitem a identificação de um indivíduo ou organização, direta ou indiretamente.

Dados não validados nem tratados: Os dados que não têm qualquer validação na fonte nem são tratados estatisticamente pela entidade responsável pela sua divulgação.

Dados não tratados: Os dados antes da aplicação de métodos de controlo de divulgação estatístico (SDC). Também designado de dados originais "*raw data*" ou "original data".

Dados Pessoais: qualquer informação, de qualquer natureza e independentemente do respetivo suporte, incluindo som e imagem, relativa a uma pessoa singular identificada ou identificável ('titular dos dados'); é considerada identificável a pessoa que possa ser identificada direta ou indiretamente, designadamente por referência a um número de identificação ou a um ou mais elementos específicos da sua identidade física, fisiológica, psíquica, económica, cultural ou social.

Divulgação (*Disclosure):* ocorre quando uma pessoa ou organização reconhece ou aprende algo que desconhecia, relativamente a uma pessoa ou organização, através dos dados disponibilizados.

Divulgação da identidade: A divulgação (*disclosure*) da identidade ocorre se um intruso associa um indivíduo conhecido ou organização com o/os registos de dados disponibilizados.

Ficheiro: um conjunto estruturado de dados pessoais acessíveis segundo critérios específicos, centralizado, descentralizado ou repartido de modo funcional ou geográfico.

Identificador: Um identificador é uma variável/ informações que podem ser usadas para estabelecer a identidade de um individuo ou organização. Os identificadores podem levar à identificação direta ou indireta de uma pessoa ou organização.

Identificador direto: Um identificador direto é uma variável que revela diretamente e inequivocamente a identidade de um cidadão. Exemplos: nomes, números de identidade social, números de documentos de identificação fiscal, Turmas associados a escolas.

Identificador indireto: Um identificador indireto é uma variável que a partir da conjugação ou relação com outras variáveis, permite conhecer a identidade de um indivíduo ou organização.

Integridade: A integridade de um documento de arquivo, refere-se a este permanecer completo e inalterado. É necessário que os documentos sejam protegidos contra

alterações não autorizadas. As políticas de gestão de arquivo devem especificar que tipo de adições ou anotações podem ser feitas a um documento depois da sua produção, em que circunstâncias essas alterações podem ser autorizadas, e quem está autorizado a fazê-las. Qualquer alteração autorizada a um documento de arquivo deve ser explicitamente indicada e reconhecível enquanto tal.

Intruso: Um utilizador que abusa de dados divulgados e tenta divulgar informações sobre um indivíduo ou organização em particular, utilizando um conjunto de características divulgadas de forma anonimizada.

Limiar (*threshold*): Um nível pré-estabelecido, valor, margem ou ponto em que os valores que estiverem acima ou abaixo deste podem considerar para efeitos de anonimato e confidencialidade os dados seguros e não seguros. Se forem considerados não seguras ações adicionais devem ser tomadas no sentido de reduzir o risco de identificação do respondente.

Métodos determinísticos: Métodos de anonimização que seguem um determinado algoritmo e produzem os mesmos resultados se aplicado várias vezes para os mesmos dados com o mesmo conjunto de parâmetros.

Microdados: Um conjunto de registos contendo informações sobre os respondentes individuais, entidades económicas ou unidades orgânicas. Tais registos podem conter respostas a um questionário, formulários administrativos ou resultarem de um envio sistemático de dados.

Pseudominização: o tratamento de dados pessoais para que deixei de poder ser atribuídos a um titular de dados específicos sem recurso a informações suplementares, desde que estas sejam mantidas separadamente e sujeitas a medidas técnicas e organizativas para assegurar que os dados pessoais não possam ser atribuídos a uma pessoa singular identificada ou identificável.

Quase-identificadores: Um conjunto de variáveis que, em combinação, podem ser ligados a informações externas para poder identificar os inquiridos no conjunto de dados disponibilizados. Quase-identificadores também são chamados de “variáveis-chave”. Exemplo de variáveis chave com elevado grau de probabilidade de identificação

do respondente são a combinação de código postal a 7 dígitos, data de nascimento e sexo.

Supressão: Supressão de dados envolve a não divulgação de informações, por ser considerado inseguro quando são aplicadas regras de confidencialidade aos dados.

***Statistical Disclosure Control* (SDC):** Técnica estatística de controlo da divulgação (acrónimo em inglês (SDC) pode ser definida como um conjunto de métodos, para reduzir o risco de informações que identifique indivíduos, empresas ou outras organizações. Tais métodos são apenas relacionados com a etapa de divulgação e normalmente são baseados em restrições aplicados à quantidade de informação a disponibilizar ou a modificar.

Variável: Qualquer característica, número ou quantidade que pode ser observada, ou contabilizada para cada unidade de observação.

Variáveis -Chave: Um conjunto de variáveis que, em combinação, podem ser ligados a informações externas para permitir identificar os inquiridos no conjunto de dados disponibilizados. Variáveis-chave são também chamados de “quase-identificadores”.

Variáveis sensíveis: Variáveis sensíveis são aquelas cujos valores não devem ser descobertos/ respondente por quaisquer entrevistados no conjunto de dados. 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.

13 Bibliografia

Brand, R. (2002). *Microdata protection through noise addition. Lecture Notes in Computer Science London: Springer*;

Castro, J. (2012). "*Recent advances in optimization techniques for statistical tabular data protection*"; European Journal of Operational Research, 2012;

Chang, C.; Li, Y.; Huang,W. (2007), TFRP: *An efficient microaggregation algorithm for statistical disclosure control*. J. Syst. Softw.;

Chettri, S; Paul, B.; Dutta, A., (2012); "*A comparative study on microaggregation techniques for microdata protection*"; International Journal of Data Mining Knowledge Management Process; Vol2, Nº6; 2012.;

Defays, D., and Nanopoulos, P., (1993). *Panels of enterprises and confidentiality: the small aggregates method*. In Proc. of 92 Symposium on Design and Analysis of Longitudinal Surveys, Ottawa, 1993.;

DGEEC (2021), Direção Geral De Estatísticas da Educação e Ciência, acedido em 24-03-2021, em https://www.dgeec.mec.pt/np4/glossario.html

Domingo-Ferrer, J. e Torra, V., (2001). *Disclosure Control Methods and Information Loss for Microdata*. *Cap. 5, of*: Doyle, P., Lane, J.I., Theeuwes, J.J.M. e Zayatz, L.V.(eds), *Confidentiality, Disclosure and Data Access: Theory and Practical Applications for Statistical Agencies.;*

Domingo-Ferrer, J., and J.M. Mateo-Sanz., (2002). 'Practical Data-Oriented Microaggregation for Statistical Disclosure Control', Transactions on Knowledge and Data Engineering (forthcoming).;

Domingo-Ferrer, J.; Martinez-Ballesté, A.; Mateo-Sanz, J.; Sebé, F., (2006). *Efficient multivariate data-oriented microaggregation*. VLDB J.;

Domingo-Ferrer, J.; Sebé, F.; Solanas, A., (2008). *A polynomial-time approximation to optimal multivariate microaggregation.* Comput.;

Drechsler, J., Reiter, J. (2010); “Sampling with synthesis: A new approach for releasing public use census microdata”; Journal of the American Statistical; 2010.;

Fayyoumi, E.; Oommen, B.J. (2010). *A survey on statistical disclosure control and micro-aggregation techniques for secure statistical databases*. Softw. Pract. Exp.;

Fung, B. C. M., Wang, K., Chen, R., and Yu, P. S., (2010). *Privacy-preserving data publishing: A survey of recent developments*. ACM Computer Survey.;

Greenberg, B., (1987). *Rank swapping for ordinal data*, Washington, DC: U. S. Bureau of the Census (unpublished manuscript).;

Heaton, B.; Mukherjee, S. (2011). *Record Ordering Heuristics for Disclosure Control through Microaggregation*. In Proceedings of the International Conference on Advances in Communication and Information Technology, Amsterdam, The Netherlands, 1–2 December.;

Hout, A. van den, Elamir, E. A. H., (2006). "*Statistical Disclosure Control Using Post Randomisation : Variants and Measures for Disclosure*", Journal of Official Statistics 22.;

Huang e Williamson., (2001). *A comparison of synthetic reconstruction and combinatorial optimisation approaches to the creation of small-area microdata*.;

Hundepool, A., Domingo-Ferrer, J., Franconi, L., Giessing, S., Lenz, R., Longhurst, J.,Nordholt, E.S., Seri, G. e Wolf, P.P. (2009).;

Hundepool, A., et al., (2007). *Handbook on statistical disclosure control*. Brand, R. (2002). *Microdata protection through noise addition*. Lecture Notes in Computer Science London: Springer.;

Jarque, C.M. and Bera, A.K. (1987) *A Test for Normality of Observations and Regression Residuals*. International Statistical Review, 55.;

K. Muralidhar and R. Sarathy., (2006). ”*Data shuffling - A new masking approach for numerical data*,” Management Science.;

Laszlo, M.; Mukherjee, S., (2005). *Minimum spanning tree partitioning algorithm for microaggregation*. IEEE Trans. Knowl.;

Lin, J.L.; Wen, T.H.; Hsieh, J.C.; Chang, P.C., (2010). *Density-based microaggregation for statistical disclosure control.* Expert Syst.;

Loukides, G., & Lomax, N., (2021). *Privacy-preserving data publishing through anonymization, statistical disclosure control*, *and de-identification*, acedido em https://doi.org/10.31219/osf.io/2fvj7

Matthias Templ., (2017). "*Statistical disclosure control for Microdata: methods and applications in R*"

Mendes, E., (2010). " *Confidencialidade dos Dados: Aplicação e Comparação de Técnicas de Controlo da Divulgação Estatística",* Dissertação de Mestrado, Faculdade de Economia da Universidade do Porto

Moore, R., (1996). *Controlled data swapping techniques for masking public use microdata sets*, 1996. U. S. Bureau of the Census, Washington.;

Mortazavi, R.; Jalili, S.; Gohargazi, H., (2013). *Multivariate microaggregation by iterative optimization*.;

Oganian, A:, and Domingo-Ferrer, J., (2001). *On the complexity of optimal microaggregation for statistical Microdata* - 85 - disclosure control. Statistical Journal of the United Nations Economic Comission for Europe, 2001.;

Panagiotakis, C.; Tziritas, G., (2011). *Successive group selection for microaggregation*. IEEE Trans. Knowl. Data Eng.;

Reiss, S. P., (1984). *Practical data-swapping: the first steps. ACM Transactions on Database Systems*, 1984.;

Reiss, S. P., Post, M. J., and Dalenius, T., (1982). *Nonreversible privacy transformations. In Proceedings of the ACM Symposium on Principles of Database Systems*, pages 139ñ146, Los Angeles, CA, 1982. ACM.;

SHAPIRO, S. S., & WILK, M. B. (1965). *An analysis of variance test for normality* (complete samples). Biometrika, 52(3–4), 591–611. https://doi.org/10.1093/biomet/52.3-4.591

Shmoys, D.B.; Lenstra, J.; Kan, A.R.; Lawler, E.L., (1985). *The Traveling Salesman Problem; John Wiley & Sons*, Incorporated: Hoboken, NJ, USA,; Volume 12.;

Solanas, A., (2008), "*Privacy protection with genetic algorithms*", Studies in Computational Intellegence *237* (2008).;

Solanas, A.; Martinez, A., (2021). VMDAV: *A Multivariate Microaggregation With Variable Group Size*. In Proceedings of the 17th COMPSTAT Symposium of the IASC.;

Solé, M.; Muntés-Mulero, V.; Nin, J., (2012). *Efficient microaggregation techniques for large numerical data volumes*.;

Waal, t. e Willenborg, L., (2001). *Elements of Statistical Disclosure Control*, Lecture Notes in Statistics, 155. *(Springer).;*

Wolf, P.P., Gouweleeuw, J. M., Kooiman e P., Willenborg, L., (1998), *Reflections on PRAM. Proceedings of the conference: Statistical Data Protection.* Lisboa.;

Wong, R. C. and Fu, A. W., (2010). *Privacy-Preserving Data Publishing: An Overview*. Synthesis Lectures on Data Management. Morgan & Claypool Publishers.;

Yang, G.; Ye, X.; Fang, X.; Wu, R.; Wang, L., (2020). *Associated attribute-aware differentially private data publishing via microaggregation*.;

Zigomitros, A.; Casino, F.; Solanas, A.; Patsakis, C., (2020). *A Survey on Privacy Properties for Data Publishing of Relational Data*.;

14 Anexo

Variáveis da base dos dados

tina:

Descrição: Número de Identificação fiscal anonimizado

Natureza: Variável categórica

bina:

Descrição: Número de identificação anonimizado do Banco

Natureza: Variável categórica

cina:

Descrição: Número de identificação anonimizado do crédito

Natureza: Variável categórica

date:

Descrição: Data de referência (Mês/ Ano)

Natureza: Variável categórica

devediridbp_r :

Descrição: Código de identificação do devedor

Natureza: Variável categórica

nivelresponsabilidade_r :

Descrição: Nível de responsabilidade

Natureza: Variável categórica

Níveis:

Classificação	**Definição**
1	Crédito individual
2	Crédito conjunto – 1º Devedor
3	Crédito conjunto - Restantes
4	Fiador individual
5	Fiador conjunto

situacaocredito_r :

Descrição: Situação do crédito

Natureza: Variável categórica

Classificação	**Definição**
1	Crédito regular
2	Crédito potencial
3	Crédito vencido
4	‘credito baixado’
5	Crédito renegociado
6	Crédito vencido em litígio

classecreditovencido_r :

Descrição: Classe de crédito vencido

Natureza: Variável categórica

Classificação	**Definição**
1	Até 1 mês
2	De 1 a 2 meses
3	De 2 a 3 meses
4	De 3 a 6 meses
5	De 6 a 9 meses
6	De 9 a 12 meses
7	De 12 a 15 meses
8	De 15 a 18 meses
9	De 18 a 24 meses
10	De 24 a 30 meses
11	De 30 a 36 meses
12	De 36 a 48 meses
13	De 48 a 60 meses
14	Mais de 60 meses

prazooriginal_r/ prazoresidual_r:

Descrição:

- Prazo original: Maturidade do crédito definido no momento do acordo contratual
- Prazo residual: Intervalo de tempo entre a data de referência e o momento do acordo contratual

Natureza: Variável categórica

Classificação	**Definição**
1	Indefinido
2	Até 90 dias
3	De 90 a 180 dias
4	De 180 dias a 1 ano
51	De 1 a 2 anos
52	De 2 a 3 anos
53	De 3 a 4 anos
54	De 4 a 5 anos
61	De 5 a 6 anos
62	De 6 a 7 anos
63	De 7 a 8 anos
64	De 8 a 9 anos
65	De 9 a 10 anos
71	De 10 a 15 anos
72	De 15 a 20 anos
8	De 20 a 25 anos
9	De 25 a 30 anos
10	Mais de 30 anos

Paisbalcaoid_r:

Descrição: Código do país de residência do devedor

Natureza: Variável Categórica

Nivéis:

Código do País	Nome do País
CYM	Ilhas Caimão
ESP	Espanha
FRA	França
GBR	Reino Unido
LUX	Luxemburgo
PRT	Portugal

produto_r :

Descrição: Produto financeiro

Natureza: Variável categórica

Classificação	**Definição**
1	Descontos e outros créditos
2	Conta corrente (linhas de crédito)
3	Descontos em depósito à ordem
4	Factorização do recurso
5	Factorização sem recurso
6	Aluguer de bens imobiliários
7	Aluguer de bens não imobiliários
8	Financiamento da atividade empresarial
9	Cartão de crédito
10	Crédito hipotecário
11	Crédito consumo

12	Crédito automóvel
13	Outros créditos
14	Garantias bancárias de outras entidades participantes
15	Garantias bancárias de outros participantes

Moeda_r:

Descrição: Código da moeda do país de residência do devedor

Natureza: Variável Categórica

Nivéis:

Código da Moeda	Nome da moeda
EUR	Euro

valor_r:

Descrição: Montante total da dívida em circulação, em euros

Natureza: Variável contínua